DIFFUSION TRANSFORMERS ARE PROVABLY OPTIMAL IN-CONTEXT GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative foundation models are attracting interest for their ability to produce desired outputs from demonstrations given at inference time, without updating parameters. However, since a few demonstrations cannot uniquely identify the intended task, the challenge is how to learn and sample from an output distribution that reflects this task uncertainty. In this work, we theoretically analyze how a Diffusion Transformer (DiT), pretrained across diverse tasks, learns and generates predictive distributions for a new query from demonstrations. We first show that the natural target to generate from finite demonstrations is not an output derived from estimating a single task, but rather a predictive distribution that captures the task uncertainty remaining after observing the demonstrations. We then prove that a DiT can learn this predictive distribution through score estimation, using attention to aggregate information from demonstrations and diffusion to generate samples. Owing to this property, with sufficient pretraining resources and diffusion sampling steps, the resulting DiT achieves the minimax optimal rate over a Hölder class of test-time tasks. These results imply that DiT acts as a statistically grounded in-context generator capable of generating distributions adapted to new tasks while retaining the uncertainty inherent in finite demonstrations.

1 INTRODUCTION

Generative models have advanced rapidly over the past few years with diffusion models in particular achieving remarkable success in synthesizing complex high-dimensional data such as images (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022; Chi et al., 2023). At the same time, in-context learning (ICL) has shown that a pretrained model can adapt to a new task from a few demonstrations without updating its parameters (Brown et al., 2020; Dong et al., 2024). These developments motivate *in-context generation*: using demonstrations to specify what kinds of outputs should be generated for a new query (Bar et al., 2022; Wang et al., 2023). Fig. 1 illustrates this idea in a robotic manipulation example: a user provides several examples of acceptable behaviors and asks the model to generate an appropriate action for a new object. More broadly, in image generation and other structured domains (Fu et al., 2025; Gu et al., 2024), the desired behavior may depend on visual, geometric, or stylistic criteria that are cumbersome to specify perfectly in natural language or mathematical form. In such cases, demonstrations serve as a natural interface for specifying what should be generated, and in-context generation is the mechanism that turns this interface into a working system.

Yet demonstrations are not necessarily a complete specification of the underlying task. With only finitely many examples, multiple latent rules may remain compatible with the same context while implying different outputs for a new query, leaving uncertainty about what should be generated (Xie et al., 2022; Panwar et al., 2024). The output distribution that the context induces at a new query is then frequently non-degenerate or even multimodal. Collapsing this distribution to its mean can yield implausible outputs (Fig. 1; see also Blau & Michaeli, 2018; Chi et al., 2023). Selecting a single plausible candidate instead discards the relative probabilities of alternatives, preventing faithful reproduction of the distribution and potentially compromising downstream decision-making under uncertainty (Kohl et al., 2018; Hoeting et al., 1999; Gneiting & Raftery, 2007). Thus, what in-context generation should deliver is not a single prediction but samples drawn from the predictive distribution itself, so that the uncertainty left by the finite demonstrations is preserved.

Despite substantial theoretical progress on both in-context learning and diffusion models, existing frameworks do not cover this generative setting. Most finite-sample statistical analyses of ICL

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

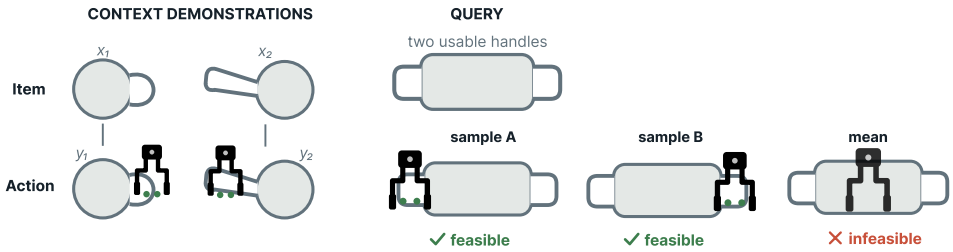

Figure 1: **In-context generation from demonstrations.** A user specifies a desired behavior through a few examples rather than an explicit rule. Because a finite context may be consistent with multiple underlying rules, different outputs can remain plausible for the same query; collapsing the plausible outputs to a single average can produce an invalid result.

focus on point prediction under regression-type losses, studying how pretrained Transformers output optimal regression rules or Bayes predictors from a finite context (Bai et al., 2023; Kim et al., 2024; Wakayama & Suzuki, 2026). In contrast, conditional generative modeling and conditional diffusion models aim to learn an entire output distribution given an observed covariate or conditioning variable (Fu et al., 2024; Hu et al., 2025). These settings typically assume that, at the population level, the observed condition indexes a well-defined conditional distribution. In-context generation is different: the condition itself is a finite collection of demonstrations generated from an unobserved task, and provides only partial information about the test distribution. Consequently, it remains unclear how a predictive distribution should be formed from finitely many demonstrations, and how a pretrained generative model can learn and sample from this distribution across tasks.

In this paper, we formulate in-context generation as sampling from the distribution of outputs that remain plausible after finite demonstrations, rather than committing to a single inferred task or output. We identify this target with the posterior-predictive distribution and give a finite-sample theory of how a pretrained diffusion Transformer (DiT; Peebles & Xie, 2023) learns the corresponding prompt-to-distribution map and realizes it through diffusion sampling. Our main findings are as follows.

- **Pretraining learns the shared rule; demonstrations reduce task uncertainty.** We identify two distinct error sources: error from learning and sampling the shared prompt-to-distribution mapping, and the uncertainty left by the current prompt. More pretraining data and computation can reduce the former, but cannot in general eliminate the latter.
- **Attention infers; diffusion generates.** Our construction shows that attention aggregates evidence from demonstrations into weights over plausible tasks, while diffusion converts the resulting predictive mixture into samples. With finite pretraining data and no task labels, a sufficiently large DiT and sampler can approximate the posterior-predictive distribution to any prescribed accuracy under our assumptions.
- **Minimax-optimal in-context generation via DiT.** We establish matching upper and lower bounds for finite-prompt uncertainty and show that, over the nondegenerate Hölder task class, the learned DiT attains the minimax-optimal prompt-length rate $\sqrt{\log n/n}$ in task-conditional total-variation error, given sufficient model capacity, pretraining episodes, and diffusion steps. This demonstrates the test-time adaptive generation of DiT.

The paper is organized as follows. Section 2 defines the data-generating process and evaluation criteria, then introduces the DiT, its training objective, and its reverse sampler. Section 3 explains the task-family assumptions and the score-learning, prompt-uncertainty, and task-conditional generation guarantees, together with the main ideas behind their proofs. Section 4 presents a summary and practical implications. The appendices contain the proofs and empirical illustrations.

1.1 Related Work

(A) ICL theory. Existing statistical theories of ICL primarily study point prediction, often under regression-type losses, and characterize how pretrained Transformers implement or approximate task-adaptive estimators (Bai et al., 2023; Li et al., 2023; Zhang et al., 2024b; Kim et al., 2024). Bayesian analyses further interpret ICL through posterior prediction over latent tasks (Wakayama & Suzuki, 2026; Ma et al., 2025). Our focus is different: we study generation of the full posterior-predictive output distribution and its finite-sample error under distributional loss. Kang et al. (2026) give a

constructive Transformer approximation of the posterior-predictive distribution in Gaussian-process (GP) regression. However, their theory assumes a known GP model with fixed kernel and noise parameters and does not establish generalization from finite pretraining data. By contrast, our result learns a prompt-conditioned diffusion score from finitely many independent pretraining examples without knowing the Hölder task family or task prior, and controls expected posterior-predictive TV error of the resulting finite-step reverse diffusion sampler.

(B) Score learning with Transformers. Prior work shows Transformers can represent score estimators or denoising updates (Wu et al., 2025; Smart et al., 2025; Rosu et al., 2025); Adibi et al. (2026) construct frozen Transformers simulating closed-form diffusion recursions and approximating an estimation-free empirical sampler. Our focus is statistical rather than merely representational: we analyze a DiT learned from finite pretraining data and show that finite-step reverse diffusion achieves posterior-predictive accuracy, achieving the minimax-optimal rate with respect to prompt length.

Further related work on diffusion model theory is discussed in Appendix A.

2 In-Context Generation Setup and Diffusion Transformer

We first specify the data-generating process and the two output distributions used to evaluate in-context generation. We then introduce the prompt-conditioned Diffusion Transformer, its training objective, and its reverse diffusion sampler. The task-family assumptions used to prove the guarantees are stated separately in Section 3. Appendix B gives the formal probability conventions.

2.1 Data-generating process

Fix a prompt length $n \geq 1$ and an output dimension $d_z \geq 1$, and let $\mathcal{Z} = [0,1]^{d_z}$ be the output domain. Let $\mathcal{T}$ be a standard Borel task space, and let Π be the Borel probability distribution over tasks encountered during pretraining. The same distribution serves as the task prior in posterior prediction. Adapting the task-mixture prompt-generating formulation (Ching et al., 2026; Wakayama & Suzuki, 2026; Ma et al., 2025) to distributional generation, we consider the following two models.

Definition 2.1 (Prompt-generating process)**.** *Each prompt and its target are generated as follows.*

1. *Sample a latent task* $\Theta \sim \Pi$.
2. *Conditional on* Θ*, generate* $n+1$ *complete examples independently from the task-specific distribution:*
   - *For* ***sample-only prompts****, a task* θ *specifies a distribution* P_θ *on* $\mathcal{Z}$*, where* $\theta \mapsto P_\theta$ *is a Borel probability kernel, and* $Z_1, \ldots, Z_n, Z_\star \overset{\text{iid}}{\sim} P_\Theta$.
   - *For* ***input–output prompts****, fix an input dimension* $d_x \geq 1$ *and let* $\mathcal{X} = [0,1]^{d_x}$*. A task* θ *specifies a joint distribution* $P_\theta^{X,Z}$ *on* $\mathcal{X} \times \mathcal{Z}$*, where* $\theta \mapsto P_\theta^{X,Z}$ *is a Borel probability kernel, and* $(X_1, Z_1), \ldots, (X_n, Z_n), (X_\star, Z_\star) \overset{\text{iid}}{\sim} P_\Theta^{X,Z}$.
3. *Form the context from the first* n *examples and hold out the output in the final example as the target:* $C_n = Z_{1:n}, H_n = C_n$ *for sample-only prompts and* $C_n = ((X_i, Z_i))_{i=1}^n, H_n = (C_n, X_\star)$ *for input–output prompts.*

Write $\mathcal{H}_n := \mathcal{Z}^n$ and $\mathcal{H}_n := (\mathcal{X} \times \mathcal{Z})^n \times \mathcal{X}$ for sample-only and input–output prompts, respectively, so that $H_n \in \mathcal{H}_n$. The prompt never includes the target output $Z_\star$. In the input–output model, it includes the observed query $X_\star$, which can itself provide information about Θ when the input marginal is task-dependent. We call one draw of a latent task together with its prompt and held-out target an **episode**; its observed part is the prompt–target pair $(H_n, Z_\star)$. Under either prompt model, the $n+1$ complete examples are conditionally i.i.d. given Θ and are generally dependent after marginalizing over their shared latent task. In the input–output prompt model, H_n contains $X_\star$, so H_n and $Z_\star$ need not be conditionally independent given Θ.

Pretraining and test-time generation. Repeating this procedure independently N times yields the pretraining prompt–output pairs $((H_n^{(e)}, Z_\star^{(e)}))_{e=1}^N$. The model is pretrained on these pairs without observing the task labels. At test time, a fresh episode is generated independently using the same meta-prior Π, and the pretrained model generates an output conditioned on the observed prompt H_n. For a fixed pretrained model and prompt, let $\widehat{P}(\cdot \mid H_n)$ denote the output distribution induced by the generation procedure's randomness.

2.2 Risk decomposition via a posterior-predictive oracle

Task-conditional generation risk. Our goal is distributional: generate from the task's output distribution underlying the prompt rather than recover its latent parameter Θ:

$$P(\cdot \mid H_n, \Theta) = \mathsf{Law}(Z_\star \mid H_n, \Theta).$$

The task-conditional distribution equals P_Θ for sample-only prompts and $P_\Theta^{Z \mid X}(\cdot \mid X_\star)$ for input–output prompts, almost surely. For any Borel probability kernel $Q(\cdot \mid H_n)$, define its generation risk by the total variation (TV) error averaged over a fresh test episode[1]:

$$\mathcal{R}_n(Q) := \mathbb{E}\mathsf{TV}\big(Q(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big). \tag{1}$$

The generator observes H_n but not Θ, and a finite prompt need not identify the task that produced it.

A posterior-predictive oracle. The posterior predictive distribution $P(\cdot \mid H_n) = \mathsf{Law}(Z_\star \mid H_n)$ averages the task-conditional output distributions given the observed prompt:

$$P(\cdot \mid h) = \int P(\cdot \mid h, \theta)\, \Pi(\mathrm{d}\theta \mid h), \qquad \text{for } \mathsf{Law}(H_n)\text{-almost every } h. \tag{2}$$

In the input–output model, this posterior conditions on both C_n and $X_\star$, so the query can update the task probabilities when the input marginal depends on the task. For a fixed task family and prior, convexity of TV and the triangle inequality give $\inf_Q \mathcal{R}_n(Q) \le \mathcal{R}_n(P) \le 2 \inf_Q \mathcal{R}_n(Q)$, where the infimum is over all prompt-based probability kernels, which may know the task family and prior Π. Thus, the posterior predictive distribution is a constant-factor oracle under this information constraint.

This oracle motivates (implicitly) learning the shared map $h \mapsto P(\cdot \mid h)$ as a latent teacher during pretraining. For DiT, this means learning the diffusion score of the posterior predictive distribution, which is also the target of the population denoising objective in Section 2.4. Using this implicit target in the error bound separates learning and sampling errors from the uncertainty left by a finite test prompt; i.e., for the learned generator $\widehat{P}$, the triangle inequality gives the principled decomposition

$$\underbrace{\mathbb{E}\mathsf{TV}\big(\widehat{P}(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big)}_{\text{task-conditional generation error}} \le \underbrace{\mathbb{E}\mathsf{TV}\big(\widehat{P}(\cdot \mid H_n), P(\cdot \mid H_n)\big)}_{\text{posterior-predictive generation error}} + \underbrace{\mathbb{E}\mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big)}_{\text{finite-prompt task uncertainty}}. \tag{3}$$

The first term measures how accurately the pretrained model and its sampler approximate the posterior-predictive target; it includes model approximation, estimation from finite pretraining data, optimization, and sampling errors. The second term equals the oracle risk $\mathcal{R}_n(P)$ and measures the uncertainty about the task's output distribution left by the finite test prompt. It depends on the task's difficulty and prompt length n, and can remain positive even with exact posterior-predictive generation. This uncertainty concerns output distributions rather than task-label identification: distinct tasks may induce the same output distribution.

2.3 Architecture: Prompt-conditioned Diffusion Transformer

We use a prompt-conditioned DiT to learn the score for posterior-predictive generation. Following prior theoretical work on Transformer-based ICL (e.g., Bai et al., 2023; Ahn et al., 2023; Kim et al., 2024), we analyze a simplified Transformer architecture. Details are deferred to Appendices E and G.

Conditional diffusion process. Keeping the prompt fixed, we perturb $Z_0 \equiv Z_\star$ through the Ornstein–Uhlenbeck (OU) process:

$$\mathrm{d}Z_t = -Z_t\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t, \qquad Z_t \overset{\mathrm{d}}{=} \alpha_t Z_0 + \sigma_t \xi, \quad \alpha_t = e^{-t}, \quad \sigma_t^2 = 1 - e^{-2t}. \tag{4}$$

Here W_t is standard d_z-dimensional Brownian motion independent of (Z_0, H_n), and $\xi \sim \mathcal{N}(0, I_{d_z})$ is independent of the episode. Write $P_{0|h} = P(\cdot \mid h)$ and $P_{t|h} = \mathsf{Law}(Z_t \mid H_n = h)$. For $t > 0$, the latter has a smooth density $p_t(\cdot \mid h)$ with score

$$s^\star(z, t \mid h) = \nabla_z \log p_t(z \mid h), \qquad z \in \mathbb{R}^{d_z}. \tag{5}$$

This score determines the exact reverse drift (Haussmann & Pardoux, 1986). As $t \to \infty$, $P_{t|h}$ approaches $\mathcal{N}(0, I_{d_z})$; starting from this Gaussian at finite T incurs an initialization error.

[1] Unless stated otherwise, $\mathbb{E}$ in a test-risk bound conditions on the completed training run and averages over the fresh task Θ and prompt H_n, including the query $X_\star$ in the input–output model.

Prompt-conditioned DiT model. For the variance-balanced denoising objective in Section 2.4, the network predicts the scaled score $\widetilde{s}^\star = \sigma_t s^\star$. Feedforward ReLU networks encode each context example and the query–state–time input $(X_\star, z, t)$, with $X_\star$ omitted for sample-only prompts.

The encoded token array $\mathbf{E}^{(0)}$ passes through four masked softmax-attention blocks, each with a residual connection followed by a token-wise coordinate update:

$$\mathbf{E}^{(k)} = \mathsf{R}_{k,n}\big(\mathbf{E}^{(k-1)} + \mathrm{Att}_{\mathbf{v},k}(\mathbf{E}^{(k-1)})\big), \qquad k = 1, \dots, 4.$$

Here $\mathrm{Att}_{\mathbf{v},k}$ uses trainable query, key, value, and output projections with a fixed attention mask, whereas $\mathsf{R}_{k,n}$ is a fixed affine–ReLU map applied separately to each token. A fixed ReLU readout maps the final output token to a d_z-dimensional vector, and radial projection produces $\widetilde{s}_{\mathbf{v}}(z, t, h)$. We defer the specific interpretation of the attention blocks to Section 3.1.

Transformer score class. Fix $0 < t_0 < T$ and a deterministic output radius $R_{\mathrm{out},n} > 0$. Each admissible parameter vector $\mathbf{v}$ defines a jointly Borel map

$$\widetilde{s}_{\mathbf{v}} : \mathbb{R}^{d_z} \times [t_0, T] \times \mathcal{H}_n \longrightarrow \{u \in \mathbb{R}^{d_z} : \|u\|_2 \le R_{\mathrm{out},n}\}, \qquad s_{\mathbf{v}}(z, t, h) = \frac{\widetilde{s}_{\mathbf{v}}(z, t, h)}{\sigma_t}.$$

The trainable coordinates comprise the designated encoder coefficients, prior logits, and attention projection matrices, and $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$ denotes their total number. We denote by $\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon} = \{s_{\mathbf{v}}\}$ the class whose architecture and parameter bounds are chosen using only (n, ε) and the fixed dimensions and regularity bounds in Assumption 3.1. The class is independent of the unknown task family, prior, training data, and confidence level. The reverse sampler uses the unscaled score $s_{\mathbf{v}}$.

2.4 DiT TRAINING AND REVERSE DIFFUSION SAMPLING

We connect DiT pretraining to posterior-predictive generation through three stages: a population variance-balanced DSM (bDSM) objective, its clipped finite-episode empirical counterpart, and randomized reverse diffusion sampling. Details are deferred to Appendices F and G.

Population score error and bDSM. For $0 < t_0 < T$, define the integrated score-estimation error

$$\mathcal{E}_{\mathrm{score}}(s) := \mathbb{E}_{H_n} \int_{t_0}^{T} \big\| s(\cdot, t, H_n) - s^\star(\cdot, t \mid H_n) \big\|^2_{L^2(P_{t|H_n})}\, dt. \tag{6}$$

Since $s^\star$ is unknown, denoising score matching (DSM) instead uses the Gaussian perturbation as the regression target (Vincent, 2011; Song et al., 2021b). Predicting the scaled score $\tilde{s} = \sigma_t s$ replaces $-\xi_{\mathrm{dsm}}/\sigma_t$ with the unit-scale target $-\xi_{\mathrm{dsm}}$, but weights squared score error by σ_t^2. We compensate for this weighting by sampling the diffusion time with density $\varpi_{t_0,T}(t) := (A_{t_0,T}\sigma_t^2)^{-1}$, where $A_{t_0,T} := \int_{t_0}^{T} \sigma_t^{-2}\, dt$. The positive cutoff $t_0 > 0$ makes this density normalizable.

Draw $\tau \sim \varpi_{t_0,T}$ and $\xi_{\mathrm{dsm}} \sim \mathcal{N}(0, I_{d_z})$ independently of each other and the episode, and set $Z_\tau = \alpha_\tau Z_0 + \sigma_\tau \xi_{\mathrm{dsm}}$. The bDSM loss is

$$\mathcal{R}_{\mathrm{bDSM}}(s) := A_{t_0,T}\, \mathbb{E}\big\| \tilde{s}(Z_\tau, \tau, H_n) + \xi_{\mathrm{dsm}} \big\|_2^2. \tag{7}$$

Under the finite-risk conditions in Proposition C.1, we have

$$\mathcal{R}_{\mathrm{bDSM}}(s) - \mathcal{R}_{\mathrm{bDSM}}(s^\star) = \mathcal{E}_{\mathrm{score}}(s). \tag{8}$$

Thus, excess bDSM risk controls the integrated score error required by the reverse-sampling analysis.

Finite-episode pretraining. For each of the N independent training episodes, independently draw $\tau_e \sim \varpi_{t_0,T}$ and $\xi^{(e)}_{\mathrm{dsm}} \sim \mathcal{N}(0, I_{d_z})$, and form $Z^{(e)}_{\tau_e} = \alpha_{\tau_e} Z^{(e)}_0 + \sigma_{\tau_e} \xi^{(e)}_{\mathrm{dsm}}$. To obtain a bounded empirical loss, define $\xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}} := \mathrm{clip}_{R_{\mathrm{lab}}}(-\xi^{(e)}_{\mathrm{dsm}})$, where clip_R is radial projection onto the Euclidean ball of radius R. The empirical objective is

$$\widehat{\mathcal{R}}^{(R_{\mathrm{lab}})}_{N,\mathrm{bDSM}}(s) = \frac{A_{t_0,T}}{N} \sum_{e=1}^{N} \big\| \tilde{s}(Z^{(e)}_{\tau_e}, \tau_e, H^{(e)}_n) - \xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}} \big\|_2^2.$$

Let $\widehat{s}_{N,n,\varepsilon}$ be an $\varepsilon_{\mathrm{opt}}$-approximate minimizer over $\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}$ of this $A_{t_0,T}$-scaled objective.

Randomized reverse sampler. Given a new prompt h, we use an exponential–Euler reverse sampler on a randomly shifted time grid. Averaging over the shift connects score errors at the sampler's evaluation times to the integrated error controlled by bDSM.

Define the inverse-variance clock $u(t) := \int_{t_0}^{t} \sigma_r^{-2}\,\mathrm{d}r$, so its total length is $A_{t_0,T} = u(T)$. Choose an integer M satisfying $M \geq \max\{\lceil 2A_{t_0,T}\rceil, \lceil A_{t_0,T}/(T-t_0)\rceil\}$, as in (136), sample a uniform shift over one u-clock interval of length $A_{t_0,T}/M$, and let the resulting forward-time nodes be $v_{J_{\mathrm{grid}}} > \cdots > v_0 > v_{-1} = t_0$, where the sampler makes $J_{\mathrm{grid}} + 1 \leq M$ score evaluations almost surely. Set $\Delta_j = v_j - v_{j-1}$. Initialize $\widehat{Z}_{v_{J_{\mathrm{grid}}}} \sim \mathcal{N}(0, I_{d_z})$ and iterate

$$\widehat{Z}_{v_{j-1}} = e^{\Delta_j}\widehat{Z}_{v_j} + 2(e^{\Delta_j}-1)\widehat{s}_{N,n,\varepsilon}(\widehat{Z}_{v_j}, v_j, h) + \sqrt{e^{2\Delta_j}-1}\,\xi_j, \qquad j = J_{\mathrm{grid}}, \ldots, 0.$$

The vectors ξ_j are independent and standard Gaussian. The law of $\mathrm{Proj}_{\mathcal{Z}}(\widehat{Z}_{v_{-1}})$, averaged over all sampler randomness, is $\widehat{P}^{(M)}_{0|h} = \widehat{P}(\cdot \mid h)$. The sampler uses at most M score evaluations almost surely. Because the reverse process stops at the positive time t_0, the projected endpoint incurs an early-stopping error, controlled in Appendix G.

3 Theoretical Guarantees

Our theoretical guarantees follow the decomposition in (3).

$$\underbrace{\mathcal{R}_n(\widehat{P})}_{\text{total risk}} \leq \underbrace{\mathbb{E}\mathsf{TV}\big(\widehat{P}(\cdot \mid H_n), P(\cdot \mid H_n)\big)}_{\substack{\text{pretraining and sampling} \\ \to \text{ Section 3.1}}} + \underbrace{\mathbb{E}\mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big)}_{\substack{\text{finite-prompt uncertainty} \\ \to \text{ Section 3.2}}} \lesssim \underbrace{\begin{pmatrix}\text{minimax-optimal} \\ \text{prompt rate}\end{pmatrix}}_{\to \text{ Section 3.3}}$$

Section 3.1 shows that a DiT can approximate the posterior-predictive oracle from finitely many pretraining episodes without knowing the task family or prior. The analysis controls score learning from pretraining and translates score error into generation error through finite-step reverse diffusion. Also, finite prompts can leave task uncertainty even under exact posterior-predictive generation. Section 3.2 bounds this uncertainty and establishes a matching worst-case lower bound, even for generators that know the task family and prior. Section 3.3 then combines the bounds at the prompt-uncertainty scale to obtain a task-conditional generation guarantee, establishing minimax optimality in the prompt length under the assumption below.

Task-family assumptions. The episodic setup in Section 2 does not itself require densities or regularity conditions. The guarantees below specialize it to an idealized class of positive task densities with regularity in both the observations and the latent task parameter. The corresponding density domain has dimension $d_{\mathrm{stat}} = d_z$ in the sample-only model and $d_{\mathrm{stat}} = d_x + d_z$ in the input–output model. Write $\mathcal{F}^{d}_{\beta,+}(L_\beta, \underline{p})$ for the densities on $[0,1]^d$ with β-Hölder norm at most L_β and lower bound $\underline{p} > 0$ (Appendix B.4).

Assumption 3.1 (Smooth latent tasks with positive densities)**.** *Tasks are indexed by a parameter $\theta \in \mathcal{T} = [0,1]^{d_\theta}$, with $d_\theta \in \mathbb{N}$ fixed across the model class and independent of n, but the sampled Θ remains latent and Π may be any Borel probability measure on $\mathcal{T}$. Let g_θ denote the Lebesgue density of P_θ or $P_\theta^{X,Z}$ in the sample-only or input–output model, respectively. Fix $\beta > 0$, $L_\beta > 1$, $\underline{p} \in (0,1]$, $\beta_\theta \in (0,1]$, and $H_\theta \in [0,\infty)$. The map $(\theta, u) \mapsto g_\theta(u)$ is jointly Borel and, for all $\theta, \theta' \in \mathcal{T}$,*

$$g_\theta \in \mathcal{F}^{d_{\mathrm{stat}}}_{\beta,+}(L_\beta, \underline{p}), \qquad \sup_{u \in [0,1]^{d_{\mathrm{stat}}}} |g_\theta(u) - g_{\theta'}(u)| \leq H_\theta \|\theta - \theta'\|_2^{\beta_\theta}.$$

Fix either the sample-only or input–output prompt model. Let $\mathfrak{M}_{\mathrm{lat}}$ denote the class of task families

$$\mathcal{M} = \big(\{g_\theta\}_{\theta \in \mathcal{T}}, \Pi\big)$$

that satisfy Assumption 3.1 with the same displayed dimensions and regularity bounds. Each $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ induces a hierarchical distribution of $(\Theta, H_n, Z_\star)$ through Definition 2.1. Pretraining observes independent copies of $(H_n, Z_\star)$, while test risk uses an independent fresh copy of $(\Theta, H_n, Z_\star)$ and gives the trained generator only H_n. The learner knows the common dimensions and regularity bounds, but not $\mathcal{M}$, its task-density family, its task prior, or the sampled task parameters.

This assumption is a model-level analogue of the Hölder regularity imposed on the Bayes predictor in Wakayama & Suzuki (2026). We instead impose the regularity on the task-generating densities rather than on the prompt-to-score map itself, making the condition explicit and readily verifiable for concrete distribution families; see Appendix I for examples including a multi-modal task family. The sampling result also holds under weaker spatial conditions stated locally in Corollary G.6.

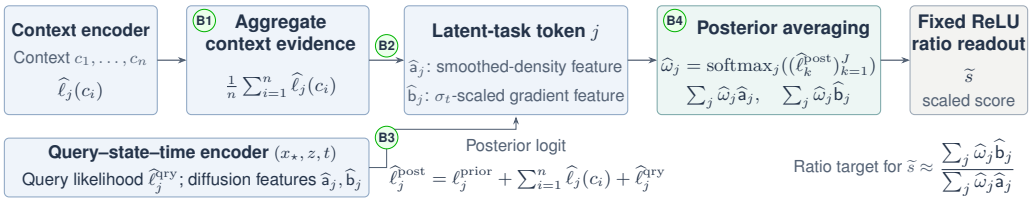

Figure 2: Conceptual view of the representability construction. The four attention blocks: B1 pools context log likelihoods; B2 copies the pooled values to each latent-task token; B3 forms task logits and attaches diffusion features; and B4 forms posterior-weighted averages of the density and σ_t-scaled gradient features. A fixed ReLU readout approximates their ratio to produce $\widetilde{s}$.

3.1 DiTs learn posterior-predictive generation during pretraining

During pretraining, a DiT can learn to generate the posterior-predictive distribution for a new prompt.

Theorem 3.1 (Posterior-predictive generation; informal version of Theorem G.8)**.** *Under Assumption 3.1, for each $n \geq 2$ and sufficiently small $\varepsilon > 0$, there exists a common DiT class and a choice of diffusion parameters with model size and forward-pass cost polynomial in (n, ε^{-1}). For every fixed $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ and $\delta \in (0,1)$, pretraining on N independent episodes to empirical optimization gap $\varepsilon_{\mathrm{opt}} \lesssim \varepsilon^2$ and sampling with M admissible reverse steps, as in Section 2.4, yield, with probability at least $1-\delta$ over the training episodes and their diffusion augmentations,*

$$\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0|H_n}, P(\cdot \mid H_n)\big) \lesssim \varepsilon + \log(e/\varepsilon)\sqrt{\frac{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}\log(en/\varepsilon) + \log(2/\delta)}{N}} + \frac{\log(e/\varepsilon)}{\sqrt{M}}, \tag{9}$$

where $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$ is the number of trainable parameters.

The result goes beyond known-model representability results (e.g., Kang et al., 2026): it gives a finite-sample guarantee for learning and generating the posterior-predictive distribution from finite pretraining data, *without task labels or knowledge of the task family or prior*, and with explicit TV control for finite-step reverse diffusion. Thus, posterior-predictive generation is not merely representable under a known model; it is statistically learnable as an implicit target during pretraining.

For $\mathcal{O}(\varepsilon)$ predictive TV error, it suffices to choose

$$N \gtrsim \varepsilon^{-2}\log^2(e/\varepsilon)\big\{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}\log(en/\varepsilon) + \log(2/\delta)\big\} \quad \text{and} \quad M \gtrsim \varepsilon^{-2}\log^2(e/\varepsilon). \tag{10}$$

Up to logarithmic factors, pretraining episodes scale linearly with parameter count, while score evaluations scale as ε^{-2}. Thus, all required resources are polynomial.

Under the task regularities (Assumption 3.1), the main technical challenge is to obtain a common score class with resource bounds polynomial in (n, ε^{-1}) as the prompt length grows. In the following two paragraphs, we explain how DiT realizes posterior averaging and why polynomially many representative tasks suffice.

The structural alignment between Bayes' rule and the attention circuit. A key property of DiT in in-context generation is that attention can implement posterior averaging by pooling example log likelihoods and using softmax to obtain posterior task weights. For a prompt of length n, discretize the latent space into $J_{\theta,n}$ representative tasks $\{\theta_{j,n}\}_{j=1}^{J_{\theta,n}}$ at resolution η. For $\zeta \in (0,1]$, add $\zeta/J_{\theta,n}$ to each prior cell weight and renormalize. For brevity, set $J := J_{\theta,n}$ and $\theta_j := \theta_{j,n}$, and below omit the n-index from other grid quantities and the ζ-dependence from the logits and weights. Set $c_i = z_i$ in the sample-only model and $c_i = (x_i, z_i)$ in the input–output model. Let $\ell_j^{\mathrm{qry}}(x_\star)$ denote the query-input log marginal density, taking it to be zero in the sample-only model, and let ℓ_j^{prior} denote the corresponding prior logit. Bayes' rule gives the task logits:

$$\ell_j^{\mathrm{post}}(h) = \ell_j^{\mathrm{prior}} + \sum_{i=1}^n \log g_{\theta_j}(c_i) + \ell_j^{\mathrm{qry}}(x_\star), \qquad \omega_j(h) = \mathrm{softmax}_j\big((\ell_k^{\mathrm{post}}(h))_{k=1}^J\big). \tag{11}$$

Attention pools the example log likelihoods, while softmax converts the resulting logits into posterior weights. Let $q_{j,t} := q_{\theta_j,t}$ denote the diffused task-conditional density under representative task θ_j,

with conditioning on $x_\star$ omitted in the sample-only model. The finite model's predictive score is

$$s^{\eta,\zeta}(z,t \mid h) = \frac{\sum_{j=1}^{J} \omega_j(h) \nabla_z q_{j,t}(z \mid x_\star)}{\sum_{j=1}^{J} \omega_j(h) q_{j,t}(z \mid x_\star)}. \tag{12}$$

The construction in Figure 2 uses four attention blocks (B1–B4) and a fixed readout to approximate this computation. The final attention block averages density and σ_t-scaled gradient features rather than task scores; the readout approximates their ratio to produce $\widetilde{s}^{\eta,\zeta} = \sigma_t s^{\eta,\zeta}$. Training over the full class need not preserve this internal posterior interpretation. Appendices D.4 and E.1.3 give the formal episode-level discretization, neural approximation, and token-level construction, respectively.

Why polynomially many representative tasks suffice. Uniformly approximating the posterior for every prompt is stronger than the integrated accuracy required by score matching and reverse diffusion. Instead, we compare the original and discretized episode distributions. The conditional KL chain rule and OU entropy dissipation yield

$$\int_{t_0}^{T} \mathbb{E}\big\|s^\star(Z_t,t \mid H_n) - s^{\eta,\zeta}(Z_t,t \mid H_n)\big\|_2^2 \,\mathrm{d}t \lesssim (n+1)\eta^{2\beta_\theta} + \log(1+\zeta), \tag{13}$$

where the expectation uses the original model. Conditional independence makes the KL cost additive over the $n+1$ examples in an episode. Then, choosing $\eta^{\beta_\theta} \asymp \varepsilon/\sqrt{n+1}$ and $\zeta \asymp \varepsilon^2$ gives $\mathcal{O}(\varepsilon^2)$ integrated score error with only $J = J_{\theta,n} \lesssim (1+\sqrt{n+1}/\varepsilon)^{d_\theta/\beta_\theta}$ representative tasks. Combining this bound with the parameter counts for the prior logits, the ReLU approximation of the fixed-dimensional likelihood and diffusion features, and all trainable attention projections, Appendix E.4 shows that the resulting total number of trainable scalar parameters $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$ is polynomial in (n,ε^{-1}). Crucially, $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$ controls the covering number of the same score class and yields the pretraining requirement in (10). Finally, the randomized inverse-variance grid converts the integrated score error controlled by bDSM into finite-step TV accuracy.

3.2 A SHARP INFORMATION RATE FROM FINITE PROMPTS AT TEST TIME

A sharp rate for prompt-based generation. A finite prompt can leave several task-specific output distributions consistent with the demonstrations. The next result gives the sharp worst-case task-conditional accuracy available to any prompt-based generator. The lower bound lets the generator know the full task family and prior and includes generation randomness in its output distribution.

Theorem 3.2 (Sharp finite-prompt rate; informal version of Theorem H.1)**.** *Under Assumption 3.1, the following statements hold.*

- ***Uniform upper bound.*** *Every $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ and every $n \geq 2$ satisfy*
$$\mathbb{E}_{\mathcal{M}}\, \mathrm{TV}(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)) \lesssim \sqrt{\log n/n}.$$
- ***Matching worst-case lower bound.*** *If, in addition, $H_\theta > 0$, $\underline{p} < 1$, and $L_\beta > 1$, then for every sufficiently large n, there is a model $\mathcal{M}_n \in \mathfrak{M}_{\mathrm{lat}}$ for which*
$$\inf_{Q^{\mathcal{M}_n}} \mathbb{E}_{\mathcal{M}_n}\, \mathrm{TV}\big(Q^{\mathcal{M}_n}(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \gtrsim \sqrt{\log n/n},$$
where the infimum is over prompt-only Borel probability kernels from $\mathcal{H}_n$ to $\mathcal{Z}$ that may know $\mathcal{M}_n$.

Posterior prediction makes oracle use of the prompt. The lower bound is an information limit: it remains valid with arbitrarily many independent pretraining episodes and unrestricted computation at generation time. For each fixed task family and prior, the posterior predictive has expected task-conditional TV risk at most twice that of the best prompt-only generator (see (156)), implying it is an oracle prompt-only generator up to a constant factor. *What matters is uncertainty about the output distribution, not the task label alone.* Distinct tasks cause no generation error when they induce the same output distribution, while multimodality within a task should remain. The controlled experiment in Appendix J illustrates the posterior-predictive behavior: the generated distribution is broadly bimodal under a short, ambiguous context and contracts toward a narrow, essentially unimodal distribution as the context grows.

Sharpness in a rich latent-task class. Fixed latent dimension still allows many weakly separated output distributions. The matching lower bound embeds about $\log n$ weakly distinguishable changes along a single latent coordinate, showing that the logarithmic factor reflects the richness of the task class. This is a worst-case result: the hard family and prior may depend on n. A fixed family may have a faster rate or zero uncertainty.

3.3 THE MINIMAX-OPTIMAL RATE WITH A PRETRAINED DIT

Reaching the known-model rate from data. Section 3.2 grants the oracle generator the task family and prior. The DiT learning rule receives neither and observes no task labels. Yet Theorems 3.1 and 3.2 together give the same worst-case prompt-length rate.

Corollary 3.3 (Task-conditional generation at the sharp prompt rate; informal version of Corollary H.4)**.** *For sufficiently large n, set $\varepsilon = \sqrt{\log n/n}$, use the score class and estimator of Theorem 3.1 with $\varepsilon_{\mathrm{opt}} \lesssim \log n/n$, and let N and M satisfy (10) at this value of ε. For every fixed $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over pretraining and its diffusion augmentations,*

$$\mathbb{E}\,\mathrm{TV}\Big(\widehat{P}^{(M)}_{0|H_n}, P(\cdot \mid H_n, \Theta)\Big) \lesssim \sqrt{\frac{\log n}{n}}. \tag{14}$$

Here the expectation is over a fresh test episode conditional on the completed training run.

Matching learning and sampling accuracy to prompt information. Pretraining episodes estimate the shared prompt-to-distribution rule, model capacity controls its approximation, and diffusion steps control sampling accuracy. Additional demonstrations instead provide more information about the fresh task. Keeping the learning and sampling errors at $\mathcal{O}(\sqrt{\log n/n})$ lets the DiT attain the optimal rate for test-time generation.

A geometric view of in-context generation. Assumption 3.1 defines a Hölder-parameterized family of task distributions. A finite prompt induces a posterior over this family, and the posterior predictive is the corresponding posterior-weighted barycenter of the distributions, which need not itself belong to the family (see Example I.1, where task densities lie on a circle, while the posterior predictive can lie in its interior). In our finite representation, attention realizes this geometry by assigning posterior weights to representative tasks, so that its softmax weights can be viewed as barycentric weights over the latent task family. The same low-dimensional regularity controls statistical complexity: only polynomially many representative tasks are needed to approximate the predictive score, while the cumulative information about the latent task grows only logarithmically with the prompt length. This yields the $\mathcal{O}(\sqrt{\log n/n})$ convergence of the posterior predictive toward the realized task distribution in expected total variation.

4 CONCLUSION AND DISCUSSION

We established finite-sample guarantees for posterior-predictive generation with a pretrained DiT, separating learning and sampling errors from finite-prompt task uncertainty. Under our assumptions, sufficient model capacity, pretraining data, optimization accuracy, and sampling resources yield the minimax-optimal prompt-length rate in task-conditional total variation.

Preserving uncertainty is part of generative accuracy. A finite prompt may leave several task hypotheses plausible, so multimodality may be appropriate and mode collapse may constitute error (cf. Kohl et al., 2018). A sharper output distribution should not automatically count as improved quality.

- As an architectural implication, because the posterior predictive need not contract to any single task distribution, our results motivate soft representations of plausible tasks without mandatory commitment like hard routing or switching (Fedus et al., 2022).
- Our theory naturally raises a future research question: *how to modify the test-time distribution?* Guidance (Ho & Salimans, 2021; Fu et al., 2024), reward-based reranking or best-of-K selection (Amini et al., 2025) can improve sample utility while changing the probability mass assigned to plausible alternatives. How to improve utility while preserving, or explicitly controlling departures from, posterior-predictive calibration remains an open direction for in-context generation.

Resource allocation should follow the source of uncertainty. Once learning and sampling errors are small relative to prompt uncertainty, additional diffusion steps cannot supply the missing task information. If the remaining multimodality reflects ambiguity between tasks, requesting another informative demonstration may therefore be more useful, motivating an extension to active demonstration acquisition (Rainforth et al., 2024). For pretraining, our generalization analysis counts independent episodes: observations within an episode share a latent task and are generally dependent. Total token count alone consequently does not characterize the effective training sample size. The appendix summarizes the corresponding bottlenecks and interventions.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on experiments, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. No other tasks with required disclosure are applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, search for and identify relevant literature, assist with LaTeX formatting, and proofread the manuscript to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arman Adibi, Alireza Jafari, Mohammad Ghavamzadeh, and Hadi Daneshmand. Transformers as In-Context samplers: From Closed-Form diffusion to Estimation-Free sampling, 2026. URL https://arxiv.org/abs/2609.08981.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 45614–45650. Curran Associates, Inc., 2023. doi: 10.52202/075280-1977. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8ed3d610ea4b68e7afb30ea7d01422c6-Paper-Conference.pdf.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin, Heidelberg, 3rd edition, 2006. ISBN 978-3-540-29586-0. doi: 10.1007/3-540-29587-9. URL https://doi.org/10.1007/3-540-29587-9.

Afra Amini, Tim Vieira, Elliott Ash, and Ryan Cotterell. Variational best-of-N alignment. In *International Conference on Learning Representations*, pp. 25717–25737, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/402542c2341e5d2eadc1dd0891275901-Abstract-Conference.html.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions, 2024. URL https://arxiv.org/abs/2409.18804.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211. Curran Associates, Inc., 2023. doi: 10.52202/075280-2498. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b2e63e36c57e153b9015fece2352a9f9-Paper-Conference.pdf.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017. Curran Associates, Inc., 2022. doi: 10.52202/068431-1813. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*,

volume 2024, pp. 36916–36936, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/9edce07129d2ec4c511a8783eb3d81b5-Paper-Conference.pdf.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, June 2018. doi: 10.1109/CVPR.2018.00652. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

L. D. Brown and R. Purves. Measurable selections of extrema. *The Annals of Statistics*, 1(5):902–912, September 1973. doi: 10.1214/aos/1176342510. URL https://doi.org/10.1214/aos/1176342510.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 2023a. URL https://proceedings.mlr.press/v202/chen23q.html.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 2023b. URL https://proceedings.mlr.press/v202/chen23o.html.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023c. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Stochastic control, entropic interpolation and gradient flows on Wasserstein product spaces, 2016. URL https://arxiv.org/abs/1601.04891.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3hD1gzThtY.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A Survey on In-context Learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Guoji Fu and Wee Sun Lee. Approximation and generalization abilities of score-based neural network generative models for Sub-Gaussian distributions. In Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nan Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 77963–78061. Curran Associates, Inc., 2025. doi: 10.52202/085713-2613. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/70702e8cbb4890b4a467b984ae59828a-Paper-Conference.pdf.

Guoji Fu, Taiji Suzuki, Wee Sun Lee, and Atsushi Nitanda. Intrinsic Wasserstein rates for score-based generative models on smooth manifolds, 2026. URL https://arxiv.org/abs/2605.15822.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory, 2024. URL https://arxiv.org/abs/2403.11968.

Max Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, Will Panitch, Fangchen Liu, Hui Li, and Ken Goldberg. ICRT: In-context imitation learning via next-token prediction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5937–5944. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11128272. URL https://doi.org/10.1109/ICRA55743.2025.11128272.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In Sanmi Koyejo, Seyed Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2013. ISBN 9781439840955. doi: 10.1201/b16018. URL https://sites.stat.columbia.edu/gelman/book/.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Trans. Graph.*, 43(4):1–15, July 2024. doi: 10.1145/3658136. URL https://doi.org/10.1145/3658136.

Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, October 1986. doi: 10.1214/aop/1176992362. URL https://doi.org/10.1214/aop/1176992362.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999. doi: 10.1214/ss/1009212519. URL https://doi.org/10.1214/ss/1009212519.

Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 12591–12675, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/21a33eba893ca3890e395651b38810df-Paper-Conference.pdf.

Gyeonghun Kang, Changwoo J. Lee, and Xiang Cheng. Transformers can learn posterior predictive distributions in-context. In *Forty-third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62122.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer New York, NY, 2 edition, 1991. ISBN 978-0-387-97655-6. doi: 10.1007/978-1-4612-0949-2. URL https://doi.org/10.1007/978-1-4612-0949-2.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are Minimax Optimal Nonparametric In-Context Learners. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 106667–106713. Curran Associates, Inc., 2024. doi: 10.52202/079017-3387. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c11daad0a48ea5f3c5c6390c7b060720-Paper-Conference.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. In Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 6965–6975. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 1st edition, 2011. doi: 10.1007/978-3-642-22147-7.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023. URL https://proceedings.mlr.press/v201/lee23a.html.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 51686–51697,

2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e24d9d028e3c7f6f13e6032919ee021e-Abstract-Conference.html.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23l.html.

Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes I: General Theory*, volume 5 of *Stochastic Modelling and Applied Probability*. Springer Berlin, Heidelberg, 2 edition, 2001. ISBN 978-3-540-63929-9. doi: 10.1007/978-3-662-13043-8. URL https://doi.org/10.1007/978-3-662-13043-8.

Tianyi Ma, Tengyao Wang, and Richard J. Samworth. Optimal in-context adaptivity and distributional robustness of transformers, 2025. URL https://arxiv.org/abs/2510.23254.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-Context Learning through the Bayesian Prism. In *International Conference on Learning Representations*, volume 2024, pp. 49789–49843, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d81cd83e7f6748af351485d73f305483-Abstract-Conference.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915. URL https://doi.org/10.1214/23-STS915.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Paul Rosu, Lawrence Carin, and Xiang Cheng. From softmax to score: Transformers can effectively implement in-context denoising steps. In *Advances in Neural Information Processing Systems*, volume 38, pp. 142994–143017. Curran Associates, Inc., 2025. doi: 10.52202/085713-4787. URL https://papers.nips.cc/paper_files/paper/2025/hash/d27af2bebeab8a3e6f3848fc71736235-Abstract-Conference.html.

Matthew Smart, Alberto Bietti, and Anirvan M. Sengupta. In-context denoising with one-layer transformers: Connections between attention and associative memory retrieval. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55950–55971. PMLR, 2025. URL https://proceedings.mlr.press/v267/smart25a.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33416–33447. PMLR, 2023. URL https://proceedings.mlr.press/v202/takakura23a.html.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1648–1656. PMLR, 2024. URL https://proceedings.mlr.press/v238/tang24a.html.

Rong Tang, Lizhen Lin, and Yun Yang. Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *International Conference on Learning Representations*, volume 2025, pp. 66346–66358, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a5e2853be2ff6218a8d87a5ce4b65062-Abstract-Conference.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 1st edition, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, second edition, 2023. ISBN 978-3-031-29038-1. doi: 10.1007/978-3-031-29040-4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. doi: 10.1162/NECO_a_00142. URL https://doi.org/10.1162/NECO_a_00142.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. doi: 10.1017/9781108627771. URL https://doi.org/10.1017/9781108627771.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: A generalization theory for meta-learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.10981.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8542–8562. Curran Associates, Inc., 2023. doi: 10.52202/075280-0374. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b3750390ca8b931fb9ca988647940cb-Paper-Conference.pdf.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4958–4991. PMLR, 2024. URL https://proceedings.mlr.press/v247/wibisono24a.html.

Weimin Wu, Maojiang Su, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. In-context deep learning via transformer models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67670–67718. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25aa.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891. PMLR, 2025. URL https://proceedings.mlr.press/v291/yakovlev25a.html.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, October 2017. doi: 10.1016/j.neunet.2017.07.002. URL https://doi.org/10.1016/j.neunet.2017.07.002.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60134–60178. PMLR, 2024a. URL https://proceedings.mlr.press/v235/zhang24bv.html.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024b. URL http://jmlr.org/papers/v25/23-1042.html.

APPENDIX CONTENTS

A Additional Related Work **19**

B Formal Setup and Conventions **20**

B.1 Probability metrics and conditioning conventions 20

B.2 Latent-task prompt models 20

B.3 Independent-episode pretraining and test-time generation 21

B.4 Hölder density and task-family conventions 22

C Conditional OU Scores and Variance-Balanced DSM **23**

C.1 OU smoothing and conditional score identities 23

C.2 Variance-balanced DSM and exact excess risk 23

D Latent-Task Regularity and Episode-KL Approximation **25**

D.1 Compact preprocessing and score tails 25

D.2 Likelihood and conditional-density regularity 26

D.3 Diffusion features and latent-grid quadrature 27

D.4 Episode-level KL and integrated score approximation 30

E Transformer Construction, Approximation, and Class Complexity **32**

E.1 Score Metric, Transformer Class, and Four-Block Realization 32

E.2 Component Approximation and Stability Bounds 36

E.3 Integrated Transformer Approximation Guarantee 42

E.4 Class Complexity: Parameters, Sensitivity, and Covering 43

F Pretraining and Score Learning from Independent Episodes **47**

F.1 Clipped Labels and the Empirical bDSM Objective 47

F.2 Scaled-Score Covering Metric 48

F.3 Finite-Episode Score-Learning Theorem 49

F.4 Error Decomposition and Tuning Guidance 50

G Reverse Diffusion Sampling and Posterior-Predictive Generation **52**

G.1 Randomized Inverse-Variance Time Grid 52

G.2 Exact Reverse Process on Positive Times 53

G.3 Endpoint Errors: Gaussian Initialization and Early Stopping 53

G.4 Finite-Energy KL Comparison for Reverse Diffusions 55

G.5 Positive-Time Score Dynamics and Discretization 56

G.6 End-to-End Parameter Choices for Target Accuracy 60

G.7 Tail Bounds for the Selected Radii 61

G.8 Proof of the Posterior-Predictive Generation Guarantee 62

G.9 Scalar Sample-Only Complexity Specialization 63

H Sharp Finite-Prompt Task Uncertainty **65**

H.1 Finite-Prompt Theorem and Oracle Risk . . . 65
H.2 Posterior Averaging and Oracle Comparison . . . 66
H.3 Upper Bound via Predictive Information . . . 66
H.4 Lower Bound via a Smooth Latent Assouad Family . . . 68
H.5 Task-Conditional Consequence and Minimaxity . . . 70

I Examples of Smooth Latent Task Families **72**
I.1 Direct Sample-Only and Input–Output Families . . . 72
I.2 General Construction Templates . . . 72

J Numerical Illustrations **74**
J.1 Posterior-Predictive Generation in a Synthetic Image Task . . . 74
J.2 Oracle Diagnostics for Predictive and Task-Conditional Error . . . 77

K Qualitative Diagnostics and Resource Allocation **80**

A Additional Related Work

Section 1.1 reviews the closely related work on ICL and Transformer-based score estimation. Here, we highlight the two diffusion-specific ingredients needed for our end-to-end analysis: learning a prompt-conditioned score from finitely many independent pretraining episodes, and separating posterior-predictive approximation error from the uncertainty that remains with a finite prompt.

Diffusion-model theory commonly separates score estimation from the stability and discretization analysis of the reverse-time process. Score-to-sampling results translate suitable L^2 score errors along the forward marginals into distributional sampling guarantees under various moment and regularity conditions (Chen et al., 2023c; Benton et al., 2024; Lee et al., 2023; Chen et al., 2023a; Li et al., 2024). Complementary minimax analyses ask whether empirical score matching followed by reverse diffusion estimates a fixed data distribution at an optimal statistical rate (Oko et al., 2023; Wibisono et al., 2024; Chen et al., 2023b; Zhang et al., 2024a; Fu & Lee, 2025). Related work also exploits structural or manifold assumptions to obtain generalization, convergence, or intrinsic-dimensional guarantees (Yakovlev & Puchkin, 2025; Tang & Yang, 2024; Azangulov et al., 2024; Fu et al., 2026). Our analysis uses the same broad score-to-distribution interface, but its input is the integrated prompt-conditioned risk of a score estimator learned from independent pretraining episodes. The randomly shifted inverse-variance time grid converts this integrated-in-time control into an expected total-variation guarantee for a finite-step reverse sampler, without requiring a score-error bound that holds uniformly over time.

Conditional diffusion theory typically studies families of distributions indexed by an observed covariate (Fu et al., 2024; Tang et al., 2025), including approximation and estimation guarantees for conditional DiTs (Hu et al., 2025). In our episodic model, by contrast, the conditioning object is a random prompt H_n generated from an unobserved task. The prompt provides evidence about the latent task but does not directly index a single realized-task distribution. Consequently, the posterior-predictive distribution $P(\cdot \mid H_n)$ is the score-learning target, while its discrepancy from the realized-task distribution $P(\cdot \mid H_n, \Theta)$ is analyzed separately as finite-prompt task uncertainty. Under the stated nondegeneracy conditions, our minimax prompt-length lower bound quantifies the unavoidable finite-prompt contribution to task-conditional total-variation risk.

B Formal Setup and Conventions

Section 2 defines the episodic problem, the sample-only and input–output prompt models, evaluation distributions, and error decomposition. This appendix fixes the formal conventions used throughout the proofs. It proceeds from probability metrics and regular conditioning to the two latent-task prompt models, then specifies independent-episode pretraining and test-time generation, and finally records the Hölder classes used in the regularity assumptions.

B.1 Probability metrics and conditioning conventions

All spaces are standard Borel unless stated otherwise. For a space E, $\mathscr{B}(E)$ and $\mathcal{P}(E)$ denote its Borel σ-field and Borel probability measures, respectively. We equip $\mathcal{P}(E)$ with the canonical evaluation σ-field generated by the maps $\mu \mapsto \mu(A)$ for $A \in \mathscr{B}(E)$; this makes $\mathcal{P}(E)$ standard Borel and agrees with the weak Borel σ-field under any compatible Polish realization of E. For a measurable map G, $G_{\#}P$ denotes the pushforward of P. For $P, Q \in \mathcal{P}(E)$, we use

$$\mathsf{TV}(P,Q) = \sup_{A \in \mathscr{B}(E)} |P(A) - Q(A)|, \qquad \mathrm{KL}(P\|Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, \tag{15}$$

with $\mathrm{KL}(P\|Q) = +\infty$ when $P \not\ll Q$. For densities p, q relative to a common measure μ, $\mathsf{TV}(P,Q) = \frac{1}{2}\int |p-q|\,\mathrm{d}\mu$. We also write $\mathsf{TV}(p,q)$ and $\mathrm{KL}(p\|q)$ for the TV distance and KL divergence between their induced measures; the density 1 denotes the uniform law when the domain is a unit-volume cube. We write $\mathcal{N}(m,\Sigma)$ for a Gaussian distribution and, for positive-definite Σ, $\mathcal{N}(u; m, \Sigma)$ for its density evaluated at u. For nonnegative quantities a and b, we write $a \lesssim b$ when $a \le cb$ for a positive constant c that may depend only on the fixed dimensions, regularity bounds, and auxiliary constructions stated in the result. The hidden constant is independent of the varying quantities n, N, M, ε, and δ, as well as the particular model, prior, and observed data. We define $\gtrsim$ and $\asymp$ analogously; subscripts on these symbols record any additional allowed dependencies. Named constants remain explicit when they are fixed and reused in a construction or in later parameter choices.

Upper-case P, Q denote distributions or kernels and lower-case p, q their densities; for task densities, p_θ denotes the sample-only density, f_θ the joint (X, Z)-density, and g_θ either one according to the prompt model. Euclidean densities are relative to Lebesgue measure unless stated otherwise. Lower-case h and $\boldsymbol{c}_n = (c_1, \dots, c_n)$ denote realizations of H_n and C_n, with $c_i = z_i$ in the sample-only prompt model and $c_i = (x_i, z_i)$ in the input–output prompt model. We use $\mathbb{1}\{A\}$ for an indicator and $\mathbf{1}_d$ for the all-ones vector. Expectations average the stated randomness; test-risk expectations have the conditional meaning specified in Section B.3.

Fix $d_z \ge 1$, $\mathcal{Z} = [0,1]^{d_z}$, an integer prompt length $n \ge 1$, and a prompt H_n taking values in a standard Borel space $\mathcal{H}_n$. When comparing a law on $\mathcal{Z}$ with an OU-smoothed law on $\mathbb{R}^{d_z}$, we identify the former with its pushforward under the inclusion $\mathcal{Z} \hookrightarrow \mathbb{R}^{d_z}$. Write $\mathbb{P}^{(n)}_{\mathrm{epi}} := \mathsf{Law}(H_n, Z_\star)$ for the distribution of observed episodes. Standard-Borel regular conditioning permits us to choose a probability kernel

$$h \mapsto P(\cdot \,|\, h) := \mathsf{Law}(Z_\star \,|\, H_n = h). \tag{16}$$

Because $Z_\star \in \mathcal{Z}$ almost surely, we choose this version to satisfy $P(\mathcal{Z} \,|\, h) = 1$ for every h, replacing it by a fixed point mass on the measurable exceptional set if necessary. For $A \in \mathscr{B}(\mathcal{Z})$, $P(A \,|\, H_n)$ denotes the measurable map $h \mapsto P(A \,|\, h)$ evaluated at H_n, so $P(\cdot \,|\, H_n)$ is the corresponding random probability measure. Density, smoothness, positivity, and within-episode conditional-independence conditions are imposed only by the task-family assumptions used for a given result.

B.2 Latent-task prompt models

We formalize the latent-task construction for the sample-only and input–output prompt models in Section 2.1. Fix a standard Borel task space $\mathcal{T}$, a task prior $\Pi \in \mathcal{P}(\mathcal{T})$, and Borel kernels $\theta \mapsto P_{H_n \,|\, \theta}$ and $(h, \theta) \mapsto P_{Z_\star \,|\, H_n, \theta}(\cdot \,|\, h)$. The prior Π is also the distribution over tasks in the pretraining data-generating process. The induced distribution of observed episodes is

$$\mathbb{P}^{(n)}_{\mathrm{epi}}(\mathrm{d}h, \mathrm{d}z_\star) = \int_{\mathcal{T}} P_{H_n \,|\, \theta}(\mathrm{d}h) P_{Z_\star \,|\, H_n, \theta}(\mathrm{d}z_\star \,|\, h)\, \Pi(\mathrm{d}\theta). \tag{17}$$

Only $(H_n, Z_\star)$ is observed, while Θ is retained for task-conditional evaluation. The standard-Borel assumptions give a regular conditional distribution $\Pi(\mathrm{d}\theta \mid h)$ of the task given the prompt. Define the task-conditional output kernel by

$$P(\cdot \mid h, \theta) := P_{Z_\star \mid H_n, \theta}(\cdot \mid h). \tag{18}$$

By the tower property, we may choose the kernel in (16) so that the posterior-mixture identity (2) holds as an identity of probability measures for $\mathsf{Law}(H_n)$-almost every h. A countable determining algebra and the monotone-class theorem give a single common exceptional null set for all Borel subsets of $\mathcal{Z}$.

B.2.1 Sample-only in-context generation

Without inputs, let $\theta \mapsto P_\theta$ be a Borel probability kernel on $\mathcal{Z}$ and set

$$Z_1, \dots, Z_n, Z_\star \mid \Theta \overset{\text{i.i.d.}}{\sim} P_\Theta, \qquad H_n = Z_{1:n}. \tag{19}$$

The task-conditional output distribution is $P(\cdot \mid h, \theta) = P_\theta$ for $\mathsf{Law}(H_n, \Theta)$-almost every (h, θ). If $P_\theta(\mathrm{d}z) = p_\theta(z)\,\mathrm{d}z$ with a jointly Borel density, then the posterior-predictive density is $p(z \mid h) = \int_{\mathcal{T}} p_\theta(z)\,\Pi(\mathrm{d}\theta \mid h)$ for $\mathsf{Law}(H_n)$-almost every h. The episodic formulation itself does not require densities.

B.2.2 Input–output in-context generation

Let $\mathcal{X}$ be standard Borel and let $\theta \mapsto P_\theta^{X,Z}$ be a Borel probability kernel on $\mathcal{X} \times \mathcal{Z}$. Conditional on the task, set

$$(X_1, Z_1), \dots, (X_n, Z_n), (X_\star, Z_\star) \mid \Theta \overset{\text{i.i.d.}}{\sim} P_\Theta^{X,Z}, \qquad H_n = \big(((X_i, Z_i))_{i=1}^n, X_\star\big). \tag{20}$$

Choose a jointly measurable regular conditional kernel $(\theta, x) \mapsto P_\theta^{Z|X}(\cdot \mid x)$. For $\mathsf{Law}(H_n, \Theta)$-almost every (h, θ) with $h = (\boldsymbol{c}_n, x_\star)$, the task-conditional output distribution is

$$P(\cdot \mid h, \theta) = P_\theta^{Z|X}(\cdot \mid x_\star). \tag{21}$$

The posterior-mixture identity in (2) therefore conditions the task posterior on both the context and the observed query. The query affects both the task-conditional output distribution and the task posterior when the input marginal is task-dependent.

B.3 Independent-episode pretraining and test-time generation

Let the training sample contain $N \geq 1$ independent copies of the observed episode:

$$\mathcal{D}_N^{\text{epi}} := \big((H_n^{(e)}, Z_\star^{(e)})\big)_{e=1}^N, \qquad (H_n^{(e)}, Z_\star^{(e)}) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{epi}}^{(n)}. \tag{22}$$

For $N = 0$, define $\mathcal{D}_0^{\text{epi}}$ as the deterministic empty sample. Any declared sample splits are mutually independent, and a fresh test episode is independent of the training sample and all training randomness. A possibly randomized learning procedure takes $\mathcal{D}_N^{\text{epi}}$ and an independent auxiliary seed as inputs and is jointly Borel in the training sample, seed, and prompt. Every completed training run therefore defines a probability kernel

$$h \longmapsto \widehat{P}(\cdot \mid h) \in \mathcal{P}(\mathcal{Z}). \tag{23}$$

The kernel $\widehat{P}(\cdot \mid h)$ is the output distribution induced by the test-time sampler and therefore already integrates over generation-time randomness.

Unless stated otherwise, $\mathbb{E}$ in a test-risk bound is conditional on the completed training run and averages an independent fresh episode. Under the latent-task model, this expectation includes Θ and H_n, including the fresh query $X_\star$ in the input–output model. An outer expectation may additionally average the training sample and training randomness. We use the three expected TV distances in (3) directly: predictive error, task-conditional error, and task uncertainty. The last depends only on the episodic model and prompt length, not on training or the generator.

B.4 HÖLDER DENSITY AND TASK-FAMILY CONVENTIONS

The following conventions define the Hölder classes used in Section 3. Both the sampling and task-uncertainty analyses use these conventions.

Fix $\beta > 0$ and an integer $d \geq 1$. Write $m_\beta := \lceil \beta \rceil - 1$ and $\vartheta_\beta := \beta - m_\beta \in (0, 1]$. For a function g on $[0,1]^d$ whose derivatives through order m_β exist in the interior and extend continuously to the cube, define

$$\|g\|_{C^\beta([0,1]^d)} := \max_{|a| \leq m_\beta} \|D^a g\|_\infty + \max_{|a|=m_\beta} \sup_{u \neq v} \frac{|D^a g(u) - D^a g(v)|}{\|u - v\|_2^{\vartheta_\beta}}, \tag{24}$$

where a ranges over multi-indices. With this convention, integer smoothness β means $C^{\beta-1,1}$ regularity. Since $[0,1]^d$ has unit volume and $\|p\|_{C^\beta} \geq \|p\|_\infty \geq 1$ for every density p, the class below is empty when $0 < L_\beta < 1$. We therefore take $L_\beta \geq 1$ and let

$$\mathcal{F}_\beta^d(L_\beta) := \Big\{ p : [0,1]^d \to [0,\infty) : \int_{[0,1]^d} p(u)\,\mathrm{d}u = 1, \|p\|_{C^\beta([0,1]^d)} \leq L_\beta \Big\}. \tag{25}$$

For $\underline{p} \in (0, 1]$, define the positive-density subclass

$$\mathcal{F}_{\beta,+}^d(L_\beta, \underline{p}) := \big\{ p \in \mathcal{F}_\beta^d(L_\beta) : \inf_{u \in [0,1]^d} p(u) \geq \underline{p} \big\}.$$

Because the Hölder norm contains $\|p\|_\infty$, every member of the full class automatically satisfies $p(u) \leq L_\beta$ on the cube. At $L_\beta = 1$, the full class $\mathcal{F}_\beta^d(L_\beta)$ is the singleton containing the uniform density; when $L_\beta > 1$, that full class may contain densities that vanish and may have task-dependent supports.

The density classes above also support the density-estimation comparison. Corollaries G.2 and G.6 state their spatial regularity conditions locally; they do not require the positivity or latent regularity imposed by Assumption 3.1.

C Conditional OU Scores and Variance-Balanced DSM

This appendix supplies the diffusion identities that connect the statistical and sampling analyses. We first derive the conditional OU score and its positive-time bounds, and then show that inverse-variance time sampling makes the excess variance-balanced DSM risk exactly equal to the integrated score error. The reverse processes and sampling guarantees are deferred to Section G.

As in Section 2.3, set $Z_0 \equiv Z_\star$ and $Z_0^{(e)} \equiv Z_\star^{(e)}$; the zero subscript denotes diffusion time. Write $P_{0\,|\,h} := P(\cdot \,|\, h)$ for the same posterior-predictive distribution.

C.1 OU smoothing and conditional score identities

Fix $T > 0$ and a prompt realization h. The unit-rate OU process in (4) has the following fixed-time representation (Karatzas & Shreve, 1991; Song et al., 2021b). For each $t \geq 0$,

$$Z_t \overset{\mathrm{d}}{=} \alpha_t Z_0 + \sigma_t \xi, \qquad \alpha_t = e^{-t}, \quad \sigma_t^2 = 1 - e^{-2t}, \tag{26}$$

where $\xi \sim \mathcal{N}(0, I_{d_z})$ is independent of Z_0. For any initial distribution P, write P_t for the distribution of $\alpha_t Z + \sigma_t \xi$ with independent $Z \sim P$ and $\xi \sim \mathcal{N}(0, I_{d_z})$. In particular, $P_{t\,|\,h} := \mathsf{Law}(Z_t \,|\, H_n = h)$ is the OU evolution of $P_{0\,|\,h}$. For every $t > 0$, this distribution has an everywhere positive smooth density $p_t(\cdot \,|\, h)$, even if $P_{0\,|\,h}$ is singular. Its conditional score function is

$$s^\star(z, t \,|\, h) := \nabla_z \log p_t(z \,|\, h). \tag{27}$$

The Gaussian-convolution formula and positivity provide a jointly Borel version of $(z, t, h) \mapsto s^\star(z, t \,|\, h)$ on $\mathbb{R}^{d_z} \times (0, T] \times \mathcal{H}_n$; this version is used throughout. Gaussian differentiation and Bayes' formula further give

$$\begin{aligned} s^\star(z, t \,|\, h) &= -\frac{z - \alpha_t \mathbb{E}[Z_0 \,|\, Z_t = z, H_n = h]}{\sigma_t^2}, \\ \nabla_z s^\star(z, t \,|\, h) &= -\frac{I_{d_z}}{\sigma_t^2} + \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov}(Z_0 \,|\, Z_t = z, H_n = h). \end{aligned} \tag{28}$$

Since $P_{0\,|\,h}$ is supported on $[0,1]^{d_z}$, for every $\varepsilon \in (0, T)$ and $t \in [\varepsilon, T]$,

$$\|s^\star(z, t \,|\, h)\|_2 \leq \frac{\|z\|_2 + \sqrt{d_z}}{\sigma_\varepsilon^2}, \qquad \|\nabla_z s^\star(z, t \,|\, h)\|_{\mathrm{op}} \leq \frac{1}{\sigma_\varepsilon^2} + \frac{d_z}{\sigma_\varepsilon^4}. \tag{29}$$

The denoising identity also yields $\mathbb{E}[\|s^\star(Z_t, t \,|\, H_n)\|_2^2 \,|\, H_n] \leq d_z / \sigma_t^2$. These bounds justify positive-time reversal in Section G.2 and provide regularity estimates for score approximation. The evaluations at the random prompt are denoted by $P_{t\,|\,H_n}$ and $s^\star(\cdot, t \,|\, H_n)$.

C.2 Variance-balanced DSM and exact excess risk

Variance-balanced DSM trains the scaled-score predictor using the label $-\xi_{\mathrm{dsm}}$. The clipped empirical objective is defined in Section F.1; here we establish the unclipped population identity.

Fix $0 < t_0 < T$. For a measurable score function $s(z, t, h)$, recall the integrated squared score error introduced in Section 2.4:

$$\mathcal{E}_{\mathrm{score}}(s) := \mathbb{E}_{H_n}\left[\int_{t_0}^{T} \left\| s(\cdot, t, H_n) - s^\star(\cdot, t \,|\, H_n) \right\|_{L^2(P_{t\,|\,H_n})}^2 \mathrm{d}t\right]. \tag{30}$$

$\mathcal{E}_{\mathrm{score}}$ is not directly observable because the true score $s^\star$ is unknown.

The standard denoising identity (Vincent, 2011) states that, for every fixed $t > 0$,

$$\mathbb{E}\left[-\frac{\xi_{\mathrm{dsm}}}{\sigma_t} \,|\, Z_t, t, H_n\right] = s^\star(Z_t, t \,|\, H_n), \qquad Z_t = \alpha_t Z_0 + \sigma_t \xi_{\mathrm{dsm}}.$$

Under uniform time sampling, direct score regression uses the label $-\xi_{\mathrm{dsm}}/\sigma_t$, whose second moment grows as $t \downarrow 0$. Predicting $\sigma_t s$ instead uses the unit-scale label $-\xi_{\mathrm{dsm}}$, but yields a σ_t^2-weighted score

error. Converting that error to $\mathcal{E}_{\mathrm{score}}$ by a worst-case bound costs a factor $\sigma_{t_0}^{-2}$. Inverse-variance time sampling compensates for this weighting, so the balanced objective has excess risk exactly $\mathcal{E}_{\mathrm{score}}$.

Draw $\xi_{\mathrm{dsm}} \sim \mathcal{N}(0, I_{d_z})$ independently of an episode. For any score function s, define its scaled representation by

$$\widetilde{s}(z,t,h) := \sigma_t s(z,t,h), \qquad s(z,t,h) = \frac{\widetilde{s}(z,t,h)}{\sigma_t}. \tag{31}$$

Thus the same parameterized model predicts $\widetilde{s}$ in the scaled-score training loss and returns s for use in the reverse sampler.

Define

$$A_{t_0,T} := \int_{t_0}^{T} \frac{\mathrm{d}t}{\sigma_t^2} = \frac{1}{2}\log\left(\frac{e^{2T}-1}{e^{2t_0}-1}\right), \qquad \varpi_{t_0,T}(t) := \frac{1}{A_{t_0,T}\sigma_t^2}. \tag{32}$$

For the variance-balanced DSM analysis, draw τ with density $\varpi_{t_0,T}$, independently of the episode and $\xi_{\mathrm{dsm}} \sim \mathcal{N}(0, I_{d_z})$, and form Z_τ by

$$Z_\tau := \alpha_\tau Z_0 + \sigma_\tau \xi_{\mathrm{dsm}}. \tag{33}$$

The positive cutoff remains essential: because $\sigma_t^2 \sim 2t$ as $t \downarrow 0$, $A_{0,T} = \infty$, so the balanced time density cannot be normalized at $t_0 = 0$.

As in the main paper, bDSM denotes variance-balanced denoising score matching. Parameterize its population objective by the score function s:

$$\mathcal{R}_{\mathrm{bDSM}}(s) := A_{t_0,T}\,\mathbb{E}\left[\|\widetilde{s}(Z_\tau,\tau,H_n) + \xi_{\mathrm{dsm}}\|_2^2\right]. \tag{34}$$

Proposition C.1 (Exact variance-balanced DSM excess-risk identity)**.** *Define* $\widetilde{s}^\star(z,t\,|\,h) := \sigma_t s^\star(z,t\,|\,h)$. *Then*

$$\mathbb{E}[-\xi_{\mathrm{dsm}} \,|\, Z_\tau, \tau, H_n] = \widetilde{s}^\star(Z_\tau, \tau \,|\, H_n)$$

almost surely, and every jointly Borel score s satisfying $\mathcal{R}_{\mathrm{bDSM}}(s) < \infty$ obeys

$$\mathcal{R}_{\mathrm{bDSM}}(s) - \mathcal{R}_{\mathrm{bDSM}}(s^\star) = \mathcal{E}_{\mathrm{score}}(s). \tag{35}$$

Proof of Proposition C.1. The standard denoising identity (Vincent, 2011), multiplied by σ_t, gives the conditional mean at each fixed $t > 0$. Joint measurability and disintegration under $\varpi_{t_0,T}$ give the asserted identity at τ. Conditional Jensen and $\xi_{\mathrm{dsm}} \in L^2$ imply $\widetilde{s}^\star \in L^2$ and finite risk. For every finite-risk s, the squared-loss Pythagorean identity therefore yields

$$\begin{aligned}\mathcal{R}_{\mathrm{bDSM}}(s) - \mathcal{R}_{\mathrm{bDSM}}(s^\star) &= A_{t_0,T}\mathbb{E}\|\widetilde{s}(Z_\tau,\tau,H_n) - \widetilde{s}^\star(Z_\tau,\tau \mid H_n)\|_2^2 \\ &= \mathcal{E}_{\mathrm{score}}(s),\end{aligned}$$

where the second equality uses $A_{t_0,T}\varpi_{t_0,T}(t)\sigma_t^2 = 1$. □

The identity controls the integrated score error without the factor $\sigma_{t_0}^{-2}$ incurred by a worst-case conversion from uniform-time noise prediction. The inverse-variance time density compensates for the score scaling by σ_t; it is not claimed to minimize variance among all time-sampling schemes.

D LATENT-TASK REGULARITY AND EPISODE-KL APPROXIMATION

The score representation in Section C averages over a continuous latent task space. This appendix prepares its finite approximation in four steps: compact preprocessing and tail control; likelihood and conditional-density regularity; normalized diffusion features and latent-grid quadrature; and an episode-level KL comparison that yields integrated score error under the true prompt law. Section E then approximates the resulting finite-mixture score with neural component maps and an exact four-block attention aggregation. Unless displayed explicitly, dependence on the fixed diffusion window (t_0, T) and deterministic state radius $R_{\mathrm{state},n}$ is suppressed.

D.1 COMPACT PREPROCESSING AND SCORE TAILS

The construction uses the sample-only and input–output prompt models on Euclidean cubes from Section B.2. Recall that $\boldsymbol{c}_n = (c_1, \dots, c_n)$ denotes a realized context. We first define bounded input features and establish the tail bounds that convert uniform approximation on a compact set into integrated score-error bounds.

Define the query space

$$\mathcal{X}_\star := \begin{cases} \{0\}, & \text{sample-only}, \\ [0,1]^{d_x}, & \text{input–output}. \end{cases}$$

Write $h = \boldsymbol{c}_n$ in the sample-only prompt model and $h = (\boldsymbol{c}_n, x_\star)$ in the input–output prompt model. To use the same formulas in both cases, set $x_\star = 0$ and $X_\star \equiv 0$ in the sample-only prompt model. Under either the sample-only or the input–output prompt model, the query–state–time feature defined below has dimension $d_{\mathrm{stat}} + 1$. For $R > 0$, define radial clipping $\mathrm{clip}_R : \mathbb{R}^{d_z} \to \overline{\mathrm{Ball}}(0, R)$ by

$$\mathrm{clip}_R(v) := \begin{cases} v, & \|v\|_2 \le R, \\ Rv/\|v\|_2, & \|v\|_2 > R, \end{cases} \tag{36}$$

For each n, fix a deterministic state-clipping radius $R_{\mathrm{state},n} > \sqrt{d_z}$. Define the fixed normalized input feature

$$\mathsf{u}_n(z,t,h) := \begin{cases} \left(\mathrm{clip}_{R_{\mathrm{state},n}}(z)/R_{\mathrm{state},n}, 2\dfrac{t-t_0}{T-t_0} - 1 \right), & \text{sample-only}, \\ \left(2x_\star - \mathbf{1}_{d_x}, \mathrm{clip}_{R_{\mathrm{state},n}}(z)/R_{\mathrm{state},n}, 2\dfrac{t-t_0}{T-t_0} - 1 \right), & \text{input–output}. \end{cases} \tag{37}$$

Here $x_\star$ is the query component of h in the input–output prompt model. The fixed map u_n takes values in $[-1,1]^{d_{\mathrm{stat}}+1}$; in particular, no additional query encoder is required. The state radius enters only through the approximation domain, tail bounds, and regularity constants. Recall from Proposition C.1 that $\widetilde{s}^\star(z,t \mid h) = \sigma_t s^\star(z,t \mid h)$. The regression identity

$$\mathcal{E}_{\mathrm{score}}(s) = A_{t_0,T}\mathbb{E}\big[\|\widetilde{s}(Z_\tau, \tau, H_n) - \widetilde{s}^\star(Z_\tau, \tau \mid H_n)\|_2^2\big] \tag{38}$$

shows why the learning analysis uses the scaled score rather than the unscaled score, which becomes singular as $t \downarrow 0$.

For the smooth latent model considered below, every context array and query lies in the specified domain. The positive-likelihood posterior kernel derived in Section D.3.1 is defined for every such prompt and agrees almost surely with the conditional distribution chosen in Section B.1. Set

$$R_{\mathrm{bal},n} := \frac{R_{\mathrm{state},n} + \sqrt{d_z}}{\sigma_{t_0}}, \qquad \mathcal{G}_n := \mathcal{H}_n \times \overline{\mathrm{Ball}}(0, R_{\mathrm{state},n}) \times [t_0, T],$$

and let $\delta_{\mathrm{tail},n} := \Pr((H_n, Z_\tau, \tau) \notin \mathcal{G}_n)$.

Lemma D.1 (Scaled-score bound and truncation error)**.** *For the cube-supported posterior-predictive distributions in Section B.1, $\|\widetilde{s}^\star(z,t \mid h)\|_2 \le R_{\mathrm{bal},n}$ on $\mathcal{G}_n$, and*

$$\delta_{\mathrm{tail},n} \le \Pr\Big(\|\xi\|_2 > R_{\mathrm{state},n} - \sqrt{d_z}\Big) < 1, \qquad \xi \sim \mathcal{N}(0, I_{d_z}). \tag{39}$$

The squared-score contribution outside the truncation region is finite and satisfies

$$\varepsilon^2_{\mathrm{tail},n} := A_{t_0,T}\mathbb{E}\big[\|\widetilde{s}^\star(Z_\tau, \tau \mid H_n)\|_2^2 \, \mathbb{1}\{(H_n, Z_\tau, \tau) \notin \mathcal{G}_n\}\big] \tag{40}$$

$$\le \frac{A_{t_0,T}}{\sigma_{t_0}^2}\mathbb{E}\Big[(\|\xi\|_2 + 2\sqrt{d_z})^2 \, \mathbb{1}\{\|\xi\|_2 > R_{\mathrm{state},n} - \sqrt{d_z}\}\Big]. \tag{41}$$

Proof. By (28) and cube support,

$$\|\widetilde{s}^\star(z,t \mid h)\|_2 \le \frac{\|z\|_2 + \sqrt{d_z}}{\sigma_{t_0}}, \qquad t \in [t_0, T].$$

This proves the uniform bound on $\mathcal{G}_n$. Since $\|Z_0\|_2 \le \sqrt{d_z}$ and $\alpha_\tau, \sigma_\tau \le 1$, forward noising gives

$$\{\|Z_\tau\|_2 > R_{\mathrm{state},n}\} \subseteq \{\|\xi\|_2 > R_{\mathrm{state},n} - \sqrt{d_z}\}, \qquad \|Z_\tau\|_2 + \sqrt{d_z} \le \|\xi\|_2 + 2\sqrt{d_z}.$$

The Gaussian probability proves (39); its strict bound follows from $R_{\mathrm{state},n} > \sqrt{d_z}$. Combining the score bound with these inequalities and applying Tonelli proves (41). Joint Borel measurability follows from the chosen score version and the fixed clipping set, and Gaussian second moments give finiteness. □

D.2 Likelihood and conditional-density regularity

Under Assumption 3.1, write $g_\theta = p_\theta$ for sample-only prompts and $g_\theta = f_\theta$ for input–output prompts. The learner knows only the common dimensions, regularity bounds, and deterministic grid scheme; the unknown task family and prior enter the analysis through the existence of an approximating parameter vector. The prior may be any Borel probability measure, and the regularity and positivity bounds hold at every latent coordinate, including grid points with zero prior mass. The latent coordinates determine fixed grid indices but are neither observed nor supplied as numerical model inputs.

The next lemma derives likelihood and conditional-density bounds from the task assumptions. The episode-KL bound in Proposition D.4 uses these bounds to control latent discretization in expected integrated score error.

Set $\beta' := \beta \wedge 1$ for the spatial modulus of the likelihood and conditional-density components. In the input–output prompt model, define

$$\lambda_\theta(x) := \int_{[0,1]^{d_z}} g_\theta(x,y)\,\mathrm{d}y, \qquad q_\theta(z_0 \mid x) := \frac{g_\theta(x,z_0)}{\lambda_\theta(x)},$$

and set $\ell_\theta := \log g_\theta$ and $\ell^{\mathrm{qry}}_\theta := \log \lambda_\theta$. In the sample-only prompt model, set

$$q_\theta(z_0 \mid 0) := g_\theta(z_0), \qquad \lambda_\theta(0) := 1, \qquad \ell_\theta := \log g_\theta, \qquad \ell^{\mathrm{qry}}_\theta(0) := 0.$$

Lemma D.2 (Likelihood and conditional-density regularity)**.** *Under Assumption 3.1, the preceding maps are jointly Borel, and $q_\theta(\cdot \mid x)$ is a probability density for every (θ, x). There are fixed constants $B_{\ell,\mathrm{ctx}} \ge 1$, $B_{\ell,\mathrm{qry}} \ge 0$, $0 < \underline{q}_{\mathrm{out}} \le \overline{q}_{\mathrm{out}}$, and finite Hölder constants such that, for every $\theta, \theta' \in \mathcal{T}$, every $c, c' \in [0,1]^{d_{\mathrm{stat}}}$, every $x, x' \in \mathcal{X}_\star$, and every $z_0 \in [0,1]^{d_z}$,*

$$|\ell_\theta(c)| \le B_{\ell,\mathrm{ctx}}, \qquad |\ell^{\mathrm{qry}}_\theta(x)| \le B_{\ell,\mathrm{qry}}, \qquad \underline{q}_{\mathrm{out}} \le q_\theta(z_0 \mid x) \le \overline{q}_{\mathrm{out}}, \tag{42}$$

and

$$|\ell_\theta(c) - \ell_{\theta'}(c')| \le H_{\ell,\theta}\|\theta - \theta'\|_2^{\beta_\theta} + H_{\ell,c}\|c - c'\|_2^{\beta'}, \tag{43}$$

$$|\ell^{\mathrm{qry}}_\theta(x) - \ell^{\mathrm{qry}}_{\theta'}(x')| \le H_{\mathrm{qry},\theta}\|\theta - \theta'\|_2^{\beta_\theta} + H_{\mathrm{qry},x}\|x - x'\|_2^{\beta'}, \tag{44}$$

$$\sup_{z_0 \in [0,1]^{d_z}} |q_\theta(z_0 \mid x) - q_{\theta'}(z_0 \mid x')| \le H_{q,\theta}\|\theta - \theta'\|_2^{\beta_\theta} + H_{q,x}\|x - x'\|_2^{\beta'}. \tag{45}$$

In the sample-only prompt model,

$$(B_{\ell,\mathrm{qry}}, H_{\mathrm{qry},\theta}, H_{\mathrm{qry},x}, H_{q,x}) = (0,0,0,0).$$

The constants depend only on the parameters in Assumption 3.1; explicit choices for the sample-only and input–output prompt models are given in (46) and (47).

We prove Lemma D.2 under Assumption 3.1 and give the constants used in the construction. Recall that $\beta' = \beta \wedge 1$.

Proof of Lemma D.2. The defining Hölder bound for $\beta \le 1$, and the mean-value theorem with the Hölder norm's gradient bound for $\beta > 1$, give a constant

$$H_{\mathrm{dens}} := C_{\beta, d_{\mathrm{stat}}} L_\beta$$

such that every g_θ in Assumption 3.1 satisfies

$$|g_\theta(u) - g_\theta(u')| \le H_{\mathrm{dens}} \|u - u'\|_2^{\beta'}.$$

Put

$$B_{\log} := 1 \vee |\log \underline{p}| \vee \log L_\beta, \qquad K_q := \underline{p}^{-1} + L_\beta \underline{p}^{-2}.$$

In the input–output prompt model, one may choose

$$\begin{aligned} B_{\ell,\mathrm{ctx}} = B_{\ell,\mathrm{qry}} &:= B_{\log}, & (\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}) &:= \left(\frac{\underline{p}}{L_\beta}, \frac{L_\beta}{\underline{p}} \right), \\ H_{\ell,\theta} = H_{\mathrm{qry},\theta} &:= \frac{H_\theta}{\underline{p}}, & H_{\ell,c} = H_{\mathrm{qry},x} &:= \frac{H_{\mathrm{dens}}}{\underline{p}}, \\ H_{q,\theta} &:= K_q H_\theta, & H_{q,x} &:= K_q H_{\mathrm{dens}}. \end{aligned} \tag{46}$$

In the sample-only prompt model, one may instead choose

$$\begin{gathered} B_{\ell,\mathrm{ctx}} := B_{\log}, \qquad (\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}) := (\underline{p}, L_\beta), \qquad H_{\ell,\theta} := H_\theta / \underline{p}, \qquad H_{\ell,c} := H_{\mathrm{dens}} / \underline{p}, \\ H_{q,\theta} := H_\theta, \qquad (B_{\ell,\mathrm{qry}}, H_{\mathrm{qry},\theta}, H_{\mathrm{qry},x}, H_{q,x}) := (0,0,0,0). \end{gathered} \tag{47}$$

We verify the input–output case. Because the output cube has unit volume,

$$\underline{p} \le \lambda_\theta(x) \le L_\beta$$

and

$$|\lambda_\theta(x) - \lambda_{\theta'}(x')| \le H_\theta \|\theta - \theta'\|_2^{\beta_\theta} + H_{\mathrm{dens}} \|x - x'\|_2^{\beta'}.$$

The mean-value theorem for the logarithm on $[\underline{p}, L_\beta]$ proves the asserted bounds for ℓ_θ and $\ell_\theta^{\mathrm{qry}}$. The bounds on g_θ and λ_θ give the displayed values of $\underline{q}_{\mathrm{out}}$ and $\overline{q}_{\mathrm{out}}$. Moreover, for $u = (x, z_0)$ and $u' = (x', z_0)$,

$$\begin{aligned} |q_\theta(z_0 \mid x) - q_{\theta'}(z_0 \mid x')| &\le \frac{|g_\theta(u) - g_{\theta'}(u')|}{\underline{p}} + \frac{L_\beta |\lambda_\theta(x) - \lambda_{\theta'}(x')|}{\underline{p}^2} \\ &\le K_q \left\{ H_\theta \|\theta - \theta'\|_2^{\beta_\theta} + H_{\mathrm{dens}} \|x - x'\|_2^{\beta'} \right\}. \end{aligned}$$

This proves (42) to (45) with (46). The sample-only case follows directly from $q_\theta = g_\theta$, the density bounds, the latent modulus, and the mean-value bound for the logarithm, giving (47). Joint Borel measurability follows from parameterized integration and composition with continuous functions on intervals bounded away from zero. □

D.3 Diffusion features and latent-grid quadrature

We now express the scaled score using the conditional densities just bounded. The common normalization below keeps the denominator bounded away from zero, which allows us to compare the exact latent average with its grid approximation.

For $t > 0$, define

$$G_t(z, y) := (2\pi\sigma_t^2)^{-d_z/2} \exp\left\{ -\frac{\|z - \alpha_t y\|_2^2}{2\sigma_t^2} \right\}, \qquad I_t(z) := \int_{[0,1]^{d_z}} G_t(z, y)\, \mathrm{d}y.$$

For $\theta \in \mathcal{T}$, set

$$\begin{aligned} \mathsf{a}_\theta(x, z, t) &:= \frac{\int q_\theta(y \mid x) G_t(z, y)\, \mathrm{d}y}{I_t(z)}, \\ \mathsf{b}_\theta(x, z, t) &:= \frac{\sigma_t \int q_\theta(y \mid x) \nabla_z G_t(z, y)\, \mathrm{d}y}{I_t(z)}. \end{aligned} \tag{48}$$

Here a_θ is the ratio of the smoothed density to $I_t(z)$, and b_θ is the ratio of its σ_t-scaled gradient to $I_t(z)$. The density bounds imply $\underline{q}_{\mathrm{out}} \le \mathsf{a}_\theta \le \overline{q}_{\mathrm{out}}$. On $\mathcal{G}_n$, put

$$R^{\mathrm{lat}}_{\mathrm{num},n} := \overline{q}_{\mathrm{out}} R_{\mathrm{bal},n}.$$

For every $\theta, \theta' \in \mathcal{T}$ and every $(h, z, t) \in \mathcal{G}_n$, with $x_\star$ denoting the query component of h, the basis functions satisfy

$$\begin{aligned} \|\mathsf{b}_\theta(x_\star, z, t)\|_2 &\le R^{\mathrm{lat}}_{\mathrm{num},n}, \\ |\mathsf{a}_\theta(x_\star, z, t) - \mathsf{a}_{\theta'}(x_\star, z, t)| + \|\mathsf{b}_\theta(x_\star, z, t) - \mathsf{b}_{\theta'}(x_\star, z, t)\|_2 &\le (1 + R_{\mathrm{bal},n}) H_{q,\theta} \|\theta - \theta'\|_2^{\beta_\theta}. \end{aligned} \tag{49}$$

For an integer $m_{\theta,n} \ge 1$, partition $\mathcal{T}$ into $J_{\theta,n} := m_{\theta,n}^{d_\theta}$ half-open cubes $(\mathcal{C}_{j,n})_{j=1}^{J_{\theta,n}}$ of side length $m_{\theta,n}^{-1}$, with terminal cells including the outer boundary. Let $\theta_{j,n}$ be the midpoint of $\mathcal{C}_{j,n}$, and set

$$\eta_{\theta,n} := \frac{\sqrt{d_\theta}}{m_{\theta,n}}, \qquad w_{j,n} := \Pi(\mathcal{C}_{j,n}).$$

Define the Borel cell-midpoint map $\pi_{\eta_{\theta,n}}(\theta) := \theta_{j,n}$ for $\theta \in \mathcal{C}_{j,n}$. Then $\|\theta - \pi_{\eta_{\theta,n}}(\theta)\|_2 \le \eta_{\theta,n}$ for every $\theta \in \mathcal{T}$, so the regular grid below is the discretization in Proposition D.4 with $\eta = \eta_{\theta,n}$. For a prompt h, let $\boldsymbol{c}_n = (c_1, \dots, c_n)$ and $x_\star$ denote its encoded context and query components, with $x_\star = 0$ in the sample-only prompt model, and define

$$\begin{aligned} \mathsf{Lik}_{j,n}(h) &:= \exp\left\{ \ell^{\mathrm{qry}}_{\theta_{j,n}}(x_\star) + \sum_{i=1}^{n} \ell_{\theta_{j,n}}(c_i) \right\}, \\ \omega_{j,n}(h) &:= \frac{w_{j,n} \mathsf{Lik}_{j,n}(h)}{\sum_{k=1}^{J_{\theta,n}} w_{k,n} \mathsf{Lik}_{k,n}(h)}. \end{aligned}$$

Here $\mathsf{Lik}_{j,n}(h)$ is the prompt likelihood at the jth grid point, and $\omega_{j,n}(h)$ is the approximate posterior mass assigned to latent cell j. The resulting quadrature approximation to the scaled score is

$$\widetilde{s}^{\mathrm{quad}}_n(z, t \,|\, h) := \frac{\sum_j \omega_{j,n}(h) \mathsf{b}_{\theta_{j,n}}(x_\star, z, t)}{\sum_j \omega_{j,n}(h) \mathsf{a}_{\theta_{j,n}}(x_\star, z, t)}. \tag{50}$$

D.3.1 Score identity and regularity of the diffusion features

For $z \in \mathbb{R}^{d_z}$, $t > 0$, and $y \in [0,1]^{d_z}$, define

$$\nu_{z,t}(\mathrm{d}y) := \frac{G_t(z, y)\, \mathrm{d}y}{I_t(z)}, \qquad \Lambda_t(z, y) := -\frac{z - \alpha_t y}{\sigma_t}.$$

Then $\nu_{z,t}$ is a probability distribution and

$$\mathsf{a}_\theta(x, z, t) = \mathbb{E}_{\nu_{z,t}} q_\theta(Y \mid x), \qquad \mathsf{b}_\theta(x, z, t) = \mathbb{E}_{\nu_{z,t}}[q_\theta(Y \mid x) \Lambda_t(z, Y)]. \tag{51}$$

On $\mathcal{G}_n$,

$$\|\Lambda_t(z, y)\|_2 \le \frac{R_{\mathrm{state},n} + \sqrt{d_z}}{\sigma_{t_0}} = R_{\mathrm{bal},n}.$$

The density bounds and (45) therefore imply (49).

Conditional independence given the task makes the prompt likelihood the product of the context-example densities and, for input–output prompts, the query density. Bayes' formula therefore gives a task posterior proportional to

$$\exp\left\{ \ell^{\mathrm{qry}}_\theta(x_\star) + \sum_{i=1}^{n} \ell_\theta(c_i) \right\} \Pi(\mathrm{d}\theta).$$

This formula is defined for every prompt and agrees almost surely with the task posterior. Forward OU convolution of the posterior-predictive density and differentiation under the integral are justified by the bounded density and the Gaussian kernel. The common factors $I_t(z)$ and the task-posterior normalizer cancel from the score ratio. Consequently,

$$\widetilde{s}^\star(z, t \mid h) = \frac{\int_{\mathcal{T}} \mathsf{Lik}_n(\theta; h) \mathsf{b}_\theta(x_\star, z, t)\, \Pi(\mathrm{d}\theta)}{\int_{\mathcal{T}} \mathsf{Lik}_n(\theta; h) \mathsf{a}_\theta(x_\star, z, t)\, \Pi(\mathrm{d}\theta)}, \quad \mathsf{Lik}_n(\theta; h) := \exp\Big\{ \ell^{\mathrm{qry}}_\theta(x_\star) + \sum_{i=1}^{n} \ell_\theta(c_i) \Big\}. \tag{52}$$

Lemma D.3 (Normalized OU-basis regularity). *Under Assumption 3.1, put* $\mathcal{U}_{\mathrm{feat}} := [-1,1]^{d_{\mathrm{stat}}+1}$ *and* $R_{\mathrm{geom},n} := 1 + R_{\mathrm{state},n} + \sqrt{d_z}$. *For every* $\theta \in \mathcal{T}$, *there exist continuous maps*

$$\overline{\mathsf{a}}_{\theta,n} : \mathcal{U}_{\mathrm{feat}} \to [\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}], \qquad \overline{\mathsf{b}}_{\theta,n} : \mathcal{U}_{\mathrm{feat}} \to \overline{\mathrm{Ball}}(0, R^{\mathrm{lat}}_{\mathrm{num},n}),$$

such that, for every $(h,z,t) \in \mathcal{G}_n$,

$$\overline{\mathsf{a}}_{\theta,n}(\mathsf{u}_n(z,t,h)) = \mathsf{a}_\theta(x_\star, z, t), \qquad \overline{\mathsf{b}}_{\theta,n}(\mathsf{u}_n(z,t,h)) = \mathsf{b}_\theta(x_\star, z, t).$$

Moreover, there is a single finite constant $H_{\mathrm{bas},n}$, *independent of* θ, *such that, for all* $u, u' \in \mathcal{U}_{\mathrm{feat}}$,

$$|\overline{\mathsf{a}}_{\theta,n}(u) - \overline{\mathsf{a}}_{\theta,n}(u')| + \|\overline{\mathsf{b}}_{\theta,n}(u) - \overline{\mathsf{b}}_{\theta,n}(u')\|_2 \le H_{\mathrm{bas},n}\|u - u'\|_2^{\beta'}.$$

The common Hölder constant satisfies

$$\begin{aligned} H_{\mathrm{bas},n} \le C_{\mathrm{bas}}\big[1 + (1 + R_{\mathrm{bal},n})H_{q,x} \\ + R_{\mathrm{state},n}\{R_{\mathrm{geom},n}\sigma_{t_0}^{-2} + \sigma_{t_0}^{-1} + R_{\mathrm{geom},n}^2\sigma_{t_0}^{-3}\} \\ + T\{R_{\mathrm{geom},n}^2\sigma_{t_0}^{-4} + \sigma_{t_0}^{-1} + R_{\mathrm{geom},n}\sigma_{t_0}^{-3} + R_{\mathrm{geom},n}^3\sigma_{t_0}^{-5}\}\big]. \end{aligned} \tag{53}$$

Here $0 < C_{\mathrm{bas}} < \infty$ *depends only on the fixed dimensions,* β', *and the common density bounds.*

Proof of Lemma D.3. The Gaussian identities

$$\begin{aligned} \nabla_z \Lambda_t(z,y) &= -\sigma_t^{-1} I_{d_z}, \\ \nabla_z \log G_t(z,y) &= \sigma_t^{-1}\Lambda_t(z,y), \\ \partial_t \Lambda_t(z,y) &= -\frac{\alpha_t y}{\sigma_t} - \frac{\alpha_t^2}{\sigma_t^2}\Lambda_t(z,y), \\ \partial_t \log G_t(z,y) &= -\frac{d_z\alpha_t^2}{\sigma_t^2} - \frac{(z-\alpha_t y)^\top \alpha_t y}{\sigma_t^2} + \frac{\alpha_t^2\|z-\alpha_t y\|_2^2}{\sigma_t^4} \end{aligned}$$

and the covariance identity below imply the derivative bounds

$$\begin{aligned} \|\nabla_z \mathsf{a}_\theta\|_2 &\lesssim R_{\mathrm{geom},n}\sigma_{t_0}^{-2}, & \|\nabla_z \mathsf{b}_\theta\|_{\mathrm{op}} &\lesssim \sigma_{t_0}^{-1} + R_{\mathrm{geom},n}^2\sigma_{t_0}^{-3}, \\ |\partial_t \mathsf{a}_\theta| &\lesssim R_{\mathrm{geom},n}^2\sigma_{t_0}^{-4}, & \|\partial_t \mathsf{b}_\theta\|_2 &\lesssim \sigma_{t_0}^{-1} + R_{\mathrm{geom},n}\sigma_{t_0}^{-3} + R_{\mathrm{geom},n}^3\sigma_{t_0}^{-5}. \end{aligned}$$

The implicit constants depend only on the fixed dimensions and density bounds. The denominator $I_t(z)$ causes no extra small-denominator factor because, for every bounded differentiable scalar- or vector-valued integrand F and each coordinate derivative

$$D \in \{\partial_t, \partial_{z_1}, \dots, \partial_{z_{d_z}}\},$$

one has, componentwise in the vector-valued case,

$$D\mathbb{E}_{\nu_{z,t}}[F(z,t,Y)] = \mathbb{E}_{\nu_{z,t}}[DF(z,t,Y)] + \mathrm{Cov}_{\nu_{z,t}}\big(F(z,t,Y), D\log G_t(z,Y)\big). \tag{54}$$

Differentiation is justified by domination on the compact domain for (x,z,t) defined above. For the query coordinate, (45) and (51) give a β'-Hölder modulus at most

$$(1 + R_{\mathrm{bal},n})H_{q,x}.$$

To define the functions approximated by the network on the entire normalized cube, let $u \in \mathcal{U}_{\mathrm{feat}}$, written as $u = (u_x, u_z, u_t)$ in the input–output prompt model and $u = (u_z, u_t)$ in the sample-only prompt model, and set

$$x(u) := \begin{cases} 0, & \text{sample-only}, \\ (u_x + \mathbf{1}_{d_x})/2, & \text{input–output}, \end{cases}$$

$$z(u) := R_{\mathrm{state},n}\mathrm{clip}_1(u_z), \qquad t(u) := t_0 + \frac{T - t_0}{2}(u_t + 1).$$

Define the extensions by

$$\overline{\mathsf{a}}_{\theta,n}(u) := \mathsf{a}_\theta(x(u), z(u), t(u)), \qquad \overline{\mathsf{b}}_{\theta,n}(u) := \mathsf{b}_\theta(x(u), z(u), t(u)).$$

Since clip_1 is one-Lipschitz, the three coordinate maps above have Lipschitz constants bounded by a fixed constant, $R_{\mathrm{state},n}$, and T, respectively. Multiplying the state and time derivative bounds by the Lipschitz constants of these coordinate maps and adding the query modulus gives the asserted Hölder inequality, with $H_{\mathrm{bas},n}$ satisfying (53), after increasing C_{bas} by a fixed dimension-dependent factor. By construction, the extensions take values in the ranges stated in the lemma. For every $(h, z, t) \in \mathcal{G}_n$, the state block of $\mathsf{u}_n(z, t, h)$ lies in the unit ball and represents the unclipped state; substituting the feature coordinates into the definitions above therefore gives $\bar{\mathsf{a}}_{\theta,n}(\mathsf{u}_n(z, t, h)) = \mathsf{a}_\theta(x_\star, z, t)$ and $\bar{\mathsf{b}}_{\theta,n}(\mathsf{u}_n(z, t, h)) = \mathsf{b}_\theta(x_\star, z, t)$. Since $\mathcal{U}_{\mathrm{feat}}$ has bounded diameter and $\beta' \le 1$, the state–time Lipschitz bounds are also β'-Hölder after increasing their constant. This proves all claims of the lemma. □

D.4 Episode-level KL and integrated score approximation

Uniform approximation over every prompt is stronger than the integrated error required by DSM and the sampler. We therefore compare entire episode distributions before conditioning on the prompt. This comparison also controls the addition of uniform prior mass. All expectations in this subsection use the original model, including when the score being evaluated comes from a discretized model.

Proposition D.4 (Episode-level KL control of integrated score error)**.** *Under Assumption 3.1, fix an integer $J \ge 1$, pairwise distinct grid points $\theta_1, \dots, \theta_J \in \mathcal{T}$, and $\eta \ge 0$, and let $\pi_\eta : \mathcal{T} \to \{\theta_1, \dots, \theta_J\}$ be Borel with $\|\theta - \pi_\eta(\theta)\|_2 \le \eta$ for every θ. Set $w_j = \Pr(\pi_\eta(\Theta) = \theta_j)$ and, for $\zeta \in [0, 1]$, set*

$$w_j^\zeta := \frac{w_j + \zeta/J}{1 + \zeta}. \tag{55}$$

Let $Q_{\mathrm{epi}}^{\eta,\zeta}$ be the observed-episode distribution obtained by drawing a grid task with weights w_j^ζ and then drawing $n + 1$ conditionally independent examples from g_{θ_j}. In the input–output model, an example is a full pair (X, Z) and the observed episode is $(C_n, X_\star, Z_\star)$. Because every grid density is positive, let $Q_{0|h}^{\eta,\zeta}$ denote the everywhere-defined Bayes conditional law of $Z_\star$ given $H_n = h$ under $Q_{\mathrm{epi}}^{\eta,\zeta}$. If $q_t^{\eta,\zeta}(\cdot \mid h)$ denotes its OU-smoothed density, define $s^{\eta,\zeta}(z, t \mid h) := \nabla_z \log q_t^{\eta,\zeta}(z \mid h)$. Define $D_n(\eta, \zeta) = (n + 1)H_\theta^2 \eta^{2\beta_\theta}/\underline{p} + \log(1 + \zeta)$. Then

$$\mathrm{KL}\Big(\mathbb{P}_{\mathrm{epi}}^{(n)} \Big\| Q_{\mathrm{epi}}^{\eta,\zeta}\Big) \le D_n(\eta, \zeta), \tag{56}$$

$$\mathbb{E}_{H_n} \mathrm{KL}(P_{0|H_n} \| Q_{0|H_n}^{\eta,\zeta}) \le D_n(\eta, \zeta). \tag{57}$$

Moreover, for every $0 < t_0 < T < \infty$,

$$\int_{t_0}^{T} \mathbb{E}\left[\|s^\star(Z_t, t \mid H_n) - s^{\eta,\zeta}(Z_t, t \mid H_n)\|_2^2\right] \mathrm{d}t \le D_n(\eta, \zeta), \tag{58}$$

where (H_n, Z_t) has the original prompt and conditional diffusion law. The result is uniform over Borel task priors, including priors with zero cell masses.

Proof. For densities g, g' on a unit-volume cube with $g' \ge \underline{p}$, $\log u \le u - 1$ gives

$$\mathrm{KL}(g \| g') \le \int \frac{(g - g')^2}{g'} \le \underline{p}^{-1} \|g - g'\|_\infty^2.$$

Set $U_i := Z_i$ in the sample-only model and $U_i := (X_i, Z_i)$ in the input–output model. Compare the two joint laws of $(\Theta, U_1, \dots, U_{n+1})$ with the same prior Π, using g_Θ and $g_{\pi_\eta(\Theta)}$ as the conditional example densities. Tensorization of conditional KL gives a bound $(n + 1)H_\theta^2 \eta^{2\beta_\theta}/\underline{p}$. Marginalizing Θ gives the same bound for the full example sequence. Reordering that sequence into $(H_n, Z_\star)$ is a measurable bijection under either the sample-only or the input–output prompt model, so it preserves KL. This proves (56) for $\zeta = 0$.

Let $Q_{\mathrm{epi}}^{\eta,\mathrm{unif}}$ use the uniform distribution over the same grid. The weights in (55) give the identity

$$Q_{\mathrm{epi}}^{\eta,\zeta} = \frac{Q_{\mathrm{epi}}^{\eta,0} + \zeta Q_{\mathrm{epi}}^{\eta,\mathrm{unif}}}{1 + \zeta}.$$

Its density is at least $(1+\zeta)^{-1}$ times the density of $Q^{\eta,0}_{\mathrm{epi}}$. Integrating the resulting log-density inequality proves (56). The KL chain rule at $(H_n, Z_\star)$ proves (57): the conditional KL is averaged under the original prompt law, and the chain rule also contains a nonnegative KL term for the prompt marginal that can be dropped.

Fix h in a $\mathsf{Law}(H_n)$-full-measure set on which the conditional laws are defined, and write p_t, q_t for the OU-smoothed densities of $P_{0|h}, Q^{\eta,\zeta}_{0|h}$. Their initial densities on the output cube lie in $[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}]$ by Lemma D.2. Hence their ratio after convolution is bounded above and below by positive constants. For $t \in [t_0, T]$, convolution with the OU kernel gives smooth positive densities with Gaussian tails. Their spatial derivatives are bounded by Gaussian envelopes times polynomials in $\|z\|$, uniformly on this time interval. In particular, the score identity from differentiating the kernel gives $\|\nabla \log p_t(z)\| + \|\nabla \log q_t(z)\| \lesssim_{t_0,T,d_z} 1 + \|z\|$.

Both densities solve $\partial_t p = \Delta p + \nabla \cdot (zp)$. Differentiating $\int p_t \log(p_t/q_t)$ and integrating by parts therefore gives the relative-entropy dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t\|q_t) = -\int p_t \|\nabla \log p_t - \nabla \log q_t\|_2^2 \,\mathrm{d}z. \tag{59}$$

The drift terms cancel and the coefficient is one for diffusion coefficient $\sqrt{2}$; compare Chen et al. (2016). To justify the calculation on $\mathbb{R}^{d_z}$, insert smooth cutoffs equal to one on a ball of radius R, with first and second derivatives of order R^{-1} and R^{-2}, respectively, perform the calculation on their compact support, and let $R \to \infty$. The bounded density ratio, the score growth bound, and the Gaussian bounds on the derivatives give an integrable upper bound for every differentiated and boundary term, so dominated convergence applies. Integrating (59) gives

$$\begin{aligned}\int_{t_0}^{T}\int_{\mathbb{R}^{d_z}} p_t(z)\|s^\star(z,t\mid h) - s^{\eta,\zeta}(z,t\mid h)\|_2^2 \,\mathrm{d}z\,\mathrm{d}t &= \mathrm{KL}(p_{t_0}\|q_{t_0}) - \mathrm{KL}(p_T\|q_T)\\ &\le \mathrm{KL}(P_{0|h}\|Q^{\eta,\zeta}_{0|h}),\end{aligned}$$

where the last inequality uses nonnegativity and contraction of KL under the OU kernel. Tonelli's theorem and (57) finish the proof. □

For the regular grid in Section D.3, take $\eta = \eta_{\theta,n}$ and $\pi_\eta = \pi_{\eta_{\theta,n}}$. If $\omega^\zeta_{j,n}(h)$ denotes the posterior weight obtained from $w^\zeta_{j,n} = (w_{j,n} + \zeta/J_{\theta,n})/(1+\zeta)$ and $\mathsf{Lik}_{j,n}(h)$, then the everywhere-defined Bayes formula gives

$$\sigma_t s^{\eta_{\theta,n},\zeta}(z,t\mid h) = \frac{\sum_j \omega^\zeta_{j,n}(h)\mathsf{b}_{\theta_{j,n}}(x_\star,z,t)}{\sum_j \omega^\zeta_{j,n}(h)\mathsf{a}_{\theta_{j,n}}(x_\star,z,t)}. \tag{60}$$

Thus the floored quadrature score used in Section E is exactly the scaled score of the discretized model; only its neural component maps and final readout are approximated.

The piecewise-constant map $g_{\pi_\eta(\theta)}$ is used only to approximate the original task family; it need not be Hölder as a function of θ. Each grid task has the same spatial regularity and positivity as the original task family. The neural construction in Section E approximates the component maps and ratio while its four attention blocks exactly aggregate the resulting finite-mixture terms.

E TRANSFORMER CONSTRUCTION, APPROXIMATION, AND CLASS COMPLEXITY

The finite latent average from Section D specifies the score computation to approximate. This section proceeds in three stages: it first defines the scaled-score metric and the prompt-conditioned Transformer class, then uses four attention blocks to aggregate the approximated finite-mixture components exactly, and finally derives the integrated approximation and class-complexity bounds. The last stage is needed because training need not return the specifically constructed parameters.

E.1 SCORE METRIC, TRANSFORMER CLASS, AND FOUR-BLOCK REALIZATION

E.1.1 SCALED-SCORE METRIC AND COMPONENT CLASSES

For a nonempty class $\mathcal{S}$ of jointly Borel score functions, define its population approximation error by

$$\mathcal{E}_{\mathrm{app}}(\mathcal{S}) := \inf_{s\in\mathcal{S}} \mathcal{E}_{\mathrm{score}}(s). \tag{61}$$

For $s \in \mathcal{S}$, write $\widetilde{s} := \sigma_t s$ and define

$$\widetilde{\mathcal{S}} := \{\widetilde{s} = \sigma_t s : s \in \mathcal{S}\}, \qquad d_{\mathrm{bal},\infty}(s,s') := \sup_{(z,t,h)\in\mathbb{R}^{d_z}\times[t_0,T]\times\mathcal{H}_n} \|\sigma_t s(z,t,h) - \sigma_t s'(z,t,h)\|_2. \tag{62}$$

The value of $d_{\mathrm{bal},\infty}$ may be $+\infty$ for an unbounded class, so it is an extended metric until a finite uniform bound on the scaled-score output is imposed. For a vector-valued function f on this score domain, write $\|f\|_{\infty,2} := \sup_{z,t,h} \|f(z,t,h)\|_2$. Thus $d_{\mathrm{bal},\infty}(s,s') = \|\sigma_t(s-s')\|_{\infty,2}$. For finite-dimensional vectors, $\|\cdot\|_\infty$ denotes the maximum absolute coordinate; for matrices, $\|A\|_{\max} := \max_{i,j}|A_{ij}|$. For a set $\mathcal{U}$ equipped with an extended metric d, the covering number $N_{\mathrm{cov}}(\varepsilon,\mathcal{U},d)$ is the smallest cardinality of an ε-cover whose centers belong to $\mathcal{U}$ and equals $+\infty$ if no finite cover exists (cf. van der Vaart & Wellner, 2023; Wainwright, 2019). This metric uses the same scaling as the variance-balanced DSM loss in Section F.

For fixed input and output dimensions, $\mathcal{NN}(L,W,S,B)$ denotes the class of feedforward ReLU networks with L affine layers, maximum hidden-layer width W, at most S nonzero weights and biases, and parameter magnitudes at most B. The output affine layer is included in L, biases are included in S, and the subsequent projection onto the prescribed output range is not part of $\mathcal{NN}(L,W,S,B)$. Every component class used below has a fixed sparse wiring pattern, so its covering number does not include a separate search over sparsity patterns. Basis, coordinate-update, and arithmetic weights that do not depend on the approximated function may be fixed as part of that wiring.

E.1.2 PROMPT-CONDITIONED TRANSFORMER CLASS

The class takes a prompt, a noisy output, and diffusion time as inputs and returns a scaled score, followed by division by σ_t. The grid and component architectures are fixed before training; the definition below specifies which coefficients can vary. Its references to the component-approximation lemmas specify the network classes, while Section E.1.3 gives the token layout and verifies the four-block computation at the constructed parameters.

The tolerances specify the required accuracy of each component: $\varepsilon_{\ell,n}$ and $\varepsilon_{\mathrm{qry},n}$ bound uniform log-likelihood errors, $\varepsilon_{a,n}$ and $\varepsilon_{b,n}$ bound errors in the scalar OU denominator and vector numerator, and $\varepsilon_{\mathrm{rat},n}$ bounds the ratio-readout error. The vector errors are measured in Euclidean norm.

Definition E.1 (Prompt-conditioned score Transformer)**.** *Fix n and a latent grid, and choose tolerances*

$$\varepsilon_{\mathrm{floor},n},\ \varepsilon_{\ell,n},\ \varepsilon_{a,n},\ \varepsilon_{b,n},\ \varepsilon_{\mathrm{rat},n} \in (0,1].$$

Also fix a deterministic final output radius $0 < R_{\mathrm{out},n} \le R_{\mathrm{bal},n}$. The radius $R_{\mathrm{bal},n}$ remains the uniform bound used for latent quadrature and internal basis estimates; $R_{\mathrm{out},n}$ controls the score-network output and the training loss. In the input–output prompt model, also fix $\varepsilon_{\mathrm{qry},n} \in (0,1]$; in the sample-only prompt model, set $\varepsilon_{\mathrm{qry},n} := 0$ and omit the query factor below. Index the context-example tokens by $I_{\mathrm{ctx}} := \{1,\dots,n\}$ and the latent tokens by $I_{\mathrm{lat}} := \{i_1^{\mathrm{lat}},\dots,i_{J_{\theta,n}}^{\mathrm{lat}}\}$. Let i_{sum}, i_{qry}, i_{out}, and $i_\varnothing$ denote the summary, query–state–time, output, and null tokens, respectively. Thus

the sequence has $n + J_{\theta,n} + 4$ positions. Choose the explicit token width in (67), namely

$$d^{\mathrm{lat}}_{\mathrm{emb},n} = (d_z + 5)J_{\theta,n} + d_{\mathrm{stat}} + 2d_z + 11 \le C_{d_z}\{J_{\theta,n} + d_{\mathrm{stat}} + 1\}.$$

Trainable parameter space. *Let $\mathcal{V}^{\mathrm{fit}}_n$ be the product of the compact parameter cubes for the likelihood approximators with bounded outputs in Lemma E.3, the time-local OU-basis approximators in Lemma E.4, and the prior logits with radius (81). For attention, include every entry of each of the sixteen matrices*

$$W_{Q,k}, W_{K,k}, W_{V,k}, W_{O,k} \in [-B_{\mathrm{att},n}, B_{\mathrm{att},n}]^{d^{\mathrm{lat}}_{\mathrm{emb},n} \times d^{\mathrm{lat}}_{\mathrm{emb},n}}, \quad 1 \le k \le 4,$$

$$B_{\mathrm{att},n} = 1 \vee \sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}.$$

Thus the attention part has exactly $16(d^{\mathrm{lat}}_{\mathrm{emb},n})^2$ trainable coordinates. The sparse matrices in Table 1 specify a reference parameter within this entire score class. The table describes the computation at that parameter; arbitrary trainable projections need not preserve its internal posterior interpretation or role coordinates. The hard masks and coordinate-update maps are fixed for the entire class. The query log-likelihood factor is omitted in the sample-only prompt model. The fixed approximator architecture, latent identifiers, role codes, hard masks, coordinate-update maps, ratio readout, and radial output projection are not coordinates of $\mathcal{V}^{\mathrm{fit}}_n$. Let $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n} := \dim(\mathcal{V}^{\mathrm{fit}}_n)$, let $B_{\mathrm{lat},n} \ge 1$ dominate its coordinate radii, and let

$$\mathrm{Proj}^{\mathrm{par}}_n : [-B_{\mathrm{lat},n}, B_{\mathrm{lat},n}]^{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n}} \longrightarrow \mathcal{V}^{\mathrm{fit}}_n$$

be coordinatewise projection onto the corresponding factor ranges. This map is continuous, nonexpansive in ℓ_∞, and onto.

Token representation. *For a parameter vector $\mathbf{v}$ in the displayed cube, put $\widehat{\mathbf{v}} := \mathrm{Proj}^{\mathrm{par}}_n(\mathbf{v})$. We identify $\mathcal{V}^{\mathrm{fit}}_n$ with its natural subset of the enclosing cube; on this subset, $\mathrm{Proj}^{\mathrm{par}}_n$ is the identity. Let $\mathbf{E}^{(0)}_n(\widehat{\mathbf{v}}; z, t, h)$ be the encoded array of the n context-example tokens, one summary token, one query–state–time token, $J_{\theta,n}$ latent tokens, one output token, and one null token. In particular, every trainable component encoder is evaluated at the projected parameter vector $\widehat{\mathbf{v}}$. The network-component specification in Appendix E.1.3 gives the initial value and shape of every token type, with ordered coordinate decomposition (68). The context-example network uses the fixed affine rescaling $c \mapsto 2c - \mathbf{1}_{d_{\mathrm{stat}}}$. In the input–output prompt model, the query log-likelihood network uses $x_\star \mapsto 2x_\star - \mathbf{1}_{d_x}$. The OU-basis network uses the globally bounded feature $\mathsf{u}_n(z, t, h)$ from (37). Its fixed time-local encoder combines overlapping time patches with a continuous ReLU partition of unity, as specified in Lemma E.4; the four attention blocks are unchanged. All occurrences of the noisy state in this encoded array use $\bar{z} = \mathrm{clip}_{R_{\mathrm{state},n}}(z)$; there is no unprocessed state input. Consequently, for every parameter in the class,*

$$\widetilde{s}_{\mathbf{v}}(z, t, h) = \widetilde{s}_{\mathbf{v}}(\bar{z}, t, h). \tag{63}$$

This identity reduces the full-state supremum in $d_{\mathrm{bal},\infty}$ to the compact state ball. We suppress the common arguments $(\widehat{\mathbf{v}}; z, t, h)$ from the token arrays in the recursion below.

Attention blocks. *For $1 \le k \le 4$ and every token row i, define*

$$\mathbf{E}^{(k)}_{n,i} := \mathsf{R}_{k,n}\left[\mathbf{E}^{(k-1)}_{n,i} + \mathrm{Att}_{\mathbf{A}_k(\widehat{\mathbf{v}})}\left(\mathbf{E}^{(k-1)}_{n,i}; (\mathbf{E}^{(k-1)}_{n,j})_{j \in \mathsf{M}_{k,n}(i)}\right)\right]. \tag{64}$$

Here $\mathbf{A}_k(\widehat{\mathbf{v}})$ denotes the four dense bounded attention projection matrices, and $\mathsf{R}_{k,n}$ is the fixed affine–ReLU coordinate-update map after block k. The non-null masks are

$$\mathsf{M}_{1,n}(i_{\mathrm{sum}}) = I_{\mathrm{ctx}}, \qquad \mathsf{M}_{2,n}(i^{\mathrm{lat}}_j) = \{i_{\mathrm{sum}}\}, \qquad \mathsf{M}_{3,n}(i^{\mathrm{lat}}_j) = \{i_{\mathrm{qry}}\}, \qquad \mathsf{M}_{4,n}(i_{\mathrm{out}}) = I_{\mathrm{lat}},$$

for $i^{\mathrm{lat}}_j \in I_{\mathrm{lat}}$, and every undeclared row attends only to $i_\varnothing$. The projection matrices of the constructed network, residual destinations, and states after $\mathsf{R}_{k,n}$ are listed in Table 1; bounded coordinatewise projection precedes the exact selector in (71).

Score readout and class. *Let Read_n first apply exact coordinatewise projection onto*

$$[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}] \times [-R^{\mathrm{lat}}_{\mathrm{num},n} - 1, R^{\mathrm{lat}}_{\mathrm{num},n} + 1]^{d_z},$$

and then apply the fixed affine normalization and ReLU ratio approximator constructed below. This preliminary projection is part of the readout for every parameter; it is inactive on the reference network's averaged features. Let $\mathsf{P}_{\mathrm{out}}$ *extract the output-token coordinates. Define*

$$\widetilde{s}_{\mathbf{v}}(z,t,h) := \mathrm{clip}_{R_{\mathrm{out},n}}\Big[\mathsf{Read}_n\Big(\mathsf{P}_{\mathrm{out}}\mathbf{E}_n^{(4)}(\widehat{\mathbf{v}};z,t,h)\Big)\Big], \qquad s_{\mathbf{v}}(z,t,h) := \frac{\widetilde{s}_{\mathbf{v}}(z,t,h)}{\sigma_t}, \tag{65}$$

and set

$$\mathcal{S}_n^{\mathrm{lat}} := \Big\{ s_{\mathbf{v}} : \mathbf{v} \in [-B_{\mathrm{lat},n}, B_{\mathrm{lat},n}]^{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n}} \Big\}. \tag{66}$$

Writing $\mathbf{A}_k(\widehat{\mathbf{v}}) = (W_{Q,k}, W_{K,k}, W_{V,k}, W_{O,k})$*, the attention map in (64) is* $\mathrm{Att}_{W_{Q,k},W_{K,k},W_{V,k},W_{O,k}}$ *from Lemma E.2 restricted to the displayed deterministic mask. Here "Transformer" means the explicitly defined prompt-conditioned architecture: external ReLU input and readout modules surround four hard-masked residual softmax-attention blocks with fixed affine–ReLU coordinate updates and no layer normalization. Exact containment in a fixed-width conventional Transformer with mandatory feed-forward sublayers or layer normalization is not claimed. In particular, the definition does not cover a standard pre-normalized or post-normalized Transformer with fully learned coordinate updates.*

E.1.3 Token layout and four-block computation

The following specification completes the class definition by giving the token coordinates and their initialization. The construction lemma identifies parameter values at which the four attention blocks perform the required averaging and copy the relevant coordinates. The integrated approximation error is analyzed in Proposition E.6.

Trainable and fixed components. Fix $n \geq 1$, write $J := J_{\theta,n}$, and let

$$d^{\mathrm{lat}}_{\mathrm{emb},n} := (d_z+5)J + d_{\mathrm{stat}} + 2d_z + 11. \tag{67}$$

For $\mathbf{v} \in [-B_{\mathrm{lat},n}, B_{\mathrm{lat},n}]^{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n}}$, write $\widehat{\mathbf{v}} := \mathrm{Proj}_n^{\mathrm{par}}(\mathbf{v})$. Use the ordered token decomposition

$$\mathbb{R}^{d^{\mathrm{lat}}_{\mathrm{emb},n}} = \mathsf{V}_{\mathrm{role}} \oplus \mathsf{V}_{\mathrm{raw}} \oplus \mathsf{V}_{\ell} \oplus \mathsf{V}_{\mathrm{pool}} \oplus \mathsf{V}_{\mathrm{qry}} \oplus \mathsf{V}_{\mathrm{id}} \oplus \mathsf{V}_{\mathrm{prior}} \oplus \mathsf{V}_{\mathrm{lat}} \oplus \mathsf{V}_{\mathrm{out}}, \tag{68}$$

whose block dimensions, in the displayed order, are

$$\big(6, d_{\mathrm{stat}}+1, J, J, J(d_z+2), J, 1, d_z+2, d_z+1\big).$$

Index the rows by

$$I_{\mathrm{ctx}} := \{1,\dots,n\}, \qquad i_{\mathrm{sum}} := n+1, \qquad i_{\mathrm{qry}} := n+2,$$
$$i_j^{\mathrm{lat}} := n+2+j, \qquad I_{\mathrm{lat}} := \{i_j^{\mathrm{lat}} : 1 \leq j \leq J\},$$

with $i_{\mathrm{out}} := n+J+3$ and $i_{\varnothing} := n+J+4$. Let $\mathbf{E}_n^{(0)}(\widehat{\mathbf{v}};z,t,h) \in \mathbb{R}^{(n+J+4)\times d^{\mathrm{lat}}_{\mathrm{emb},n}}$ have one-hot role codes for the context-example, summary, query, latent, output, and null rows. Let $e_j \in \mathbb{R}^J$ be the jth standard coordinate vector. The array's nonzero non-role blocks are

Rows	Initialized blocks
Context example $i \in I_{\mathrm{ctx}}$	$\mathsf{V}_{\mathrm{raw}}$: padded normalized c_i; V_{ℓ}: $\widehat{\boldsymbol{\ell}}(c_i)$
Query	$\mathsf{V}_{\mathrm{raw}}$: $\mathsf{u}_n(z,t,h)$; $\mathsf{V}_{\mathrm{qry}}$: $\widehat{\mathbf{q}}(x_\star,z,t)$
Latent i_j^{lat}, $1 \leq j \leq J$	V_{id}: e_j; $\mathsf{V}_{\mathrm{prior}}$: $\ell^{\mathrm{prior}}_{j,n}$
Summary, output, and null	none

Here

$$\widehat{\boldsymbol{\ell}}(c_i) := (\widehat{\ell}_j(c_i))_{j=1}^J \quad \text{and} \quad \widehat{\mathbf{q}}(x_\star,z,t) := (\widehat{\ell}_j^{\mathrm{qry}}(x_\star), \widehat{\mathsf{a}}_j(x_\star,z,t), \widehat{\mathsf{b}}_j(x_\star,z,t))_{j=1}^J.$$

All omitted coordinates are zero.

Lemma E.1 (Four-block Transformer construction)**.** *For the encoded array above, there exist four deterministic nonempty masks, projection matrices for the constructed network*

$$W_{Q,k}, W_{K,k}, W_{V,k}, W_{O,k} \in \mathbb{R}^{d^{\mathrm{lat}}_{\mathrm{emb},n} \times d^{\mathrm{lat}}_{\mathrm{emb},n}}, \qquad 1 \leq k \leq 4,$$

and shared rowwise affine–ReLU maps

$$\mathsf{R}_{k,n} : \mathbb{R}^{d^{\mathrm{lat}}_{\mathrm{emb},n}} \to \mathbb{R}^{d^{\mathrm{lat}}_{\mathrm{emb},n}}$$

such that each residual block maps $\mathbb{R}^{(n+J+4)\times d^{\mathrm{lat}}_{\mathrm{emb},n}}$ *to itself and implements the computation in Table 1. Every row not listed in that table attends only to the null token. After block 4, the output block equals*

$$\left(\sum_{j=1}^{J} \widehat{\omega}_{j,n}\widehat{\mathsf{a}}_j, \sum_{j=1}^{J} \widehat{\omega}_{j,n}\widehat{\mathsf{b}}_j\right) \in \mathsf{V}_{\mathrm{out}}, \qquad \widehat{\omega}_{j,n} = \mathrm{softmax}_j((\widehat{\ell}^{\mathrm{post}}_{k,n})_{k=1}^{J}), \tag{69}$$

where

$$\widehat{\ell}^{\mathrm{post}}_{j,n} = \ell^{\mathrm{prior}}_{j,n} + n\bar{\widehat{\ell}}_j + \widehat{\ell}^{\mathrm{qry}}_j, \qquad \bar{\widehat{\ell}}_j := \frac{1}{n}\sum_{i=1}^{n} \widehat{\ell}_j(c_i). \tag{70}$$

The four residual updates preserve all role codes and latent identifiers. Every projection and coordinate-update coefficient of the constructed network has magnitude at most

$$1 \vee n \vee B^{\mathrm{lat}}_{\mathrm{gate},n} \vee \sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}},$$

where $B^{\mathrm{lat}}_{\mathrm{gate},n}$ *is defined in* (83)*. Hence the construction is an explicit member of the score class in Definition E.1.*

Table 1: Complete computation implemented by the four masked attention blocks. A singleton mask makes its softmax weight exactly one. Values are projected from the indicated source block and injected into the residual destination in (68).

k	Active query rows and keys	Constructed attention logits	Value and residual destination	State after fixed update $\mathsf{R}_{k,n}$
1	Summary row; keys I_{ctx}	All zero, hence weights $1/n$	$\widehat{\boldsymbol{\ell}}(c_i) \in \mathbb{R}^J$ to $\mathsf{V}_{\mathrm{pool}}$	Summary pool is $\bar{\widehat{\boldsymbol{\ell}}} = n^{-1}\sum_i \widehat{\boldsymbol{\ell}}(c_i) \in \mathbb{R}^J$
2	Every latent row; singleton key i_{sum}	Singleton, hence weight 1	$\bar{\widehat{\boldsymbol{\ell}}} \in \mathbb{R}^J$ to the latent pool block	Every latent row contains the same pooled vector $\bar{\widehat{\boldsymbol{\ell}}} \in \mathbb{R}^J$
3	Every latent row; singleton key i_{qry}	Singleton, hence weight 1	$\widehat{\mathbf{q}} \in \mathbb{R}^{J(d_z+2)}$ to the latent query block	Identifier e_j selects $(\widehat{\ell}^{\mathrm{qry}}_j, \widehat{\mathsf{a}}_j, \widehat{\mathsf{b}}_j)$, and $\mathsf{V}_{\mathrm{lat}}$ becomes $(\widehat{\ell}^{\mathrm{post}}_{j,n}, \widehat{\mathsf{a}}_j, \widehat{\mathsf{b}}_j) \in \mathbb{R}^{d_z+2}$
4	Output row; keys I_{lat}	$\widehat{\ell}^{\mathrm{post}}_{j,n}$ for latent key j	$(\widehat{\mathsf{a}}_j, \widehat{\mathsf{b}}_j) \in \mathbb{R}^{d_z+1}$ to $\mathsf{V}_{\mathrm{out}}$	Output block is the posterior-weighted pair in (69)

Proof. All reference matrices are compositions of coordinate projections and injections associated with (68), except for the scalar factors n and $\sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}$ described below. For block 1, set the active query and key projections to zero, project the context-example values from V_ℓ, and inject the attention output into $\mathsf{V}_{\mathrm{pool}}$. The summary mask contains exactly the n context-example rows, so softmax attention returns their mean. For blocks 2 and 3, the active key set of each latent row is a singleton, so the attention weight is exactly one independently of its logit. Coordinate projection and injection therefore copy the summary pool and the full query–basis array into the indicated blocks of every latent row.

The fixed map $\mathsf{R}_{3,n}$ projects every selector input onto $[-B^{\mathrm{lat}}_{\mathrm{gate},n}, B^{\mathrm{lat}}_{\mathrm{gate},n}]$ and uses the one-hot identifier e_j. For $e \in \{0,1\}$ and $|x| \le B^{\mathrm{lat}}_{\mathrm{gate},n}$, the exact affine–ReLU identity

$$ex = \mathrm{ReLU}(x) - \mathrm{ReLU}(x - B^{\mathrm{lat}}_{\mathrm{gate},n}e) - \mathrm{ReLU}(-x) + \mathrm{ReLU}(-x - B^{\mathrm{lat}}_{\mathrm{gate},n}e) \tag{71}$$

selects one coordinate exactly. Parallel copies select the jth pooled likelihood, query likelihood, scalar basis value, and vector basis value. An affine output layer then forms (70) and writes the resulting (d_z+2)-vector into $\mathsf{V}_{\mathrm{lat}}$. The six one-hot role codes and the same binary-gate identity make these tokenwise coordinate updates shared across rows while leaving all other coordinate blocks unchanged.

For block 4, map the output-token role coordinate to $\sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}e_1$ and map latent token j to $\widehat{\ell}^{\mathrm{post}}_{j,n}e_1$ in key space. The scaled dot product is therefore exactly $\widehat{\ell}^{\mathrm{post}}_{j,n}$. Map the remaining latent coordinates to $(\widehat{\mathsf{a}}_j, \widehat{\mathsf{b}}_j)$ and inject the weighted attention output into $\mathsf{V}_{\mathrm{out}}$. This proves (69) and every row of Table 1. The null-token value is zero for all four value projections of the constructed network, so unlisted rows receive zero attention updates. The direct-sum dimensions add to (67), and the displayed coefficient bound follows from the coordinate maps, selector threshold, posterior-logit multiplier, and scaling used in the dot product. □

E.2 Component Approximation and Stability Bounds

Fix the common parameters in Assumption 3.1 and the regular grid, diffusion interval, clipping radii, and component tolerances of Definition E.1. Take $\varepsilon_{\mathrm{floor},n}, \varepsilon_{\ell,n}, \varepsilon_{a,n}, \varepsilon_{b,n}, \varepsilon_{\mathrm{rat},n} \in (0,1]$ and $0 < R_{\mathrm{out},n} \le R_{\mathrm{bal},n}$; the query tolerance belongs to $(0,1]$ for input–output prompts and is zero for sample-only prompts. The following error quantities for the component networks are used by both approximation analyses. For the tolerances in Definition E.1, set

$$\begin{aligned}
\mathsf{C}_{\mathrm{mix},n} &:= \frac{R^{\mathrm{lat}}_{\mathrm{num},n}}{\underline{q}_{\mathrm{out}}} + \frac{R^{\mathrm{lat}}_{\mathrm{num},n}\overline{q}_{\mathrm{out}}}{\underline{q}^2_{\mathrm{out}}},\\
\mathsf{C}_{\mathrm{net},n} &:= 1 + 2\mathsf{C}_{\mathrm{mix},n},\\
\varepsilon_{\mathrm{net},n} &:= 2\mathsf{C}_{\mathrm{mix},n}(n\varepsilon_{\ell,n} + \varepsilon_{\mathrm{qry},n}) + \frac{\varepsilon_{b,n}}{\underline{q}_{\mathrm{out}}} + \frac{R^{\mathrm{lat}}_{\mathrm{num},n}\varepsilon_{a,n}}{\underline{q}^2_{\mathrm{out}}} + \varepsilon_{\mathrm{rat},n}.
\end{aligned}$$

The constant $\mathsf{C}_{\mathrm{mix},n}$ controls the effect of perturbing posterior weights on the score ratio.

E.2.1 Attention stability and ReLU approximation

For $d, m \in \mathbb{N}$ and a query token $v \in \mathbb{R}^d$, key–value tokens $x_1, \dots, x_m \in \mathbb{R}^d$, and matrices $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d\times d}$, define the attention block (Vaswani et al., 2017)

$$\mathrm{Att}_{W_Q,W_K,W_V,W_O}(v; x_{1:m}) := W_O \sum_{j=1}^{m} \pi_j W_V x_j, \qquad \pi_j := \frac{\exp\{(W_Q v)^\top W_K x_j/\sqrt{d}\}}{\sum_{k=1}^{m}\exp\{(W_Q v)^\top W_K x_k/\sqrt{d}\}}.$$

Lemma E.2 (Sensitivity of scaled dot-product attention)**.** *Fix $B \ge 1$. Suppose every entry of the tokens and matrices in this attention map and in a primed copy is bounded in absolute value by B. Let η_{att} be the maximum of the entrywise matrix differences and the ℓ_∞ distances between corresponding tokens. Then*

$$\left\|\mathrm{Att}_{W_Q,W_K,W_V,W_O}(v; x_{1:m}) - \mathrm{Att}_{W'_Q,W'_K,W'_V,W'_O}(v'; x'_{1:m})\right\|_\infty \le 16 d^5 B^6 \eta_{\mathrm{att}}.$$

The bound is independent of the number of keys m and remains valid after restricting both maps by the same deterministic nonempty attention mask.

The proof follows the standard perturbation argument used in self-attention stability analyses (Edelman et al., 2022; Takakura & Suzuki, 2023); we give the details because our version jointly perturbs the tokens and all four attention projections and must remain uniform in the number of masked keys.

Proof. The matrix inequality $\|Ax\|_\infty \le d\|A\|_{\max}\|x\|_\infty$ gives

$$\|W_Q v - W'_Q v'\|_\infty + \max_j \|W_K x_j - W'_K x'_j\|_\infty \le 4dB\eta_{\mathrm{att}}.$$

A two-term telescoping argument for the scaled dot products therefore gives

$$\max_j \left|\frac{(W_Q v)^\top W_K x_j - (W'_Q v')^\top W'_K x'_j}{\sqrt{d}}\right| \le 4d^{5/2}B^3\eta_{\mathrm{att}}.$$

The softmax map satisfies the following Lipschitzness (Edelman et al., 2022, Corollary A.7):

$$\|\operatorname{softmax}(a) - \operatorname{softmax}(a')\|_1 \le 2\|a - a'\|_\infty.$$

The weighted value average consequently changes by at most

$$\left\| \sum_j \pi_j W_V x_j - \sum_j \pi'_j W'_V x'_j \right\|_\infty \le \{8d^{7/2}B^5 + 2dB\}\eta_{\mathrm{att}}.$$

One final telescoping step through W_O yields

$$\|W_O - W'_O\|_{\max} \left\| \sum_j \pi_j W_V x_j \right\|_1 + dB \left\| \sum_j \pi_j W_V x_j - \sum_j \pi'_j W'_V x'_j \right\|_\infty \le 16 d^5 B^6 \eta_{\mathrm{att}}.$$

Applying the same argument to the restricted list of keys selected by the common mask proves the masked version. □

Lemma E.3 (Fixed-dimensional Hölder approximation by ReLU networks)**.** *Fix $d, d_{\mathrm{out}} \in \mathbb{N}$ and $\beta_0 \in (0,1]$. There is an exponent $0 < c < \infty$, depending only on $(d, d_{\mathrm{out}}, \beta_0)$, with the following property; all implicit constants below have the same allowed dependencies. Let $M_f, H \ge 0$, let $\varepsilon \in (0,1]$, and let $f : [-1,1]^d \to [-M_f, M_f]^{d_{\mathrm{out}}}$ satisfy*

$$\|f(x) - f(x')\|_\infty \le H \|x - x'\|_\infty^{\beta_0}.$$

There are integers $D_\varepsilon^{\mathrm{NN}}, W_\varepsilon, S_\varepsilon \ge 1$, a bound $B_\varepsilon \ge 1$, and a fixed-architecture class

$$\mathscr{N}_\varepsilon \subseteq \mathrm{Proj}_{[-M_f, M_f]^{d_{\mathrm{out}}}} \circ \mathcal{NN}(D_\varepsilon^{\mathrm{NN}}, W_\varepsilon, S_\varepsilon, B_\varepsilon),$$

chosen independently of f, that contains a member f_ε satisfying

$$\sup_{x \in [-1,1]^d} \|f_\varepsilon(x) - f(x)\|_\infty \le \varepsilon.$$

The network sizes satisfy

$$D_\varepsilon^{\mathrm{NN}} \vee W_\varepsilon \vee S_\varepsilon \lesssim \left(1 + \frac{H \vee 1}{\varepsilon}\right)^{d/\beta_0} \log^2\left(2 + \frac{1 + M_f + H}{\varepsilon}\right), \tag{72}$$

Every scalar weight and bias has magnitude at most B_ε. The tensor-hat and multiplication subnetworks are fixed, and only the output-layer interpolation coefficients vary over a compact cube. After adding functionally inactive coordinates, write the class as

$$\mathscr{N}_\varepsilon = \{f_\vartheta : \vartheta \in [-B_\varepsilon, B_\varepsilon]^{S_\varepsilon}\}.$$

The evaluation map $(x, \vartheta) \mapsto f_\vartheta(x)$ is jointly Borel and continuous in ϑ. There are constants $\mathsf{Lip}_{\mathrm{par}}, \mathsf{Lip}_{\mathrm{in}} \in [0, \infty)$ such that

$$\sup_x \|f_\vartheta(x) - f_{\vartheta'}(x)\|_\infty \le \mathsf{Lip}_{\mathrm{par}} \|\vartheta - \vartheta'\|_\infty,$$
$$\sup_{\vartheta \in [-B_\varepsilon, B_\varepsilon]^{S_\varepsilon}} \|f_\vartheta(x) - f_\vartheta(x')\|_\infty \le \mathsf{Lip}_{\mathrm{in}} \|x - x'\|_\infty,$$

and

$$B_\varepsilon + \mathsf{Lip}_{\mathrm{par}} + \mathsf{Lip}_{\mathrm{in}} \lesssim (1 + M_f + H + \varepsilon^{-1})^c. \tag{73}$$

Proof. Affinely rescale the input cube to $[0,1]^d$ and take a regular tensor grid with mesh

$$\eta_{\mathrm{grid}} := \min\left\{1, c_d \left(\frac{\varepsilon}{H \vee 1}\right)^{1/\beta_0}\right\},$$

where $c_d > 0$ is chosen below. If $u_1, \dots, u_{J_{\mathrm{hat}}}$ are the grid points and $\psi_1, \dots, \psi_{J_{\mathrm{hat}}}$ the tensor-hat partition of unity, then $J_{\mathrm{hat}} \le C_d\{1 + (H \vee 1)/\varepsilon\}^{d/\beta_0}$. The interpolant

$$I_{\eta_{\mathrm{grid}}} f(x) := \sum_{j=1}^{J_{\mathrm{hat}}} f(u_j)\psi_j(x)$$

satisfies $\|I_{\eta_{\mathrm{grid}}} f - f\|_\infty \le H(C_d \eta_{\mathrm{grid}})^{\beta_0} \le \varepsilon/2$ after decreasing c_d.

Every univariate hat is an exact ReLU function. Use the ReLU multiplication construction in Yarotsky (2017, Proposition 3 and its proof) on a fixed bounded interval, recursively approximate each product of d univariate hats, and project the result onto $[0,1]$. With

$$\varepsilon_{\mathrm{mult}} := \frac{\varepsilon}{2J_{\mathrm{hat}}(M_f \vee 1)},$$

each approximate tensor hat $\widehat{\psi}_j$ can be chosen so that $\|\widehat{\psi}_j - \psi_j\|_\infty \le \varepsilon_{\mathrm{mult}}$. Consequently,

$$\left\| \sum_{j=1}^{J_{\mathrm{hat}}} f(u_j)\widehat{\psi}_j - I_{\eta_{\mathrm{grid}}} f \right\|_\infty \le J_{\mathrm{hat}} M_f \varepsilon_{\mathrm{mult}} \le \varepsilon/2.$$

The multiplication paths have fixed width and depth $\mathcal{O}_d\big(\log(2+\varepsilon_{\mathrm{mult}}^{-1})\big)$. Parallelizing the J_{hat} subnetworks gives a fixed wiring pattern whose depth, width, and total number of scalar weights and biases obey (72); the extra logarithm allows all coordinates to share one network graph. Fix every hat and multiplication weight at the value used in these J_{hat} approximate basis functions. For a coefficient array $\mathbf{a} = (a_{kj})_{1\le k\le d_{\mathrm{out}}, 1\le j\le J_{\mathrm{hat}}}$, define

$$\widetilde{f}_{\mathbf{a}}(x) := \left(\sum_{j=1}^{J_{\mathrm{hat}}} a_{kj}\widehat{\psi}_j(x) \right)_{k=1}^{d_{\mathrm{out}}}, \qquad f_{\mathbf{a}} := \mathrm{Proj}_{[-M_f,M_f]^{d_{\mathrm{out}}}} \circ \widetilde{f}_{\mathbf{a}}.$$

Choose B_ε to bound the fixed network weights, the hat slopes, and the coefficient range $[-B_\varepsilon, B_\varepsilon]^{d_{\mathrm{out}} J_{\mathrm{hat}}}$, with $B_\varepsilon \ge M_f$. Only $\mathbf{a}$ is variable; all internal network parameters are fixed and therefore are not trainable. Pad $\mathbf{a}$ by inactive coordinates to obtain a parameter vector $\vartheta \in [-B_\varepsilon, B_\varepsilon]^{S_\varepsilon}$. The resulting class and its fixed wiring depend only on $(d, d_{\mathrm{out}}, \beta_0, M_f, H, \varepsilon)$ and not on f. Taking $a_{kj} = f_k(u_j)$ gives the approximant established above.

Because $0 \le \widehat{\psi}_j \le 1$ and Euclidean projection onto the box is nonexpansive,

$$\sup_x \|f_{\mathbf{a}}(x) - f_{\mathbf{a}'}(x)\|_\infty \le J_{\mathrm{hat}} \|\mathbf{a} - \mathbf{a}'\|_\infty.$$

Thus one may take $\mathsf{Lip}_{\mathrm{par}} = J_{\mathrm{hat}}$. Each exact univariate hat is $C_d\eta_{\mathrm{grid}}^{-1}$-Lipschitz. The multiplication paths have fixed width, uniformly bounded weights, and depth $\mathcal{O}_d(\log(2+\varepsilon_{\mathrm{mult}}^{-1}))$. Write $\mathrm{Lip}_\infty(g)$ for the Lipschitz constant of a map g with respect to the ℓ_∞ norm. For a constant $\kappa_d > 0$ depending only on d, the layerwise operator-norm bound gives

$$\max_{1\le j\le J_{\mathrm{hat}}} \mathrm{Lip}_\infty(\widehat{\psi}_j) \le C_d \eta_{\mathrm{grid}}^{-1} (2+\varepsilon_{\mathrm{mult}}^{-1})^{\kappa_d} =: \mathsf{Lip}_{\psi,\varepsilon}.$$

Consequently, uniformly over the coefficient cube,

$$\|f_{\mathbf{a}}(x) - f_{\mathbf{a}}(x')\|_\infty \le J_{\mathrm{hat}} B_\varepsilon \mathsf{Lip}_{\psi,\varepsilon} \|x - x'\|_\infty.$$

Hence one may take $\mathsf{Lip}_{\mathrm{in}} = J_{\mathrm{hat}} B_\varepsilon \mathsf{Lip}_{\psi,\varepsilon}$. The quantities J_{hat}, $\eta_{\mathrm{grid}}^{-1}$, $\varepsilon_{\mathrm{mult}}^{-1}$, and B_ε are polynomial in $(1 + M_f + H + \varepsilon^{-1})$ because d is fixed. This proves (73). All remaining measurability and continuity claims follow from finite continuous composition. □

E.2.2 Time-local approximation and output clipping

The OU-basis functions vary most rapidly near the early-stopping time. We approximate them on overlapping time intervals before combining the local approximations with a fixed partition of unity. Separate tolerances for the denominator and numerator account for their different effects on the score ratio.

Lemma E.4 (Time-local OU-basis approximation)**.** *Under Assumption 3.1, use the constants from Lemma D.2, and fix $0 < t_0 < T$, the state radius, and $\varepsilon_{a,n}, \varepsilon_{b,n} \in (0,1]$. Define*

$$K := \max\{1, \lfloor \log_2(T/t_0) \rfloor\}, \qquad t_j^{\mathrm{knot}} := 2^j t_0 \ (0 \le j < K), \qquad t_K^{\mathrm{knot}} := T.$$

For $0 \le j \le K$, set

$$a_j := t^{\mathrm{knot}}_{\max\{0,j-1\}}, \qquad b_j := t^{\mathrm{knot}}_{\min\{K,j+1\}}, \qquad h_j^{\mathrm{loc}} := b_j - a_j.$$

Let $R_{\mathrm{geom},n} := 1 + R_{\mathrm{state},n} + \sqrt{d_z}$, and fix a sufficiently large constant C_{loc} depending only on the fixed dimensions, β', and the common density bounds. Apply the extension construction of Lemma D.3 on each interval, using the query and state normalization in (37) and affine normalization of the local time interval to $[-1, 1]$. The resulting denominator and numerator maps have respective β'-Hölder moduli bounded by

$$H_{a,j,n} := C_{\mathrm{loc}}\big[1 + H_{q,x} + R_{\mathrm{state},n} R_{\mathrm{geom},n} \sigma_{a_j}^{-2} + h_j^{\mathrm{loc}} R_{\mathrm{geom},n}^2 \sigma_{a_j}^{-4}\big], \tag{74}$$

$$\begin{aligned} H_{b,j,n} := C_{\mathrm{loc}}\big[1 &+ R_{\mathrm{bal},n} H_{q,x} + R_{\mathrm{state},n}\{\sigma_{a_j}^{-1} + R_{\mathrm{geom},n}^2 \sigma_{a_j}^{-3}\} \\ &+ h_j^{\mathrm{loc}}\{\sigma_{a_j}^{-1} + R_{\mathrm{geom},n}\sigma_{a_j}^{-3} + R_{\mathrm{geom},n}^3 \sigma_{a_j}^{-5}\}\big]. \end{aligned} \tag{75}$$

Write

$$H_{\max,n} := \max_{0 \le j \le K} (H_{a,j,n} \vee H_{b,j,n}),$$

and

$$L_{\mathrm{bas},n} := \log\left(2 + K + \frac{1 + R_{\mathrm{num},n}^{\mathrm{lat}} + H_{\max,n}}{\varepsilon_{a,n} \wedge \varepsilon_{b,n}}\right).$$

For each latent-grid point, there is a shared fixed-architecture ReLU class containing approximations with uniform errors at most $\varepsilon_{a,n}$ for a and $\varepsilon_{b,n}$ in Euclidean norm for b on $\mathcal{G}_n$. Across all $J_{\theta,n}$ points, the number of trainable parameters, number of stored nonzero coefficients, and evaluation cost are bounded by

$$\mathsf{N}_{\mathrm{basis},n}^{\mathrm{lat}} := C_{\mathrm{loc}} J_{\theta,n} L_{\mathrm{bas},n}^2 \sum_{j=0}^{K} \left[\left(1 + \frac{H_{a,j,n}}{\varepsilon_{a,n}}\right)^{(d_{\mathrm{stat}}+1)/\beta'} + d_z \left(1 + \frac{H_{b,j,n}}{\varepsilon_{b,n}}\right)^{(d_{\mathrm{stat}}+1)/\beta'}\right]. \tag{76}$$

Every member is projected onto $[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}]$ in its denominator coordinate and onto $[-R_{\mathrm{num},n}^{\mathrm{lat}} - 1, R_{\mathrm{num},n}^{\mathrm{lat}} + 1]$ in each numerator coordinate. Only output-layer interpolation coefficients are trainable; they range over a compact cube, with inactive padding coordinates allowed. The parameter radius and the input and parameter Lipschitz constants are polynomial in

$$1 + K + T + t_0^{-1} + R_{\mathrm{state},n} + R_{\mathrm{num},n}^{\mathrm{lat}} + H_{\max,n} + (\varepsilon_{a,n} \wedge \varepsilon_{b,n})^{-1}.$$

The class depends only on the displayed common quantities, not on the task map or task prior.

Proof. The covariance calculation in Section D.3.1, in particular (54), applies on every interval $[a_j, b_j]$ with σ_{t_0} replaced by σ_{a_j}. Affine time normalization multiplies the time derivative by $h_j^{\mathrm{loc}}/2$. The denominator has query modulus $H_{q,x}$, whereas the numerator has query modulus at most $R_{\mathrm{bal},n} H_{q,x}$. Combining these observations with the separate state and time derivatives proves (74) and (75).

Let $\chi_0, \dots, \chi_K$ be the continuous piecewise-linear nodal hats on the knots $t_0^{\mathrm{knot}}, \dots, t_K^{\mathrm{knot}}$, restricted to $[t_0, T]$. They are nonnegative, sum to one, and χ_j is supported on $[a_j, b_j]$. Each is an exact affine–ReLU function of the globally normalized time coordinate. The final interval includes any shorter remaining interval: when $K \ge 2$, consecutive knot gaps are at least t_0 and $h_j^{\mathrm{loc}} \le 7a_j$. When $K = 1$, both supports are the whole interval and normalization cancels $T - t_0$. Thus the hats and the local coordinates after coordinatewise projection

$$u_{t,j} := 2\min\{1, \max\{0, (t - a_j)/h_j^{\mathrm{loc}}\}\} - 1$$

have coefficients and input Lipschitz constants polynomial in $1 + T + t_0^{-1}$ when expressed in the globally normalized input.

Consider one scalar function f, with range bound $B := 1 + \overline{q}_{\mathrm{out}} + R_{\mathrm{num},n}^{\mathrm{lat}}$ and desired error ε. Use $\varepsilon = \varepsilon_{a,n}$ for the denominator and $\varepsilon = \varepsilon_{b,n}/\sqrt{d_z}$ for each numerator coordinate. On the jth normalized cube, construct the tensor-hat interpolant $I_j f$ from the proof of Lemma E.3, with error at most $\varepsilon/2$. Let $H_{f,j,n}$ be $H_{a,j,n}$ for the denominator and $H_{b,j,n}$ for a numerator coordinate. Its number of grid points is $\lesssim (1 + H_{f,j,n}/\varepsilon)^{(d_{\mathrm{stat}}+1)/\beta'}$. Write its tensor hats as $\psi_{j,r}$ and its nodal coefficients as $c_{j,r}$. Then

$$If := \sum_{j,r} c_{j,r} \chi_j(t) \psi_{j,r}(u_x, u_z, u_{t,j})$$

has error at most $\varepsilon/2$: every nonzero χ_j uses a local interpolant valid at that time, and their weights sum to one.

Let J_f^{loc} be the total number of these local grid points. Each product $\chi_j\psi_{j,r}$ has $d_{\mathrm{stat}}+2$ bounded univariate factors. Use the fixed multiplication paths in the proof of Lemma E.3 and project their outputs onto $[0,1]$. Approximate each product to accuracy $\varepsilon/\{2J_f^{\mathrm{loc}}(B\vee 1)\}$. The accumulated error after weighting by the interpolant coefficients is at most $\varepsilon/2$. Only $c_{j,r}$ vary, and the final coordinatewise projections cannot increase error. The number of coefficients and operations is $\lesssim J_f^{\mathrm{loc}}\log^2(2+J_f^{\mathrm{loc}}B/\varepsilon)$. Summing the coordinates and latent-grid points gives (76) after absorbing fixed dimension factors.

For arbitrary trainable coefficients, every approximate hat remains in $[0,1]$, so parameter sensitivity is at most J_f^{loc} before the final nonexpansive projection. The univariate hat slopes and time-coordinate maps have polynomial bounds. The multiplication paths have fixed width, bounded weights, and logarithmic depth; their input Lipschitz constants are therefore polynomial in the displayed quantities, by the same layerwise bound used in (73). Summing J_f^{loc} paths with bounded output-layer coefficients preserves polynomial input sensitivity. Running the component networks in parallel and adding inactive coordinates yields the stated common compact parameter cube. All computations are continuous, so evaluation is jointly Borel and continuous in the parameters. □

Lemma E.5 (Clipping error for the scaled score)**.** *For a deterministic radius $R_{\mathrm{out},n}>0$, set*

$$\varepsilon^2_{\mathrm{out,clip},n} := A_{t_0,T}\mathbb{E}(\|\xi\|_2 - R_{\mathrm{out},n})_+^2, \qquad \xi\sim\mathcal{N}(0,I_{d_z}). \tag{77}$$

Under the episodic OU model,

$$\mathbb{E}_{H_n}\int_{t_0}^{T}\frac{\mathbb{E}_{Z_t|H_n}\|\widetilde{s}^\star(Z_t,t\mid H_n)-\mathrm{clip}_{R_{\mathrm{out},n}}\widetilde{s}^\star(Z_t,t\mid H_n)\|_2^2}{\sigma_t^2}\,\mathrm{d}t \le \varepsilon^2_{\mathrm{out,clip},n}. \tag{78}$$

If $R_{\mathrm{out},n}\ge\sqrt{d_z}$, then

$$\varepsilon^2_{\mathrm{out,clip},n} \le 2A_{t_0,T}\exp\Big\{-\tfrac{1}{2}(R_{\mathrm{out},n}-\sqrt{d_z})^2\Big\}. \tag{79}$$

Proof. The denoising identity gives $\widetilde{s}^\star(Z_t,t\mid H_n) = -\mathbb{E}[\xi\mid Z_t,H_n]$. The squared distance to a closed convex ball is convex. Conditional Jensen therefore bounds its expectation at this conditional mean by $\mathbb{E}(\|\xi\|_2-R_{\mathrm{out},n})_+^2$. Integration against $\sigma_t^{-2}\mathrm{d}t$ proves (78). More generally, Gaussian norm concentration gives, for every $R\ge\sqrt{d_z}$ and $r=R-\sqrt{d_z}$,

$$\mathbb{E}(\|\xi\|_2-R)_+^2 \le \int_0^\infty 2ue^{-(r+u)^2/2}\,\mathrm{d}u \le 2e^{-r^2/2}. \tag{80}$$

Taking $R=R_{\mathrm{out},n}$ proves (79). □

E.2.3 Prior logits and stability of the score ratio

Define

$$w^{\mathrm{floor}}_{j,n} := \frac{w_{j,n}+\varepsilon_{\mathrm{floor},n}/J_{\theta,n}}{1+\varepsilon_{\mathrm{floor},n}}, \qquad \ell^{\mathrm{prior}}_{j,n} := \log\{J_{\theta,n}w^{\mathrm{floor}}_{j,n}\}.$$

Here $\mathbf{w}_n=(w_{j,n})_{j=1}^{J_{\theta,n}}$ and $\mathbf{w}_n^{\mathrm{floor}}=(w^{\mathrm{floor}}_{j,n})_{j=1}^{J_{\theta,n}}$. Write $\boldsymbol{\omega}_n(h)$ and $\boldsymbol{\omega}_n^{\mathrm{floor}}(h)$ for the posterior-weight vectors obtained from these two prior vectors, respectively. This is the specialization of the generic prior-floor construction in Proposition D.4 to $\zeta=\varepsilon_{\mathrm{floor},n}$; thus $w^{\mathrm{floor}}_{j,n}=w^{\zeta}_{j,n}$ and $\omega^{\mathrm{floor}}_{j,n}=\omega^{\zeta}_{j,n}$ at this value of ζ. As usual, let $x_\star$ denote the query component of h, with $x_\star=0$ in the sample-only prompt model. Define the score obtained from these modified prior weights by

$$\widetilde{s}^{\mathrm{floor}}_n(z,t\mid h) := \frac{\sum_j \omega^{\mathrm{floor}}_{j,n}(h)\mathsf{b}_{\theta_{j,n}}(x_\star,z,t)}{\sum_j \omega^{\mathrm{floor}}_{j,n}(h)\mathsf{a}_{\theta_{j,n}}(x_\star,z,t)}.$$

Because

$$\frac{\varepsilon_{\mathrm{floor},n}}{2J_{\theta,n}} \le w^{\mathrm{floor}}_{j,n} \le 1,$$

one has

$$\frac{\varepsilon_{\mathrm{floor},n}}{2} \le J_{\theta,n} w_{j,n}^{\mathrm{floor}} \le J_{\theta,n}.$$

Thus

$$|\ell_{j,n}^{\mathrm{prior}}| \le \log\left(\frac{2J_{\theta,n}}{\varepsilon_{\mathrm{floor},n}}\right). \tag{81}$$

For two probability vectors ω, ω', the corresponding score ratios satisfy

$$\left\| \frac{\sum_j \omega_j \mathsf{b}_j}{\sum_j \omega_j \mathsf{a}_j} - \frac{\sum_j \omega'_j \mathsf{b}_j}{\sum_j \omega'_j \mathsf{a}_j} \right\|_2 \le \mathsf{C}_{\mathrm{mix},n} \|\omega - \omega'\|_1.$$

This follows by bounding the numerator change by $R_{\mathrm{num},n}^{\mathrm{lat}} \|\omega - \omega'\|_1$, the denominator change by $\overline{q}_{\mathrm{out}} \|\omega - \omega'\|_1$, and both denominators below by $\underline{q}_{\mathrm{out}}$.

E.2.4 Neural approximation of the finite-mixture score

Apply Lemma E.3 to the $J_{\theta,n}$ context-example log-likelihood maps and, when present, the query log-likelihood maps, using the common ranges and moduli in Lemma D.2. Use Lemma E.4 for the OU-basis coordinates, with denominator tolerance $\varepsilon_{a,n}$ and Euclidean numerator tolerance $\varepsilon_{b,n}$. Apply exact coordinatewise ReLU projection onto $[-B_{\ell,\mathrm{ctx}}, B_{\ell,\mathrm{ctx}}]$ for every context-example log-likelihood output and onto $[-B_{\ell,\mathrm{qry}}, B_{\ell,\mathrm{qry}}]$ for every query log-likelihood output. These intervals contain the corresponding log-likelihood functions, so projection cannot enlarge their approximation errors and gives a common range for the entire score class. The time-local basis class allocates scalar numerator tolerances $\varepsilon_{b,n}/\sqrt{d_z}$, with fixed dimension factors absorbed in $\mathsf{N}_{\mathrm{basis},n}^{\mathrm{lat}}$. Project every approximate a-coordinate onto $[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}]$ using its exact scalar ReLU representation. Do not radially project the approximate b-vectors inside the ReLU–attention network. Instead, project their coordinates onto

$$[-R_{\mathrm{num},n}^{\mathrm{lat}} - 1, R_{\mathrm{num},n}^{\mathrm{lat}} + 1],$$

which again cannot enlarge the approximation error because every exact numerator coordinate lies in the smaller interval with radius $R_{\mathrm{num},n}^{\mathrm{lat}}$. At the reference coefficients, their approximation guarantee and $\varepsilon_{b,n} \le 1$ give

$$\|\widehat{\mathsf{b}}_j\|_2 \le R_{\mathrm{num},n}^{\mathrm{lat}} + 1.$$

Extend the ratio map to the enclosing rectangular box

$$[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}] \times [-R_{\mathrm{num},n}^{\mathrm{lat}} - 1, R_{\mathrm{num},n}^{\mathrm{lat}} + 1]^{d_z}$$

by

$$(a, b) \longmapsto \frac{b}{\min\{\overline{q}_{\mathrm{out}}, \max\{\underline{q}_{\mathrm{out}}, a\}\}},$$

and affinely rescale this box to a unit cube. After affine normalization, this map has range bounded by

$$1 + \frac{R_{\mathrm{num},n}^{\mathrm{lat}} + 1}{\underline{q}_{\mathrm{out}}}$$

and a Lipschitz constant bounded, up to a fixed dimension factor, by

$$(1 + \overline{q}_{\mathrm{out}})(2 + R_{\mathrm{num},n}^{\mathrm{lat}})(1 + \underline{q}_{\mathrm{out}}^{-2}).$$

It agrees with b/a on the specified ball domain and satisfies the range and Lipschitz bounds used in $\mathsf{N}_{\mathrm{rat},n}^{\mathrm{lat}}$. Apply Lemma E.3 to each output coordinate with scalar tolerance $\varepsilon_{\mathrm{rat},n}/\sqrt{d_z}$, so its vector error is at most $\varepsilon_{\mathrm{rat},n}$ in Euclidean norm.

Put $\varepsilon_{\mathrm{logit},n} := n\varepsilon_{\ell,n} + \varepsilon_{\mathrm{qry},n}$. Write $\widehat{\ell}_j(c)$ for the jth context-example log-likelihood-network output, set $\overline{\widehat{\ell}}_j := n^{-1} \sum_{i=1}^n \widehat{\ell}_j(c_i)$, and write $\widehat{\ell}_j^{\mathrm{qry}}(x_\star)$ for the corresponding query log-likelihood-network

output. In the sample-only prompt model, use the convention $\widehat{\ell}_j^{\mathrm{qry}}(0) := 0$. For the modified prior weights, the exact quadrature logits and their neural approximations are

$$\ell_{j,n}^{\mathrm{post}}(h) := \ell_{j,n}^{\mathrm{prior}} + \sum_{i=1}^{n} \ell_{\theta_{j,n}}(c_i) + \ell_{\theta_{j,n}}^{\mathrm{qry}}(x_\star),$$
$$\widehat{\ell}_{j,n}^{\mathrm{post}}(h) := \ell_{j,n}^{\mathrm{prior}} + n\bar{\widehat{\ell}}_j + \widehat{\ell}_j^{\mathrm{qry}}(x_\star).$$

Their uniform coordinatewise difference is at most $\varepsilon_{\mathrm{logit},n}$. The softmax map satisfies

$$\|\operatorname{softmax}(a) - \operatorname{softmax}(a')\|_1 \le 2\|a - a'\|_\infty.$$

Hence the posterior-weight error is at most $2\varepsilon_{\mathrm{logit},n}$. The corresponding weighted numerator and denominator errors are at most

$$2R_{\mathrm{num},n}^{\mathrm{lat}}\varepsilon_{\mathrm{logit},n} + \varepsilon_{b,n}$$

and

$$2\overline{q}_{\mathrm{out}}\,\varepsilon_{\mathrm{logit},n} + \varepsilon_{a,n},$$

respectively. Both approximate denominators remain in $[\underline{q}_{\mathrm{out}}, \overline{q}_{\mathrm{out}}]$. The same decomposition of the difference between two ratios used above, followed by the ratio-readout error, gives

$$\frac{2R_{\mathrm{num},n}^{\mathrm{lat}}\varepsilon_{\mathrm{logit},n} + \varepsilon_{b,n}}{\underline{q}_{\mathrm{out}}} + \frac{R_{\mathrm{num},n}^{\mathrm{lat}}\{2\overline{q}_{\mathrm{out}}\,\varepsilon_{\mathrm{logit},n} + \varepsilon_{a,n}\}}{\underline{q}_{\mathrm{out}}^2} + \varepsilon_{\mathrm{rat},n} \le \varepsilon_{\mathrm{net},n}. \tag{82}$$

Masked Transformer implementation. Set

$$B_{\mathrm{gate},n}^{\mathrm{lat}} := 1 + \log\left(\frac{2J_{\theta,n}}{\varepsilon_{\mathrm{floor},n}}\right) + nB_{\ell,\mathrm{ctx}} + B_{\ell,\mathrm{qry}} + \overline{q}_{\mathrm{out}} + R_{\mathrm{num},n}^{\mathrm{lat}}. \tag{83}$$

Every coordinate used by the selector is bounded by (83); the preliminary projection in $\mathsf{R}_{3,n}$ is therefore inactive at the reference parameter. Apply Lemma E.1 with $J = J_{\theta,n}$, the coordinatewise projections of the component-network outputs constructed above, and the logits of the modified prior weights. The lemma specifies the exact token width, initial token shapes, attention masks, projections, residual destinations, and selector network. It follows from (69) that the output token contains the approximate posterior-weighted denominator and numerator exactly. The fixed ReLU ratio readout acts on the output token, and the fixed radial output projection completes the score map. Deterministic nonempty masks, fixed role and identifier codes, and fixed selector and ratio networks introduce no trainable coordinates. The context-example and query log-likelihood, OU-basis, prior-logit, and attention-projection coordinates all range over declared compact cubes and are jointly trainable. Their product is $\mathcal{V}_n^{\mathrm{fit}}$ in Definition E.1, and coordinatewise projection from the enclosing parameter cube gives exactly the class representation in (66). The constructed approximator coefficients, logits of the modified prior weights, and the four sets of projection matrices in Table 1 define a point of this product whose output is the constructed score. Its finite-mixture error is bounded by (82). The architectures of the approximators and the radii of their parameter cubes depend only on the shared bounds, dimensions, grid, and tolerances, so the same class applies to every task map and task prior satisfying the assumptions.

Consequently, the constructed output f before the final radial projection satisfies $\|f - \widetilde{s}_n^{\mathrm{floor}}\|_2 \le \varepsilon_{\mathrm{net},n}$ on $\mathcal{G}_n$. This estimate compares the network with the discretized score computed from the modified prior weights; it does not compare that score with the original posterior predictive. The integrated approximation guarantee below uses this same construction.

E.3 Integrated Transformer Approximation Guarantee

Proposition E.6 (Integrated score approximation by the Transformer class)**.** *Use the regular grid and cell-midpoint map from Section D.3, set $\eta = \eta_{\theta,n}$ and $\zeta = \varepsilon_{\mathrm{floor},n} > 0$ in Proposition D.4, and use the class in Definition E.1. Let $\varepsilon_{\mathrm{net},n}$ be the component-error expression in Section E.2. Then*

$$\mathcal{E}_{\mathrm{app}}(\mathcal{S}_n^{\mathrm{lat}}) \le 4D_n(\eta_{\theta,n}, \varepsilon_{\mathrm{floor},n}) + 4A_{t_0,T}\varepsilon_{\mathrm{net},n}^2 + 2\varepsilon_{\mathrm{out,clip},n}^2 + 2A_{t_0,T}R_{\mathrm{out},n}^2\delta_{\mathrm{tail},n} + 2\varepsilon_{\mathrm{tail},n}^2. \tag{84}$$

The parameters of the approximating network may depend on the task family and prior. Its architecture, coordinate bounds, and approximation guarantee depend only on the common bounds, grid, and tolerances.

Proof. By (60), the construction in Section E.2.4, in particular (82), gives an output f before the final radial projection satisfying $\|f - \sigma_t s^{\eta_{\theta,n},\zeta}\|_2 \le \varepsilon_{\mathrm{net},n}$ on $\mathcal{G}_n$. No uniform comparison of $s^{\eta_{\theta,n},\zeta}$ with $s^\star$ is used here. Recall that $\widetilde{s}^\star = \sigma_t s^\star$, and let P be radial projection onto the ball of radius $R_{\mathrm{out},n}$. On $\mathcal{G}_n$, nonexpansiveness and two applications of the squared triangle inequality give

$$\|\mathsf{P}f - \widetilde{s}^\star\|_2^2 \le 4\varepsilon_{\mathrm{net},n}^2 + 4\sigma_t^2\|s^{\eta_{\theta,n},\zeta} - s^\star\|_2^2 + 2\|\mathsf{P}\widetilde{s}^\star - \widetilde{s}^\star\|_2^2.$$

Integrate against $\sigma_t^{-2}P_{t|H_n}\,\mathrm{d}t$ and the original prompt law. The first term integrates to at most $4A_{t_0,T}\varepsilon_{\mathrm{net},n}^2$, the second is bounded by (58), and the third by Lemma E.5. On $\mathcal{G}_n^c$, use $\|\mathsf{P}f - \widetilde{s}^\star\|^2 \le 2R_{\mathrm{out},n}^2 + 2\|\widetilde{s}^\star\|^2$ and the two bounds in Lemma D.1. This proves (84). □

For a desired integrated score error of order ε^2, it is enough to take $\eta_{\theta,n}^{\beta_\theta}$ of order $\varepsilon/\sqrt{n+1}$, prior floor of order ε^2, and neural scaled-score error of order $\varepsilon/\sqrt{A_{t_0,T}}$. The conclusion is averaged under the true episode law, which is the criterion used by the score-learning and sampling analyses; it does not assert a pointwise approximation guarantee for every prompt.

E.4 Class Complexity: Parameters, Sensitivity, and Covering

Proposition E.6 constructs one accurate parameter vector for each task family and task prior. Training can select any vector in the class, so the learning analysis requires a bound on output sensitivity throughout the parameter cube, not just near that approximator. The next proposition gives this bound and the resulting covering number, together with counts of trainable parameters, stored nonzero scalar coefficients, and sparse arithmetic operations in one forward pass. Masks, role labels, latent-cell indices, and their bit encodings are excluded from scalar-coefficient storage.

Proposition E.7 (Transformer size and covering numbers)**.** *Under Assumption 3.1, fix any regular latent grid and component tolerances satisfying Definition E.1. Let* $\mathsf{N}_{\mathrm{param},n}^{\mathrm{lat}}$ *count trainable scalar coordinates, let* $\mathsf{N}_{\mathrm{store},n}^{\mathrm{lat}}$ *count stored nonzero scalar coefficients under this convention, and let* $\mathsf{N}_{\mathrm{op},n}^{\mathrm{lat}}$ *count sparse arithmetic operations in one forward evaluation. With the indexing in Definition E.1, the score class has the compact trainable-parameter representation*

$$\mathcal{S}_n^{\mathrm{lat}} = \{s_{\mathbf{v}} : \mathbf{v} \in [-B_{\mathrm{lat},n}, B_{\mathrm{lat},n}]^{\mathsf{N}_{\mathrm{param},n}^{\mathrm{lat}}}\}.$$

There is a deterministic constant $\mathsf{Lip}_{\mathrm{lat},n}^{\mathrm{par}}$ *such that*

$$d_{\mathrm{bal},\infty}(s_{\mathbf{v}}, s_{\mathbf{v}'}) \le \mathsf{Lip}_{\mathrm{lat},n}^{\mathrm{par}}\|\mathbf{v} - \mathbf{v}'\|_\infty. \tag{85}$$

This bound holds for all parameter vectors $\mathbf{v}$ *and* $\mathbf{v}'$ *in the displayed cube. Consequently, for every* $\varepsilon_{\mathrm{cov}} > 0$*,*

$$\log N_{\mathrm{cov}}\left(\varepsilon_{\mathrm{cov}}, \mathcal{S}_n^{\mathrm{lat}}, d_{\mathrm{bal},\infty}\right) \le \mathsf{N}_{\mathrm{param},n}^{\mathrm{lat}} \log\left(1 + \frac{2B_{\mathrm{lat},n}\mathsf{Lip}_{\mathrm{lat},n}^{\mathrm{par}}}{\varepsilon_{\mathrm{cov}}}\right). \tag{86}$$

The empirical variance-balanced DSM loss admits a Borel-measurable exact ERM over this compact parameter cube. Explicit bounds for $\mathsf{N}_{\mathrm{param},n}^{\mathrm{lat}}$*,* $\mathsf{N}_{\mathrm{store},n}^{\mathrm{lat}}$*, and* $\mathsf{N}_{\mathrm{op},n}^{\mathrm{lat}}$ *are given in* (87) *to* (89)*. These conclusions hold uniformly over the task maps in the fixed regularity class of Assumption 3.1 and every Borel task prior on* $\mathcal{T}$*.*

The bounds on the number of parameters, stored coefficients, and arithmetic operations use exact real arithmetic, counting each scalar arithmetic operation, comparison, exponential, logarithm, and square root at unit cost, and do not establish finite-precision stability or optimization runtime. For a larger trainable architecture, the covering analysis would need its full parameter count, radius, and sensitivity. The class is indexed by prompt length and approximation choices; no parameter sharing or extrapolation across prompt lengths is established.

E.4.1 Parameter counts and global parameter sensitivity

We count the construction's trainable coordinates, stored coefficients, and sparse operations, and then combine these bounds with global parameter sensitivity to obtain the covering-number bound used in the finite-episode learning analysis. The parameter choices for a target accuracy are deferred

to Section G.6. Section G.6 later specifies the grid and approximation tolerances for a prescribed predictive accuracy; the present bounds do not require those specific choices.

For fixed d and $\beta_0 \in (0,1]$, define

$$\mathfrak{C}_{d,\beta_0}(B,H,\varepsilon) := C_{d,\beta_0}\left(1+\frac{H\vee 1}{\varepsilon}\right)^{d/\beta_0}\log^2\left(2+\frac{1+B+H}{\varepsilon}\right).$$

This is the ReLU-network size needed for a fixed-dimensional β_0-Hölder approximation with range bound B, Hölder constant H, and approximation error ε. Set

$$\mathsf{N}^{\mathrm{lat}}_{\ell,n} := J_{\theta,n}\mathfrak{C}_{d_{\mathrm{stat}},\beta'}(B_{\ell,\mathrm{ctx}},H_{\ell,c},\varepsilon_{\ell,n}),$$

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{qry},n} := \begin{cases}0, & \text{sample-only prompt model},\\ J_{\theta,n}\mathfrak{C}_{d_x,\beta'}(B_{\ell,\mathrm{qry}},H_{\mathrm{qry},x},\varepsilon_{\mathrm{qry},n}), & \text{input–output prompt model},\end{cases}$$

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{rat},n} := d_z\mathfrak{C}_{d_z+1,1}\left(1+\frac{R^{\mathrm{lat}}_{\mathrm{num},n}+1}{\underline{q}_{\mathrm{out}}},(1+\overline{q}_{\mathrm{out}})(2+R^{\mathrm{lat}}_{\mathrm{num},n})(1+\underline{q}^{-2}_{\mathrm{out}}),\varepsilon_{\mathrm{rat},n}\right).$$

Use $\mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n}$ from (76); it counts all local hats, multiplication paths, interpolation coefficients, and final coordinatewise projections. In the sample-only prompt model, omit the query log-likelihood network. With the counting conventions of Proposition E.7, the bounds on the numbers of trainable parameters, stored coefficients, and arithmetic operations per forward pass are

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n} \lesssim (d^{\mathrm{lat}}_{\mathrm{emb},n})^2 + J_{\theta,n} + \mathsf{N}^{\mathrm{lat}}_{\ell,n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{qry},n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n}, \tag{87}$$

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{store},n} \lesssim (d^{\mathrm{lat}}_{\mathrm{emb},n})^2 + \mathsf{N}^{\mathrm{lat}}_{\ell,n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{qry},n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{rat},n}, \tag{88}$$

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{op},n} \lesssim n\mathsf{N}^{\mathrm{lat}}_{\ell,n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{qry},n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n} + (n+J_{\theta,n})(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 + \mathsf{N}^{\mathrm{lat}}_{\mathrm{rat},n}. \tag{89}$$

The parameter count includes inactive coordinates used to embed every component network into a common compact parameter cube. The storage and operation counts exclude structural zeros, masks, cell and role indices, and their bit encodings, but include every nonzero fixed or trainable scalar coefficient.

Lemma E.8 (Token bounds and parameter sensitivity of the four-attention-block network)**.** *Let $\mathbf{v},\mathbf{v}' \in \mathcal{V}^{\mathrm{fit}}_n$ collect trainable coordinates of the component networks, prior logits, and entries of the attention projection matrices. All token-array bounds below hold uniformly over $(h,z,t)\in\mathcal{G}_n$, and their common input arguments are suppressed. There are deterministic quantities $B_{0,n},\mathsf{Lip}_{0,n},B_{\mathrm{par},n}\ge 1$ and fixed-map constants $B_{\mathrm{rt},k,n},\mathsf{Lip}_{\mathrm{rt},k,n}\ge 1$, $1\le k\le 4$, such that*

$$\max_i \|\mathbf{E}^{(0)}_{n,i}(\mathbf{v})\|_\infty \le B_{0,n}, \tag{90}$$

$$\max_i \|\mathbf{E}^{(0)}_{n,i}(\mathbf{v})-\mathbf{E}^{(0)}_{n,i}(\mathbf{v}')\|_\infty \le \mathsf{Lip}_{0,n}\|\mathbf{v}-\mathbf{v}'\|_\infty, \tag{91}$$

and every trainable or fixed coefficient used by an attention block has magnitude at most $B_{\mathrm{par},n}$. Define recursively, for $1\le k\le 4$,

$$\overline{B}_{k-1,n} := \max\{1,B_{\mathrm{par},n},B_{k-1,n}\},$$

$$B_{k,n} := B_{\mathrm{rt},k,n}\left\{1+B_{k-1,n}+(d^{\mathrm{lat}}_{\mathrm{emb},n})^2\overline{B}^3_{k-1,n}\right\}, \tag{92}$$

$$\mathsf{Lip}_{k,n} := \mathsf{Lip}_{\mathrm{rt},k,n}\left[\mathsf{Lip}_{k-1,n}+16(d^{\mathrm{lat}}_{\mathrm{emb},n})^5\overline{B}^6_{k-1,n}(1+\mathsf{Lip}_{k-1,n})\right]. \tag{93}$$

Then, for every $0\le k\le 4$,

$$\max_i \|\mathbf{E}^{(k)}_{n,i}(\mathbf{v})\|_\infty \le B_{k,n}, \tag{94}$$

$$\max_i \|\mathbf{E}^{(k)}_{n,i}(\mathbf{v})-\mathbf{E}^{(k)}_{n,i}(\mathbf{v}')\|_\infty \le \mathsf{Lip}_{k,n}\|\mathbf{v}-\mathbf{v}'\|_\infty. \tag{95}$$

The quantities $B_{0,n}$, $\mathsf{Lip}_{0,n}$, $B_{\mathrm{par},n}$, $B_{\mathrm{rt},k,n}$, and $\mathsf{Lip}_{\mathrm{rt},k,n}$ are polynomial in the displayed component-network bounds, $d^{\mathrm{lat}}_{\mathrm{emb},n}$, $J_{\theta,n}$, $B^{\mathrm{lat}}_{\mathrm{gate},n}$, and $\log(2J_{\theta,n}/\varepsilon_{\mathrm{floor},n})$. If $\mathsf{Lip}_{\mathrm{out},n}$ is an input-Lipschitz constant of the fixed ratio readout followed by radial projection, then

$$d_{\mathrm{bal},\infty}(s_{\mathbf{v}},s_{\mathbf{v}'}) \le \mathsf{Lip}_{\mathrm{out},n}\mathsf{Lip}_{4,n}\|\mathbf{v}-\mathbf{v}'\|_\infty, \tag{96}$$

and $\mathsf{Lip}_{\mathrm{out},n}$ is polynomial in the same inputs and $\varepsilon^{-1}_{\mathrm{rat},n}$.

Proof. The likelihood-network bounds in (73), the time-local basis bounds in Lemma E.4, the projections of component outputs, and (81) give (90) and (91) and a polynomial choice of $B_{\mathrm{par},n}$. After each attention update, let $\mathsf{R}_{k,n}$ denote the fixed affine–ReLU coordinate-update map, with the selector from (71) included after the third update. The four coordinate-update networks have a universal depth bound D_{rt}, width at most $C_{d_z}(d^{\mathrm{lat}}_{\mathrm{emb},n})^2$, and coefficient magnitude at most

$$B_{\mathrm{rt},n} := 1 \vee n \vee B^{\mathrm{lat}}_{\mathrm{gate},n} \vee \sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}.$$

An affine layer of width at most w and coefficient bound B satisfies

$$1 + \|Wx + b\|_\infty \le 2w(B \vee 1)(1 + \|x\|_\infty), \qquad \|W(x - x')\|_\infty \le w(B \vee 1)\|x - x'\|_\infty.$$

ReLU is one-Lipschitz and does not increase the bound on token magnitudes, so iteration over the D_{rt} layers gives constants $B_{\mathrm{rt},k,n}, \mathsf{Lip}_{\mathrm{rt},k,n} \ge 1$ satisfying

$$\|\mathsf{R}_{k,n}(x)\|_\infty \le B_{\mathrm{rt},k,n}(1 + \|x\|_\infty),$$
$$\|\mathsf{R}_{k,n}(x) - \mathsf{R}_{k,n}(x')\|_\infty \le \mathsf{Lip}_{\mathrm{rt},k,n}\|x - x'\|_\infty.$$

More precisely, one may take both constants no larger than

$$\{2C_{d_z}(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 B_{\mathrm{rt},n}\}^{D_{\mathrm{rt}}}.$$

This proves the asserted polynomial dependence for the coordinate-update networks. If the tokens and attention matrices entering block k are bounded by $\overline{B}_{k-1,n}$, then its attention output is bounded coordinatewise by $(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 \overline{B}^3_{k-1,n}$. Adding the residual and applying $\mathsf{R}_{k,n}$ proves (92) and (94) by induction. At block k, the maximum perturbation among its tokens and trainable matrices is at most $(1 + \mathsf{Lip}_{k-1,n})\|\mathbf{v} - \mathbf{v}'\|_\infty$. The attention perturbation bound in Lemma E.2, followed by the residual update and the $\mathsf{Lip}_{\mathrm{rt},k,n}$-Lipschitz coordinate-update map, gives (93) and (95). The preliminary readout projection maps all inputs to the compact approximation box specified above and is nonexpansive. Thus (73) gives a global polynomial input-Lipschitz constant for the fixed ratio readout; radial output projection is also nonexpansive. The state-clipping identity (63) extends the token bounds from $\mathcal{G}_n$ to the full state domain. Absorbing the fixed conversion between ℓ_∞ and Euclidean norms into $\mathsf{Lip}_{\mathrm{out},n}$ proves (96). □

Proof of Proposition E.7. Work under Assumption 3.1, with the constants given in Lemma D.2. We first count the coefficients and operations in the attention blocks and component networks. We then use parameter sensitivity to obtain a covering bound and a measurable ERM, while tracking the dependence on the declared grid and component tolerances.

Model size and forward-pass cost. Exclude the component encoders and the fixed ratio readout, but include every nonzero fixed or trainable scalar coefficient in the four attention blocks and their coordinate-update maps. The four-attention-block network above can be stored using at most

$$16(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 + C_{d_z} d^{\mathrm{lat}}_{\mathrm{emb},n} \le C_{d_z}(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 \tag{97}$$

such coefficients and evaluated using at most

$$C_{d_z}(n + J_{\theta,n} + 4)(d^{\mathrm{lat}}_{\mathrm{emb},n})^2 \tag{98}$$

sparse arithmetic operations. Structural zeros, masks, role labels, latent-cell indices, and their bit encodings are excluded under the counting convention of Proposition E.7. Every fixed coordinate-update coefficient has magnitude at most

$$1 \vee n \vee B^{\mathrm{lat}}_{\mathrm{gate},n} \vee \sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}. \tag{99}$$

Each block has four $d^{\mathrm{lat}}_{\mathrm{emb},n} \times d^{\mathrm{lat}}_{\mathrm{emb},n}$ attention projections, so all four blocks contain at most $16(d^{\mathrm{lat}}_{\mathrm{emb},n})^2$ nonzero projection coefficients. The role and block injections, coordinate copies, residual writes, scalar projections, and posterior-logit affine sum use at most $C_{d_z} d^{\mathrm{lat}}_{\mathrm{emb},n}$ shared nonzero coefficients. For the selector, one copy of (71) is used for each of the $J_{\theta,n}$ candidate entries in each of the $d_z + 2$ selected likelihood-and-basis coordinates. It therefore uses at most $C_{d_z} J_{\theta,n} \le C_{d_z} d^{\mathrm{lat}}_{\mathrm{emb},n}$ shared nonzero coefficients. This proves (97).

For a forward evaluation, the number of operations needed to apply the attention projections in all four blocks is $\lesssim (n + J_{\theta,n} + 4)(d^{\mathrm{lat}}_{\mathrm{emb},n})^2$. The masks in Table 1 contain $n + 3J_{\theta,n}$ non-null query–key pairs, whose dot products and value aggregation require $\lesssim (n + J_{\theta,n})d^{\mathrm{lat}}_{\mathrm{emb},n}$ further operations. The shared sparse coordinate-update maps require at most $C_{d_z} d^{\mathrm{lat}}_{\mathrm{emb},n}$ operations per token row, including the $J_{\theta,n}(d_z + 2)$ selector gates on a latent row. Residual additions and fixed coordinate injections cost $\lesssim (n + J_{\theta,n} + 4)d^{\mathrm{lat}}_{\mathrm{emb},n}$ operations. Since $d^{\mathrm{lat}}_{\mathrm{emb},n} \geq 1$, these bounds imply (98). Among the fixed coordinate-update coefficients, only the factor n in the posterior logit and the projection and gate thresholds can exceed one. The trainable attention coefficients are bounded by $1 \vee \sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}$ by the parameter-space definition. This proves (99).

We now add the component networks to these attention-block counts. The trainable output-layer coefficients of the context-example log-likelihood, query log-likelihood, and OU-basis networks are bounded, up to a fixed dimension-dependent factor, by

$$\mathsf{N}^{\mathrm{lat}}_{\ell,n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{qry},n} + \mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n}.$$

The fixed nonzero internal coefficients of these three network families obey the same bound. The prior contributes $J_{\theta,n}$ trainable logits. By (97), the fixed selector and coordinate-update networks and the trainable attention matrices contribute $\lesssim (d^{\mathrm{lat}}_{\mathrm{emb},n})^2$ nonzero stored coordinates. The fixed ratio readout contributes at most $\mathsf{N}^{\mathrm{lat}}_{\mathrm{rat},n}$ nonzero stored coordinates. This proves (87) and (88). Evaluating the shared context-example network n times and the query networks and ratio readout once, and then applying (98), proves (89). In particular, only the summary row attends to all n context-example tokens, the two coordinate-copy stages use singleton-key rows, and the output row attends to the $J_{\theta,n}$ latent tokens; no dense quadratic all-token attention matrix is evaluated.

Covering numbers and measurable empirical minimization. To pass from the parameter count to a covering number, we bound the parameter radius and the sensitivity of the score to its parameters. The final posterior-attention logit uses the standard $(d^{\mathrm{lat}}_{\mathrm{emb},n})^{-1/2}$ scaled dot product, so one fixed query or key coordinate is multiplied by $\sqrt{d^{\mathrm{lat}}_{\mathrm{emb},n}}$; this factor is included in the fixed attention and coordinate-update radius. By construction, the radii of the parameter-space factors in Definition E.1 are the component-network parameter radii, the prior-logit radius in (81), and the attention-projection radius containing the reference matrices. Take $B_{\mathrm{lat},n}$ to be the largest of these radii. This choice is polynomial in the displayed inputs and $\log(\varepsilon^{-1}_{\mathrm{floor},n})$ by (73) and Lemma E.4. Because $\mathrm{Proj}^{\mathrm{par}}_n$ is nonexpansive in ℓ_∞, applying Lemma E.8 after this projection and setting $\mathsf{Lip}^{\mathrm{par}}_{\mathrm{lat},n} := \mathsf{Lip}_{\mathrm{out},n}\mathsf{Lip}_{4,n}$ proves the required bound. This proves (85) and makes the claimed polynomial dependence explicit. A grid contained in the compact trainable-parameter cube proves (86). The resulting finite network is jointly Borel in the data and continuous in its parameters. Since the parameter cube is a compact metric space and the empirical risk is Borel measurable in the training sample and continuous in the parameter (hence Carathéodory), the measurable maximum theorem (Aliprantis & Border, 2006, Theorem 18.19), applied to the negative empirical risk, yields a Borel-measurable parameter selector whose induced score is an exact empirical-risk minimizer; see also Brown & Purves (1973).

□

F Pretraining and Score Learning from Independent Episodes

This section turns the approximation class from Section E into a finite-episode learning guarantee. We first define clipped variance-balanced DSM and its population target, then establish uniform convergence in the scaled-score metric, and finally combine approximation, clipping, optimization, covering, and estimation errors for the Transformer class.

F.1 Clipped Labels and the Empirical bDSM Objective

The training sample consists of N independent episodes, each augmented with one independent diffusion time and Gaussian perturbation. Bounding the scaled-score output and clipping the Gaussian label yield a bounded squared loss. The resulting theorem separates approximation, clipping, optimization, covering, and estimation errors.

F.1.1 Population regression function and clipping bias

For a DSM-label clipping radius $R_{\mathrm{lab}} > 0$, let $\mathrm{clip}_{R_{\mathrm{lab}}}$ be the radial clipping map in (36). Define the clipped DSM label and a measurable version of its regression function by

$$\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} := \mathrm{clip}_{R_{\mathrm{lab}}}(-\xi_{\mathrm{dsm}}), \tag{100}$$

$$\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}(z,t,h) := \mathbb{E}\Big[\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} \,\Big|\, Z_\tau = z, \tau = t, H_n = h\Big]. \tag{101}$$

For a score s, define the clipped population risk and the label-clipping bias by

$$\mathcal{R}^{(R_{\mathrm{lab}})}_{\mathrm{bDSM}}(s) := A_{t_0,T}\mathbb{E}\Big[\|\widetilde{s}(Z_\tau,\tau,H_n) - \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}}\|_2^2\Big], \tag{102}$$

$$\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}) := A_{t_0,T}\mathbb{E}\big[\|\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau,\tau,H_n) - \widetilde{s}^{\star}(Z_\tau,\tau \,|\, H_n)\|_2^2\big]. \tag{103}$$

The Pythagorean identity for squared loss and the squared triangle inequality give

$$\mathcal{R}^{(R_{\mathrm{lab}})}_{\mathrm{bDSM}}(s) - A_{t_0,T}\mathbb{E}\Big[\|\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}} - \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}}\|_2^2\Big] = A_{t_0,T}\mathbb{E}\big[\|\widetilde{s} - \widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}\|_2^2\big],$$

$$\mathcal{E}_{\mathrm{score}}(s) \le 2A_{t_0,T}\mathbb{E}\big[\|\widetilde{s} - \widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}\|_2^2\big] + 2\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}). \tag{104}$$

Here, $\widetilde{s}$ and $\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}$ denote their evaluations at the random training input (Z_τ, τ, H_n). The term $\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}})$ measures the bias introduced by clipping the DSM label. This label clipping, together with the uniform bound imposed on the scaled-score predictor in (111), keeps the empirical squared loss uniformly bounded. This uniform loss bound uses the final output radius, rather than the potentially larger internal feature radius. It is separate from state clipping and from the score or scaled-score clipping built into the score network.

Proposition F.1 (Bias from clipping Gaussian DSM labels)**.** *For every $R_{\mathrm{lab}} > 0$,*

$$\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}) \le A_{t_0,T}\mathbb{E}(\|\xi\|_2 - R_{\mathrm{lab}})_+^2, \qquad \xi \sim \mathcal{N}(0, I_{d_z}). \tag{105}$$

Consequently, for every $x \ge 0$, taking $R_{\mathrm{lab}} \ge \sqrt{d_z} + x$ gives

$$\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}) \le 2A_{t_0,T}e^{-x^2/2}. \tag{106}$$

Proof of Proposition F.1. Let $\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} := \mathrm{clip}_{R_{\mathrm{lab}}}(-\xi_{\mathrm{dsm}})$. By Proposition C.1, $\widetilde{s}^{\star}(Z_\tau,\tau \,|\, H_n) = \mathbb{E}[-\xi_{\mathrm{dsm}} \,|\, Z_\tau, \tau, H_n]$, whereas $\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau,\tau,H_n) = \mathbb{E}[\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} \,|\, Z_\tau, \tau, H_n]$. Conditional Jensen's inequality therefore gives

$$\big\|\widetilde{s}^{\,\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau,\tau,H_n) - \widetilde{s}^{\star}(Z_\tau,\tau \,|\, H_n)\big\|_2^2 \le \mathbb{E}\Big[\|\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} + \xi_{\mathrm{dsm}}\|_2^2 \,|\, Z_\tau, \tau, H_n\Big].$$

For radial clipping, $\|\mathrm{clip}_R(u) - u\|_2 = (\|u\|_2 - R)_+$. Hence

$$\varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}) \le A_{t_0,T}\mathbb{E}(\|\xi_{\mathrm{dsm}}\|_2 - R_{\mathrm{lab}})_+^2,$$

which proves (105). The Gaussian tail moment bound (80), with $R = \sqrt{d_z} + x$, proves (106). □

F.1.2 Empirical risk minimization over independent pretraining episodes

For each independent training episode e, draw $\tau_e \sim \varpi_{t_0,T}$ and $\xi^{(e)}_{\mathrm{dsm}} \sim \mathcal{N}(0, I_{d_z})$, independently of the episode and of all other draws, and set

$$Z^{(e)}_{\tau_e} := \alpha_{\tau_e} Z^{(e)}_0 + \sigma_{\tau_e} \xi^{(e)}_{\mathrm{dsm}}, \qquad \xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}} := \mathrm{clip}_{R_{\mathrm{lab}}}(-\xi^{(e)}_{\mathrm{dsm}}).$$

The augmented training examples are independent copies of $(H_n, Z_\tau, \tau, \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}})$:

$$(H^{(e)}_n, Z^{(e)}_{\tau_e}, \tau_e, \xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}}) \overset{\text{i.i.d.}}{\sim} \mathsf{Law}(H_n, Z_\tau, \tau, \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}}), \qquad e = 1, \dots, N. \tag{107}$$

This augmentation transforms the episodic training sample $\mathcal{D}^{\mathrm{epi}}_N$ into

$$\mathcal{D}^{\mathrm{dsm}}_N := \big((H^{(e)}_n, Z^{(e)}_{\tau_e}, \tau_e, \xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}})\big)_{e=1}^N. \tag{108}$$

Here and below, we suppress the dependence on the fixed episode distribution and time window (t_0, T) in the notation for the augmented distribution, sample, and clipping-bias functional; we write the label radius explicitly when it is relevant. Let the score hypothesis class

$$\mathcal{S} \subseteq \{s : \mathbb{R}^{d_z} \times [t_0, T] \times \mathcal{H}_n \to \mathbb{R}^{d_z} : s \text{ is jointly Borel}\}$$

be deterministic and fixed before $\mathcal{D}^{\mathrm{dsm}}_N$ is observed. The empirical variance-balanced DSM risk is

$$\widehat{\mathcal{R}}^{(R_{\mathrm{lab}})}_{N,\mathrm{bDSM}}(s) := \frac{A_{t_0,T}}{N} \sum_{e=1}^N \|\widetilde{s}(Z^{(e)}_{\tau_e}, \tau_e, H^{(e)}_n) - \xi^{\mathrm{clip},(e)}_{R_{\mathrm{lab}}}\|_2^2. \tag{109}$$

For each fixed s, this empirical risk is unbiased for $\mathcal{R}^{(R_{\mathrm{lab}})}_{\mathrm{bDSM}}(s)$. A jointly measurable approximate ERM is a data-dependent $\widehat{s}_N \in \mathcal{S}$ satisfying

$$\widehat{\mathcal{R}}^{(R_{\mathrm{lab}})}_{N,\mathrm{bDSM}}(\widehat{s}_N) \le \inf_{s \in \mathcal{S}} \widehat{\mathcal{R}}^{(R_{\mathrm{lab}})}_{N,\mathrm{bDSM}}(s) + \varepsilon_{\mathrm{opt}}, \qquad \varepsilon_{\mathrm{opt}} \ge 0. \tag{110}$$

Here $\varepsilon_{\mathrm{opt}}$ is the optimization error in the scaled empirical objective above. Equivalently, the mean-squared-loss gap without the factor $A_{t_0,T}$ is at most $\varepsilon_{\mathrm{opt}}/A_{t_0,T}$. An exact ERM corresponds to $\varepsilon_{\mathrm{opt}} = 0$; the Transformer score theorem also allows a specified nonzero optimization gap. Joint measurability means that evaluation of $\widehat{s}_N$ is Borel jointly in the DSM training sample and (z, t, h). If training uses auxiliary randomization, impose the same requirement jointly in that randomization; because the proof event is uniform over $s \in \mathcal{S}$, the probability statement below may equivalently include the training randomization. Training evaluates the learned score $\widehat{s}_N$ through its scaled representation $\sigma_t \widehat{s}_N$. For this data-dependent score, $\mathcal{E}_{\mathrm{score}}(\widehat{s}_N)$ is evaluated on a fresh episode, conditionally on $\mathcal{D}^{\mathrm{dsm}}_N$.

F.2 Scaled-Score Covering Metric

Variance-balanced DSM evaluates the rescaled prediction $\widetilde{s} = \sigma_t s$, so it suffices to establish uniform convergence in a metric on these rescaled predictions. Recall the scaled class $\widetilde{\mathcal{S}}$ and the supremum metric on scaled scores $d_{\mathrm{bal},\infty}$ from (62). The two classes describe the same predictors: because $\sigma_t > 0$ on $[t_0, T]$, the map $s \mapsto \widetilde{s}$ is a bijection, with inverse $\widetilde{s} \mapsto \widetilde{s}/\sigma_t$. Since $\sigma_{t_0} \le \sigma_t \le 1$,

$$d_{\mathrm{bal},\infty}(s, s') \le \|s - s'\|_{\infty,2} \le \sigma_{t_0}^{-1} d_{\mathrm{bal},\infty}(s, s'), \qquad s, s' \in \mathcal{S},$$

and the best-in-class score error is unchanged by this representation:

$$\inf_{\widetilde{s} \in \widetilde{\mathcal{S}}} \mathcal{E}_{\mathrm{score}}(\widetilde{s}/\sigma_t) = \mathcal{E}_{\mathrm{app}}(\mathcal{S}).$$

Moreover, $s \mapsto \widetilde{s}$ is an isometry from $(\mathcal{S}, d_{\mathrm{bal},\infty})$ onto $(\widetilde{\mathcal{S}}, \|\cdot\|_{\infty,2})$. Hence, for every $\varepsilon \ge 0$,

$$N_{\mathrm{cov}}(\varepsilon, \mathcal{S}, d_{\mathrm{bal},\infty}) = N_{\mathrm{cov}}(\varepsilon, \widetilde{\mathcal{S}}, \|\cdot\|_{\infty,2}) \le N_{\mathrm{cov}}(\varepsilon, \mathcal{S}, \|\cdot\|_{\infty,2}).$$

The proof of Theorem F.2 uses a finite ε-cover in $d_{\mathrm{bal},\infty}$ only to establish uniform convergence. The ERM itself remains defined over the entire score class $\mathcal{S}$; the cover is used only in the proof and does not define a discretized hypothesis class or an implementation requirement. This supremum metric on scaled scores also avoids the factor $\sigma_{t_0}^{-1}$ that can arise when distances are instead measured on the unscaled score functions.

F.3 Finite-Episode Score-Learning Theorem

The next result converts approximate empirical minimization into a conditional score-error guarantee on a fresh episode. It applies to any deterministic score class whose scaled outputs are uniformly bounded and admit a finite cover in $d_{\mathrm{bal},\infty}$; the Transformer class is substituted later in Section G.8.

Theorem F.2 (Finite-episode score learning with variance-balanced DSM)**.** *Fix $N \geq 1$, $R_{\mathrm{lab}} > 0$, $0 \leq R_{\mathrm{pred}} < \infty$, and $0 \leq \varepsilon_{\mathrm{cov}} \leq R_{\mathrm{pred}} + R_{\mathrm{lab}}$. Let $\mathcal{D}_N^{\mathrm{dsm}}$ be the i.i.d. augmented training sample in (107) and (108), and let $\mathcal{S}$ be a nonempty deterministic class of jointly Borel scores satisfying*

$$\sup_{s\in\mathcal{S}} \sup_{z,t,h} \|\sigma_t s(z,t,h)\|_2 \leq R_{\mathrm{pred}}, \qquad N_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}}, \mathcal{S}, d_{\mathrm{bal},\infty}) < \infty. \tag{111}$$

Let $\widehat{s}_N \in \mathcal{S}$ be a jointly measurable approximate ERM satisfying (110) with deterministic optimization error $\varepsilon_{\mathrm{opt}}$. For every $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathcal{D}_N^{\mathrm{dsm}}$,

$$\begin{aligned}\mathcal{E}_{\mathrm{score}}(\widehat{s}_N) \lesssim\ & \mathcal{E}_{\mathrm{app}}(\mathcal{S}) + \varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab}}) + \varepsilon_{\mathrm{opt}} \\ & + A_{t_0,T}(R_{\mathrm{pred}} + R_{\mathrm{lab}})\varepsilon_{\mathrm{cov}} \\ & + A_{t_0,T}(R_{\mathrm{pred}} + R_{\mathrm{lab}})^2 \frac{\log\big(2N_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}}, \mathcal{S}, d_{\mathrm{bal},\infty})/\delta\big)}{N}.\end{aligned} \tag{112}$$

The probability event concerns the N independent augmented training episodes, while $\mathcal{E}_{\mathrm{score}}(\widehat{s}_N)$ evaluates the fitted score on a fresh episode conditional on that training sample. The inequality compares the fitted score with the best approximation in $\mathcal{S}$ and shows separately how the error depends on label clipping, empirical optimization, the finite-cover approximation used to extend the bound to the entire class, and estimation from finitely many episodes.

For the Transformer class, the measurable-selection argument in Proposition E.7 supplies a measurable exact ERM. The theorem itself permits any jointly measurable approximate ERM satisfying (110); this assumption concerns the estimator's measurability and does not provide an algorithm for finding it. This establishes the existence of a measurable minimizer, not an algorithm for finding it. The optimization gap in (110) accounts for approximate empirical minimization.

Proof of Theorem F.2. We first show that the excess loss has variance controlled by its expectation. This gives a Bernstein bound on a finite score cover, which we extend to the entire score class before applying the approximate-ERM inequality and converting the clipped regression error to score error.

The clipped regression function satisfies

$$\widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n) = \mathbb{E}[\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} \mid Z_\tau, \tau, H_n], \qquad \|\widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n)\|_2 \leq R_{\mathrm{lab}}$$

almost surely. For $s \in \mathcal{S}$, define its clipped excess loss by

$$\begin{aligned} L_s^{\mathrm{exc}} &:= \|\widetilde{s}(Z_\tau, \tau, H_n) - \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}}\|_2^2 - \|\widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n) - \xi^{\mathrm{clip}}_{R_{\mathrm{lab}}}\|_2^2, \\ \mathbb{E}[L_s^{\mathrm{exc}}] &= \mathbb{E}\|\widetilde{s}(Z_\tau, \tau, H_n) - \widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n)\|_2^2. \end{aligned}$$

The second identity is the regression Pythagorean identity, and all expectations below are over one episode augmented for variance-balanced DSM. For brevity, put

$$R_{\mathrm{loss}} := R_{\mathrm{pred}} + R_{\mathrm{lab}}, \qquad \Gamma_{\mathrm{cov}}(\delta) := \log\big(2N_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}}, \mathcal{S}, d_{\mathrm{bal},\infty})/\delta\big).$$

The difference of squared losses factors as

$$L_s^{\mathrm{exc}} = \left\langle \widetilde{s}(Z_\tau, \tau, H_n) - \widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n), \widetilde{s}(Z_\tau, \tau, H_n) + \widetilde{s}^{\mathrm{lab}}_{R_{\mathrm{lab}}}(Z_\tau, \tau, H_n) - 2\xi^{\mathrm{clip}}_{R_{\mathrm{lab}}} \right\rangle.$$

The first factor is the prediction error relative to the clipped regression function; the bounds on the scaled scores and clipped labels control the second factor. Together with the regression identity above, this gives

$$|L_s^{\mathrm{exc}}| \leq 3R_{\mathrm{loss}}^2, \qquad \mathbb{E}[(L_s^{\mathrm{exc}})^2] \leq 9R_{\mathrm{loss}}^2 \mathbb{E}[L_s^{\mathrm{exc}}].$$

In particular, $|L_s^{\mathrm{exc}} - \mathbb{E}[L_s^{\mathrm{exc}}]| \leq 6R_{\mathrm{loss}}^2$ and $\mathrm{Var}(L_s^{\mathrm{exc}}) \leq 9R_{\mathrm{loss}}^2 \mathbb{E}[L_s^{\mathrm{exc}}]$. Thus the excess losses obey the variance–expectation condition underlying standard fast-rate regression bounds (Bartlett

et al., 2005; Koltchinskii, 2011). Here a finite-cover argument suffices; we do not invoke a theorem based on local Rademacher complexity for the nonconvex Transformer class. Under the convention in Section E that cover centers belong to the score class, let $\mathcal{S}_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}}) \subseteq \mathcal{S}$ be a minimum-cardinality $\varepsilon_{\mathrm{cov}}$-net in $d_{\mathrm{bal},\infty}$. This net is used only in the proof; empirical minimization remains over the entire score class $\mathcal{S}$. Let $\widehat{\mathbb{E}}_N$ denote the empirical average over the N episodes augmented for variance-balanced DSM. Bernstein's inequality (Boucheron et al., 2013) and a union bound imply that, with probability at least $1-\delta$, simultaneously for every $u \in \mathcal{S}_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}})$,

$$|\widehat{\mathbb{E}}_N[L_u^{\mathrm{exc}}] - \mathbb{E}[L_u^{\mathrm{exc}}]| \le \sqrt{\frac{18R_{\mathrm{loss}}^2 \mathbb{E}[L_u^{\mathrm{exc}}]\Gamma_{\mathrm{cov}}(\delta)}{N}} + \frac{4R_{\mathrm{loss}}^2 \Gamma_{\mathrm{cov}}(\delta)}{N}. \tag{113}$$

For every $s \in \mathcal{S}$, choose $s^\sharp \in \mathcal{S}_{\mathrm{cov}}(\varepsilon_{\mathrm{cov}})$ satisfying

$$\|\sigma_t s - \sigma_t s^\sharp\|_{\infty,2} \le \varepsilon_{\mathrm{cov}}.$$

The difference of squares and the bounds on the scaled score maps and clipped labels give

$$|L_s^{\mathrm{exc}} - L_{s^\sharp}^{\mathrm{exc}}| \le 2R_{\mathrm{loss}}\varepsilon_{\mathrm{cov}}, \qquad \mathbb{E}[L_{s^\sharp}^{\mathrm{exc}}] \le 2\mathbb{E}[L_s^{\mathrm{exc}}] + 2\varepsilon_{\mathrm{cov}}^2. \tag{114}$$

The first inequality controls both its empirical and population averages. Combining (113) and (114), using $\varepsilon_{\mathrm{cov}} \le R_{\mathrm{loss}}$, and applying $ab \le (a^2+b^2)/2$ to the resulting cross term yields, on the same event,

$$|\widehat{\mathbb{E}}_N[L_s^{\mathrm{exc}}] - \mathbb{E}[L_s^{\mathrm{exc}}]| \lesssim \sqrt{\frac{R_{\mathrm{loss}}^2 \mathbb{E}[L_s^{\mathrm{exc}}]\Gamma_{\mathrm{cov}}(\delta)}{N}} + \frac{R_{\mathrm{loss}}^2 \Gamma_{\mathrm{cov}}(\delta)}{N} + R_{\mathrm{loss}}\varepsilon_{\mathrm{cov}}, \qquad s \in \mathcal{S}. \tag{115}$$

For any deterministic $s \in \mathcal{S}$, subtract the empirical loss of the clipped regression function from (110). Apply (115) to both $\widehat{s}_N$ and s on its simultaneous event. Young's inequality absorbs the square-root term involving $\mathbb{E}[L_{\widehat{s}_N}^{\mathrm{exc}}]$ into the left side and bounds the comparator's square-root term by its risk plus the complexity term. Multiplying by $A_{t_0,T}$ yields

$$A_{t_0,T}\mathbb{E}[L_{\widehat{s}_N}^{\mathrm{exc}}] \lesssim A_{t_0,T}\mathbb{E}[L_s^{\mathrm{exc}}] + A_{t_0,T}R_{\mathrm{loss}}^2 \frac{\Gamma_{\mathrm{cov}}(\delta)}{N} + A_{t_0,T}R_{\mathrm{loss}}\varepsilon_{\mathrm{cov}} + \varepsilon_{\mathrm{opt}}.$$

For every $s \in \mathcal{S}$, the squared triangle inequality and the inverse-variance time distribution give

$$A_{t_0,T}\mathbb{E}[L_s^{\mathrm{exc}}] \le 2\mathcal{E}_{\mathrm{score}}(s) + 2\varepsilon_{\mathrm{lab,clip}}^2(R_{\mathrm{lab}}),$$

while (104) gives

$$\mathcal{E}_{\mathrm{score}}(\widehat{s}_N) \le 2A_{t_0,T}\mathbb{E}[L_{\widehat{s}_N}^{\mathrm{exc}}] + 2\varepsilon_{\mathrm{lab,clip}}^2(R_{\mathrm{lab}}).$$

By the convention of Section B.3, for each realized training sample this population error is exactly the conditional error on a fresh episode. Combining the preceding bounds and taking the infimum over $s \in \mathcal{S}$ proves (112) with a universal implicit constant. Every concentration variable above is one complete episode augmented for variance-balanced DSM; auxiliary perturbations of the same episode do not increase the independent sample size beyond N. □

F.4 Error Decomposition and Tuning Guidance

The first three terms in (112) measure model approximation, label clipping, and optimization. The final two arise from the uniform-convergence argument: one accounts for extending the bound from a finite cover to the entire score class, and the other is the covering-number generalization term. Only these final two terms depend on the cover radius, and the generalization term also depends logarithmically on the cover cardinality.

For a desired clipping-bias tolerance $\varepsilon_{\mathrm{lab,tar}}^2 > 0$, the explicit choice

$$R_{\mathrm{lab}} = \sqrt{d_z} + \left[2\max\left\{0, \log\frac{2A_{t_0,T}}{\varepsilon_{\mathrm{lab,tar}}^2}\right\}\right]^{1/2}$$

gives $\varepsilon_{\mathrm{lab,clip}}^2(R_{\mathrm{lab}}) \le \varepsilon_{\mathrm{lab,tar}}^2$. Increasing R_{lab} reduces clipping bias but increases the bound $R_{\mathrm{pred}} + R_{\mathrm{lab}}$ on the difference between the prediction and label, and hence the bound on the

squared loss. All tuning parameters may depend deterministically on (n, N, t_0, T) and on fixed class parameters, but not on the observed training sample.

A convenient choice of the cover radius $\varepsilon_{\mathrm{cov}}$ as a function of the episode count N is

$$\varepsilon_{\mathrm{cov}} = \frac{\min\{1, R_{\mathrm{pred}} + R_{\mathrm{lab}}\}}{(N+1)^2}. \tag{116}$$

Whenever the corresponding cover is finite, substituting (116) into (112) gives an explicit score-error bound whose square root enters the reverse-SDE stability analysis linearly. With this choice, the ε-net approximation term is lower order than the N^{-1} generalization term in (112). Later, $\varepsilon_{\mathrm{cov},n}$ denotes a different radius for the Transformer class, chosen as a function of the prompt length before the N training episodes are observed.

G Reverse Diffusion Sampling and Posterior-Predictive Generation

This section converts integrated score error into a posterior-predictive generation guarantee. We first define the randomized inverse-variance time grid and establish the exact reverse-process and endpoint estimates. We then combine a finite-energy KL comparison with exact-score dynamics to obtain a positive-time discretization bound. The final four subsections choose the accuracy parameters and tail radii, combine approximation, learning, and sampling, and record the scalar sample-only specialization. The finite-step theorem applies, conditionally on training, to any finite-valued jointly Borel score with finite integrated score error; its output law averages over the independent grid shift and Gaussian sampling noise. Throughout, $M \in \mathbb{N}$ is the number, or an almost-sure upper bound on the number, of reverse updates.

A small time-integrated error does not ensure accuracy at predetermined evaluation times for a merely measurable score. The randomized grid in Section 2.4 addresses this mismatch: averaging the variance-weighted sum of score errors over its shift recovers the training integral. This makes the score-learning guarantee applicable to the sampler without an additional time-regularity assumption.

Corollary G.6 bounds predictive error by initialization, the square root of integrated score error, early stopping, and discretization. The bound concerns the sampler's output distribution averaged over its independent grid shift and uses at most M score evaluations almost surely under (136).

G.1 Randomized Inverse-Variance Time Grid

The randomly shifted inverse-variance grid is chosen to match the error controlled by score learning. The discrete score-error contribution is bounded by a random quadrature whose expectation equals the integrated squared score error. Thus, this random upper bound is unbiased for the integrated criterion, although the discrete contribution itself need not be. No Lipschitz bound on the learned score is needed.

Let $T > t_0 > 0$. For $t \in [t_0, T]$, write

$$\mathsf{A}_{t_0}(t) := A_{t_0,t} = \int_{t_0}^{t} \frac{\mathrm{d}u}{\sigma_u^2}.$$

The quantity $A_{t_0,T}$ is the length of the sampling interval in inverse-variance time.

Fix a grid spacing

$$0 < h_{\mathrm{grid}} \le \min\{1/2, A_{t_0,T}, T - t_0\}.$$

Independently of all other randomness, draw the grid shift $U_{\mathrm{grid}} \sim \mathrm{Unif}(0, h_{\mathrm{grid}})$ and put

$$J_{\mathrm{grid}} := \max\{j \in \mathbb{N}_0 : U_{\mathrm{grid}} + j h_{\mathrm{grid}} \le A_{t_0,T}\}, \qquad v_{-1} := t_0,$$
$$v_j := \mathsf{A}_{t_0}^{-1}(U_{\mathrm{grid}} + j h_{\mathrm{grid}}), \qquad \widetilde{T}_{\mathrm{grid}} := v_{J_{\mathrm{grid}}}.$$

Starting from $\widehat{Z}^{\mathrm{grid}}_{J_{\mathrm{grid}}+1} \sim \mathcal{N}(0, I_{d_z})$, for $j = J_{\mathrm{grid}}, J_{\mathrm{grid}} - 1, \dots, 0$ use the exponential–Euler update

$$\widehat{Z}^{\mathrm{grid}}_j = e^{\Delta_j} \widehat{Z}^{\mathrm{grid}}_{j+1} + 2(e^{\Delta_j} - 1)\widehat{s}(\widehat{Z}^{\mathrm{grid}}_{j+1}, v_j, h) + \sqrt{e^{2\Delta_j} - 1}\, \xi^{\mathrm{grid}}_j, \qquad \Delta_j := v_j - v_{j-1}. \tag{117}$$

Here $\widehat{Z}^{\mathrm{grid}}_{j+1}$ is the state at v_j before update j, and $\widehat{Z}^{\mathrm{grid}}_j$ is the state at v_{j-1} afterward. Thus, the initial index $J_{\mathrm{grid}} + 1$ does not denote an additional time node. The vectors ξ^{grid}_j are independent and standard Gaussian. Let $\widehat{Q}^{\mathrm{grid},h_{\mathrm{grid}}}_{t_0 \mid h}$ denote the distribution of $\widehat{Z}^{\mathrm{grid}}_0$ after averaging over U_{grid} and the Gaussian sampler noise. Joint Borel measurability of the learned score makes $h \mapsto \widehat{Q}^{\mathrm{grid},h_{\mathrm{grid}}}_{t_0 \mid h}$ a Borel probability kernel. Define the corresponding data-space output kernel by

$$\widehat{P}^{\mathrm{grid},h_{\mathrm{grid}}}_{0 \mid h} := (\mathrm{Proj}_{\mathcal{Z}})_{\#} \widehat{Q}^{\mathrm{grid},h_{\mathrm{grid}}}_{t_0 \mid h}. \tag{118}$$

The guarantees below concern this kernel, which averages over the random grid shift, and do not assert the same bound for every fixed realization of U_{grid}.

G.2 Exact Reverse Process on Positive Times

Recall the conditional OU marginals and score from Section C.1. Let $r \in [0, T)$ be increasing reverse time, so that the corresponding forward time is $t = T - r$. For fixed h, let $Z_r^{\leftarrow} := Z_{T-r}$ and equip the reversed path with its natural reversed filtration. Smooth positivity and (29), applied to the forward segment $[\varepsilon, T]$, verify the local regularity, linear-growth, and finite-energy hypotheses of the diffusion time-reversal theorem. The standard diffusion time-reversal theorem (Haussmann & Pardoux, 1986) implies that, for every $\varepsilon \in (0, T)$, its restriction to $0 \le r \le T - \varepsilon$ is a weak solution of

$$\mathrm{d}Z_r^{\leftarrow} = \{Z_r^{\leftarrow} + 2s^\star(Z_r^{\leftarrow}, T - r \,|\, h)\}\,\mathrm{d}r + \sqrt{2}\,\mathrm{d}W_r^{\leftarrow}, \qquad Z_0^{\leftarrow} \sim P_{T\,|\,h}, \tag{119}$$

and its marginal at every $0 \le r < T$ is $P_{T-r\,|\,h}$. Its marginals therefore converge weakly to the posterior-predictive distribution $P_{0\,|\,h}$ as $r \uparrow T$. No regularity of the original output distribution $P_{0\,|\,h}$ is needed because the SDE representation is asserted only on intervals bounded away from forward time zero.

For each problem instance, fix a deterministic $0 < t_0 < T$; later bounds allow (t_0, T) to depend on n. Score learning is restricted to forward diffusion times $t \in [t_0, T]$, and the approximate reverse sampler is run only through reverse time $T - t_0$. Thus t_0 is a cutoff in diffusion time, not an optimization stopping rule. This positive cutoff avoids potentially nonuniform score behavior as $t \downarrow 0$ and corresponds numerically to a minimum positive noise level.

The exact unprojected reverse endpoint has distribution $P_{t_0\,|\,h}$ rather than $P_{0\,|\,h}$; after Euclidean projection onto $\mathcal{Z}$, the exact stopped distribution is $(\mathrm{Proj}_{\mathcal{Z}})_{\#}P_{t_0\,|\,h}$.

G.3 Endpoint Errors: Gaussian Initialization and Early Stopping

The reverse sampler incurs approximation errors at both endpoints of the diffusion interval. At terminal time T, it starts from a standard Gaussian approximation to the noised posterior-predictive distribution, whereas at time t_0 it stops at positive diffusion time. We treat these errors separately because they depend on different assumptions and tuning parameters.

G.3.1 Terminal Gaussian initialization error

At terminal time T, the learned reverse process is initialized from the standard Gaussian distribution $\mathcal{N}(0, I_{d_z})$ instead of the exact noised posterior-predictive distribution $P_{T\,|\,H_n}$. Define this initialization error by

$$\varepsilon_{\mathrm{init}}(T) := \mathbb{E}\left[\mathsf{TV}\left(P_{T\,|\,H_n}, \mathcal{N}(0, I_{d_z})\right)\right]. \tag{120}$$

By the output-space convention in Section B.1, $P_{0\,|\,H_n}$ is supported on $[0,1]^{d_z}$. The resulting explicit initialization bound is

$$\varepsilon_{\mathrm{init}}(T) \le \frac{1}{2}\sqrt{-d_z \log(1 - e^{-2T})}, \qquad T > 0. \tag{121}$$

For the chosen horizons $T \ge 1$, the same calculation gives $\varepsilon_{\mathrm{init}}(T) \le \sqrt{d_z/2}\, e^{-T}$. Conditionally on $Z_0 = z$, the forward distribution at time T is

$$\mathcal{N}\{e^{-T} z, (1 - e^{-2T}) I_{d_z}\}.$$

Convexity of KL, the Gaussian KL formula, and $\|z\|_2^2 \le d_z$ give

$$\mathrm{KL}(P_{T\,|\,H_n} \| \mathcal{N}(0, I_{d_z})) \le -\frac{d_z}{2}\log(1 - e^{-2T}).$$

Pinsker's inequality and Jensen's inequality yield (121). For $T \ge 1$, the elementary inequality $-\log(1 - e^{-2T}) \le 2e^{-2T}$ gives $\varepsilon_{\mathrm{init}}(T) \le \sqrt{d_z/2}\, e^{-T}$.

G.3.2 Early-stopping error

Recall that $\beta' = \beta \wedge 1$. Reverse sampling stops at $t_0 > 0$, so projecting the positive-time distribution back to the data cube introduces an early-stopping error. The smoothness exponent is capped at one because even an infinitely smooth density generally moves mass of order $\sqrt{t}$ across the cube boundary before projection.

Lemma G.1 (Hölder early-stopping error bound on the cube). *Let $P(\mathrm{d}z) = p(z)\,\mathrm{d}z$ on $\mathcal{Z} = [0,1]^{d_z}$ with $p \in \mathcal{F}_\beta^{d_z}(L_\beta)$, and let P_t be its OU evolution in (26). There is $C_{\mathrm{stop}} = C_{\mathrm{stop}}(\beta, d_z, L_\beta) < \infty$ such that, for every $0 < t \le 1$,*

$$\mathsf{TV}\left((\mathrm{Proj}_{\mathcal{Z}})_{\#}P_t, P\right) \le \mathsf{TV}(P_t, P) \le C_{\mathrm{stop}} t^{\beta'/2}.$$

No lower density bound is required.

Proof of Lemma G.1. Write $d = d_z$, and let $\bar{p}$ be the zero extension of p from $\mathcal{Z} = [0,1]^d$ to $\mathbb{R}^d$. First note the global translation modulus

$$\|\bar{p}(\cdot - y) - \bar{p}\|_{L^1(\mathbb{R}^d)} \le H_{\mathrm{tr}} \|y\|_2^{\beta'}, \qquad y \in \mathbb{R}^d, \tag{122}$$

for a constant $H_{\mathrm{tr}} = H_{\mathrm{tr}}(d, \beta, L_\beta)$. Indeed, on $\mathcal{Z} \cap (\mathcal{Z} + y)$ the Hölder condition gives a bound of order $\|y\|_2^{\beta'}$ (the bounded gradient gives power one when $\beta > 1$), while the two boundary layers have Lebesgue measure at most $C_d \min\{1, \|y\|_2\}$. Since $\|p\|_\infty \le L_\beta$, their contribution has the same power for $\|y\|_2 \le 1$; for larger translations the trivial L^1 bound by two proves (122) after increasing H_{tr}.

For $0 < \alpha \le 1$, set $D_\alpha \bar{p}(x) := \alpha^{-d} \bar{p}(x/\alpha)$. Uniformly for $e^{-1} \le \alpha \le 1$,

$$\|D_\alpha \bar{p} - \bar{p}\|_1 \le H_{\mathrm{dil}} (1 - \alpha)^{\beta'}, \tag{123}$$

for a constant $H_{\mathrm{dil}} = H_{\mathrm{dil}}(d, \beta, L_\beta)$. To see this, split the integral over $\alpha\mathcal{Z}$ and $\mathcal{Z} \setminus \alpha\mathcal{Z}$. The normalization difference and the boundary layer are $\mathcal{O}(1 - \alpha)$. On $\alpha\mathcal{Z}$ compare $p(x/\alpha)$ with $p(x)$; their arguments are at distance at most $\sqrt{d}(\alpha^{-1} - 1)$, so Hölder continuity (or the bounded gradient) gives $\mathcal{O}((1 - \alpha)^{\beta'})$. This proves (123). A change of variables in (122) also gives

$$\|D_\alpha \bar{p}(\cdot - y) - D_\alpha \bar{p}\|_1 \le H_{\mathrm{tr}} \alpha^{-\beta'} \|y\|_2^{\beta'}. \tag{124}$$

Put $\alpha = \alpha_t = e^{-t}$ and $\sigma = \sigma_t$. Let $\varphi_\sigma^{(d)}$ be the centered d-variate Gaussian density with covariance $\sigma^2 I_d$. The OU density is $p_t = \varphi_\sigma^{(d)} * D_\alpha \bar{p}$. Minkowski's inequality, (123) and (124), and $\xi \sim \mathcal{N}(0, I_d)$ yield

$$\|p_t - \bar{p}\|_1 \le H_{\mathrm{tr}} \alpha^{-\beta'} \sigma^{\beta'} \mathbb{E}[\|\xi\|_2^{\beta'}] + H_{\mathrm{dil}} (1 - \alpha)^{\beta'}. \tag{125}$$

For $0 < t \le 1$, $\alpha^{-\beta'} \le e$, $\sigma \le \sqrt{2t}$, $1 - \alpha \le t$, and $\mathbb{E}[\|\xi\|_2^{\beta'}] \le d^{\beta'/2}$. Dividing (125) by two gives $\mathsf{TV}(P_t, P) \le C_{\mathrm{stop}} t^{\beta'/2}$. Finally, projection is a Markov kernel and fixes every point of $\mathcal{Z}$, so

$$\mathsf{TV}((\mathrm{Proj}_{\mathcal{Z}})_{\#}P_t, P) = \mathsf{TV}((\mathrm{Proj}_{\mathcal{Z}})_{\#}P_t, (\mathrm{Proj}_{\mathcal{Z}})_{\#}P) \le \mathsf{TV}(P_t, P).$$

□

Corollary G.2 (Conditional early-stopping error bound). *Fix $\beta > 0$ and $L_\beta \ge 1$, and consider either the sample-only or the input–output prompt model in (19) and (20) with any Borel task prior Π. Suppose the task laws admit jointly Borel densities $p_\theta(z)$ in the sample-only model or $f_\theta(x, z)$ in the input–output model, with $p_\Theta \in \mathcal{F}_\beta^{d_z}(L_\beta)$ or $f_\Theta \in \mathcal{F}_\beta^{d_x + d_z}(L_\beta)$, respectively, Π-almost surely. Writing $\beta' = \beta \wedge 1$, there is a constant C_{stop} depending only on (β, L_β) and the dimensions such that*

$$\mathbb{E}\left[\mathsf{TV}\left((\mathrm{Proj}_{\mathcal{Z}})_{\#}P_{t \mid H_n}, P_{0 \mid H_n}\right)\right] \le C_{\mathrm{stop}} t^{\beta'/2}, \qquad 0 < t \le 1. \tag{126}$$

For any standard Borel space $\mathcal{U}$ and Borel map $\phi_{\mathrm{pr}} : \mathcal{H}_n \to \mathcal{U}$, set $U = \phi_{\mathrm{pr}}(H_n)$. Write $P_{0|U} := \mathsf{Law}(Z_0 \mid U)$ and $P_{t|U} := \mathsf{Law}(Z_t \mid U)$. The same bound holds with H_n replaced by U. In the input–output model, the expectation is over the query drawn from its input marginal; no lower bound on its marginal density is assumed.

Proof of Corollary G.2. In the sample-only task family, for almost every prompt h, the density of $P_{0 \mid h}$ is the posterior mixture of the task densities. The Hölder ball is convex, so Lemma G.1 applies with the same constant and averaging proves (126).

For the input–output task family, condition first on the demonstrations $C_n = c$. Let

$$f_c(x,z) := \int f_\theta(x,z)\Pi(\mathrm{d}\theta \,|\, C_n = c), \qquad \lambda_c(x) := \int_{\mathcal{Z}} f_c(x,z)\,\mathrm{d}z.$$

The query has density λ_c, and where $\lambda_c(x) > 0$ the predictive output density is $f_c(x,\cdot)/\lambda_c(x)$. Let T_t^z denote unprojected OU evolution in the z coordinate and extend $f_c(x,\cdot)$ by zero outside the cube. Projection contraction, linearity, and cancellation of the query marginal give

$$\mathbb{E}\big[\mathsf{TV}((\mathrm{Proj}_{\mathcal{Z}})_\# P_{t\,|\,H_n}, P_{0\,|\,H_n}) \,|\, C_n = c\big] \le \frac{1}{2}\int_{[0,1]^{d_x}} \int_{\mathbb{R}^{d_z}} |\mathsf{T}_t^z f_c(x,z) - f_c(x,z)|\,\mathrm{d}z\,\mathrm{d}x.$$

The proof of Lemma G.1 applies verbatim to each nonnegative subdensity $f_c(x,\cdot)$: its mass and its z-Hölder norm are bounded uniformly by a constant depending only on L_β. Integrating over the unit-volume input cube proves the asserted bound without dividing by a lower bound for λ_c.

For $U = \phi_{\mathrm{pr}}(H_n)$, both conditional laws are mixtures of their counterparts given H_n. Convexity of TV and the tower property give

$$\mathbb{E}\mathsf{TV}((\mathrm{Proj}_{\mathcal{Z}})_\# P_{t|U}, P_{0|U}) \le \mathbb{E}\mathsf{TV}((\mathrm{Proj}_{\mathcal{Z}})_\# P_{t|H_n}, P_{0|H_n}),$$

proving the claim with conditioning on U. □

For a target early-stopping error $\varepsilon_{\mathrm{stop}} \in (0,1]$, the upper bound permits $t_0 \asymp \varepsilon_{\mathrm{stop}}^{2/\beta'}$.

For TV accuracy ε, (139) chooses $t_0 \asymp \varepsilon^{2/(\beta\wedge 1)}$ and fixes the constants in the early-stopping bound.

G.4 Finite-Energy KL Comparison for Reverse Diffusions

The following lemma states the stopped finite-energy form of Girsanov's theorem used below; see Liptser & Shiryaev (2001) for the standard localization approach. We include the proof because no global Novikov condition is assumed.

Lemma G.3 (Finite-energy KL bound for diffusions with matched noise)**.** *Fix $d \in \mathbb{N}$, a finite path horizon $T_{\mathrm{path}} > 0$, and an initial distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$. Let Q be the law on $C([0,T_{\mathrm{path}}];\mathbb{R}^d)$ of*

$$\mathrm{d}Y_r = b(r,Y_r)\,\mathrm{d}r + \sqrt{2}\,\mathrm{d}W_r, \qquad Y_0 \sim \mu,$$

where $b : [0,T_{\mathrm{path}}]\times\mathbb{R}^d \to \mathbb{R}^d$ is Borel, globally Lipschitz in its state variable uniformly in r, and has linear growth. Let $b^\star : [0,T_{\mathrm{path}}]\times\mathbb{R}^d \to \mathbb{R}^d$ be Borel, and let $\mathsf{P}^\star$ be the law of a weak solution with the same initial distribution and drift $b^\star$. If

$$\mathbb{E}_{\mathsf{P}^\star}\int_0^{T_{\mathrm{path}}} \|b^\star(r,Y_r) - b(r,Y_r)\|_2^2\,\mathrm{d}r < \infty,$$

then

$$\mathrm{KL}(\mathsf{P}^\star\|\mathsf{Q}) \le \frac{1}{4}\mathbb{E}_{\mathsf{P}^\star}\int_0^{T_{\mathrm{path}}} \|b^\star(r,Y_r) - b(r,Y_r)\|_2^2\,\mathrm{d}r. \tag{127}$$

The same conclusion holds when the reference process is defined by a recursively frozen nonanticipative drift. Specifically, choose a deterministic partition $0 = r_0 < \cdots < r_K = T_{\mathrm{path}}$ and finite-valued Borel maps $a_k : \mathbb{R}^d \to \mathbb{R}^d$, $0 \le k < K$, and set

$$b_{\mathrm{fr}}(r,\omega) := \omega_r + 2a_k(\omega_{r_k}), \qquad r \in [r_k, r_{k+1}),$$

with the last interval closed at T_{path}. In this alternative case, let Q denote the law of the unique interval-by-interval solution of $\mathrm{d}Y_r = b_{\mathrm{fr}}(r,Y_{[0,r]})\,\mathrm{d}r + \sqrt{2}\,\mathrm{d}W_r$ with $Y_0 \sim \mu$. In this case, replace $b(r,Y_r)$ in the energy condition and in (127) by $b_{\mathrm{fr}}(r,Y_{[0,r]})$. The construction includes $a_k(x) = \widehat{s}(x,\vartheta_k,h)$ for fixed score-evaluation times ϑ_k whenever $\widehat{s}$ is Borel and finite-valued. The reference process and every post-stopping continuation are then constructed explicitly, interval by interval, so no state regularity of the maps a_k is required.

Proof. Put $\Delta b_r = b^\star(r, Y_r) - b(r, Y_r)$ in the Markov case and $\Delta b_r = b^\star(r, Y_r) - b_{\mathrm{fr}}(r, Y_{[0,r]})$ in the recursively frozen case, and define

$$\upsilon_m := \inf\left\{ r : \int_0^r \|\Delta b_u\|_2^2 \,\mathrm{d}u \geq m \right\} \wedge T_{\mathrm{path}}.$$

On the filtered weak-solution space carrying $\mathsf{P}^\star$, the process Δb is progressively measurable, so υ_m is a stopping time. On a realization of the weak solution under $\mathsf{P}^\star$, retain its path through υ_m and, after υ_m, continue with the reference drift and the same Brownian increments. Global Lipschitzness of b makes this continuation well defined in the Markov case; the recursively frozen case is constructed successively on the remaining grid intervals. Indeed, conditional on the left endpoint of an interval $[r_k, r_{k+1}]$, its frozen map value is finite and measurable, and the resulting affine SDE has a unique explicit solution. Induction over the deterministic partition gives weak uniqueness for the recursively frozen reference process and for every post-stopping continuation. Denote its law by P_m. The coupling gives

$$\mathsf{TV}(\mathsf{P}_m, \mathsf{P}^\star) \leq \mathsf{P}^\star(\upsilon_m < T_{\mathrm{path}}) \leq \frac{1}{m}\mathbb{E}_{\mathsf{P}^\star} \int_0^{T_{\mathrm{path}}} \|\Delta b_r\|_2^2 \,\mathrm{d}r \longrightarrow 0.$$

Relative to Q, the drift change under P_m is $\mathbb{1}\{r \leq \upsilon_m\}\Delta b_r$. Its accumulated squared norm is at most m. Hence Novikov's criterion holds because

$$\mathbb{E}_{\mathsf{P}_m} \exp\left\{ \frac{1}{4} \int_0^{\upsilon_m} \|\Delta b_r\|_2^2 \,\mathrm{d}r \right\} \leq e^{m/4} < \infty.$$

If $W^{(m)}$ denotes the Brownian motion driving the process obtained by joining the two segments, define the stochastic exponential

$$D_u^{(m)} := \exp\left\{ -\frac{1}{\sqrt{2}} \int_0^{u\wedge\upsilon_m} \Delta b_r^\top \,\mathrm{d}W_r^{(m)} - \frac{1}{4}\int_0^{u\wedge\upsilon_m} \|\Delta b_r\|_2^2 \,\mathrm{d}r \right\}, \qquad 0 \leq u \leq T_{\mathrm{path}}.$$

This process is therefore a true, uniformly integrable martingale of mean one. Define Q_m by $\mathrm{d}\mathsf{Q}_m/\mathrm{d}\mathsf{P}_m = D^{(m)}_{T_{\mathrm{path}}}$. Girsanov's theorem then identifies its drift. Under Q_m the coordinate process has the reference drift b in the Markov case and b_{fr} in the recursively frozen case. Weak uniqueness for the globally Lipschitz reference equation identifies $\mathsf{Q}_m = \mathsf{Q}$ in the Markov case; the interval-by-interval construction gives the same identification in the recursively frozen case. Taking the P_m-expectation of $\log(\mathrm{d}\mathsf{P}_m/\mathrm{d}\mathsf{Q})$ eliminates the stochastic integral and gives

$$\mathrm{KL}(\mathsf{P}_m \| \mathsf{Q}) = \frac{1}{4}\mathbb{E}_{\mathsf{P}_m} \int_0^{\upsilon_m} \|\Delta b_m\|_2^2 \,\mathrm{d}r = \frac{1}{4}\mathbb{E}_{\mathsf{P}^\star} \int_0^{\upsilon_m} \|\Delta b_r\|_2^2 \,\mathrm{d}r.$$

Lower semicontinuity of relative entropy under $\mathsf{P}_m \to \mathsf{P}^\star$ in total variation, followed by monotone convergence, proves (127). □

G.5 Positive-Time Score Dynamics and Discretization

The next lemma controls the change of the exact score within a time-grid cell. Together with the KL bound for diffusions with finite drift energy, it gives the positive-time sampling bound below.

Lemma G.4 (Exact reverse-score dynamics on a positive-time interval)**.** *Fix a prompt h for which $P_{0\,|\,h}$ is supported on $[0,1]^{d_z}$, and let $p_t(\cdot\,|\,h)$ and $s_t^\star(\cdot\,|\,h) = \nabla \log p_t(\cdot\,|\,h)$ be its OU marginal density and score. For $0 < a < b < \infty$, let $(Y_r)_{0\leq r\leq b-a}$ be the exact reverse process*

$$\mathrm{d}Y_r = \{Y_r + 2s^\star_{b-r}(Y_r\,|\,h)\}\,\mathrm{d}r + \sqrt{2}\,\mathrm{d}W_r, \qquad Y_0 \sim P_{b\,|\,h}.$$

Then $Y_r \sim P_{b-r\,|\,h}$, and the forward representation $Z_t = \alpha_t Z_0 + \sigma_t \xi$ gives

$$s_t^\star(Z_t\,|\,h) = -\sigma_t^{-1}\mathbb{E}[\xi\,|\,Z_t, h], \qquad \nabla s_t^\star(Z_t\,|\,h) = \sigma_t^{-2}\{\mathrm{Cov}(\xi\,|\,Z_t, h) - I_{d_z}\}. \tag{128}$$

Consequently,

$$\mathbb{E}\|s_t^\star(Z_t\,|\,h)\|_2^2 \leq \frac{d_z}{\sigma_t^2},$$
$$\mathbb{E}\|\nabla s_t^\star(Z_t\,|\,h)\|_{\mathrm{F}}^2 \leq \frac{d_z(d_z+3)}{\sigma_t^4}. \tag{129}$$

Moreover, $S_r := s^\star_{b-r}(Y_r \mid h)$ *is a continuous semimartingale satisfying*

$$\mathrm{d}S_r = -S_r\,\mathrm{d}r + \sqrt{2}\,\nabla s^\star_{b-r}(Y_r \mid h)\,\mathrm{d}W_r. \tag{130}$$

All identities hold on the full interval $0 \le r \le b-a$ *in the usual localized Itô sense.*

Proof. At every positive time, Gaussian convolution makes $p_t(\cdot \mid h)$ strictly positive and smooth. Gaussian differentiation gives (128). Conditional Jensen proves the score bound. For $\Sigma_{t,h} := \mathrm{Cov}(\xi \mid Z_t, h)$, one has $\mathrm{Tr}(\Sigma_{t,h}^2) \le \{\mathrm{Tr}(\Sigma_{t,h})\}^2$ and

$$\mathbb{E}\{\mathrm{Tr}(\Sigma_{t,h})\}^2 \le \mathbb{E}\|\xi\|_2^4 = d_z(d_z+2).$$

Since $\|\Sigma_{t,h} - I_{d_z}\|_{\mathrm{F}}^2 \le \{\mathrm{Tr}(\Sigma_{t,h})\}^2 + d_z$, this proves the Hessian bound. The positive-time diffusion time-reversal identity gives $Y_r \sim P_{b-r\,|\,h}$. Stop the process when it leaves a compact state set and apply Itô's formula to $(r,y) \mapsto s^\star_{b-r}(y \mid h)$. Differentiating the OU Fokker–Planck equation gives

$$\partial_t s^\star_t = s^\star_t + (\nabla s^\star_t)y + \nabla\,\mathrm{div}\,s^\star_t + 2(\nabla s^\star_t)s^\star_t.$$

Substitution of the exact reverse drift cancels all finite-variation terms except $-S_r\,\mathrm{d}r$, proving (130) up to the stopping time. The moment bounds in (129), together with $b - r \ge a$, give the square integrability required to let the localization radii tend to infinity and obtain the identity on the full interval. □

G.5.1 Discretization on a positive-time interval

Proposition G.5 (Randomized time-grid discretization bound)**.** *Consider the randomized time-grid sampler in Section G.1. Let* $\widehat{s}$ *be trained independently of the test episode. Conditionally on the completed training run, assume that* $\widehat{s}$ *is finite-valued and jointly Borel and that* $\mathcal{E}_{\mathrm{score}}(\widehat{s}) < \infty$*. Then*

$$\mathbb{E}\mathsf{TV}\big(\widehat{Q}^{\mathrm{grid},h_{\mathrm{grid}}}_{t_0\,|\,H_n}, P_{t_0\,|\,H_n}\big) \le \varepsilon_{\mathrm{init}}(T - h_{\mathrm{grid}}) + \sqrt{\frac{\mathcal{E}_{\mathrm{score}}(\widehat{s})}{2}} + \varepsilon_{\mathrm{grid}}(h_{\mathrm{grid}}), \tag{131}$$

$$\varepsilon_{\mathrm{grid}}(h_{\mathrm{grid}}) := \left\{ \frac{d_z(T-t_0)h_{\mathrm{grid}}^2}{6} + d_z(d_z+3)e^{2h_{\mathrm{grid}}} h_{\mathrm{grid}} A_{t_0,T} \right\}^{1/2}, \quad h_{\mathrm{grid}} > 0. \tag{132}$$

The number of reverse updates is at most $\lceil A_{t_0,T}/h_{\mathrm{grid}} \rceil$ *almost surely.*

Proof of Proposition G.5. We first compare the exact reverse process with the process that uses the same initialization and freezes the score on each time-grid cell. The score difference splits into variation of the exact score within each cell and learned-score error at the grid nodes; averaging over the random shift controls the latter by the integrated score error. The finite-energy KL bound then gives a TV bound, after which we account for Gaussian initialization.

Condition on the learned score and fix a fresh prompt h in a common full-measure set on which the conditional integrated squared score error is finite.

To bound the variation of the exact score, consider one time-grid cell. Let Y_r be the exact reverse path initialized at its upper forward time and write $S_r = s^\star_{v_j - r}(Y_r \mid h)$ for $0 \le r \le \Delta_j = v_j - v_{j-1}$. Define the integrated squared error from freezing the exact score within each time-grid cell by

$$\mathcal{F}_h(U_{\mathrm{grid}}) := \sum_{j=0}^{J_{\mathrm{grid}}} \int_0^{\Delta_j} \mathbb{E}\|S_r - S_0\|_2^2\,\mathrm{d}r.$$

Applying Lemma G.4 with $(a,b) = (v_{j-1}, v_j)$ gives (129) and (130) on this cell.

The time grid satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{A}_{t_0}(t)} \log \sigma_t^2 = 2e^{-2t} \le 2.$$

Hence, whenever $t \in [v_{j-1}, v_j]$,

$$\Delta_j \le h_{\mathrm{grid}}\sigma_{v_j}^2 \le h_{\mathrm{grid}} e^{2h_{\mathrm{grid}}}\sigma_t^2, \qquad \Delta_j \le h_{\mathrm{grid}}. \tag{133}$$

Solving (130) on the cell gives

$$S_r - S_0 = (e^{-r} - 1)S_0 + \sqrt{2}\int_0^r e^{-(r-u)}\nabla s^\star_{v_j-u}(Y_u \mid h)\,\mathrm{d}W_u.$$

The two terms are orthogonal in L^2. Since $\int_0^{\Delta_j}(1-e^{-r})^2\,\mathrm{d}r \le \Delta_j^3/3$, $\mathbb{E}\|S_0\|_2^2 \le d_z/\sigma^2_{v_j}$, and $\Delta_j/\sigma^2_{v_j} \le h_{\mathrm{grid}}$, the summed drift contribution is at most $d_z(T-t_0)h^2_{\mathrm{grid}}/3$. Itô isometry bounds the summed martingale contribution by

$$2\sum_j \Delta_j \int_{v_{j-1}}^{v_j} \mathbb{E}\|\nabla s^\star_t(Z_t \mid h)\|_{\mathrm{F}}^2\,\mathrm{d}t.$$

Use $\Delta_j/\sigma_t^2 \le h_{\mathrm{grid}}e^{2h_{\mathrm{grid}}}$ and (129). Summing the cells gives

$$\mathcal{F}_h(U_{\mathrm{grid}}) \le \frac{d_z}{3}(T-t_0)h^2_{\mathrm{grid}} + 2d_z(d_z+3)e^{2h_{\mathrm{grid}}}h_{\mathrm{grid}}A_{t_0,T}. \tag{134}$$

This bound is uniform in h and U_{grid}.

We next control the learned-score error at the grid nodes. Since the learned score is only measurable, we use averaging over the random shift rather than a time-continuity bound. Put

$$e_h(t) := \mathbb{E}_{P_{t\mid h}}\|s^\star_t(Z_t \mid h) - \widehat{s}(Z_t, t, h)\|_2^2.$$

Joint Borel measurability of the scores and the conditional OU kernel makes e_h Lebesgue measurable, and the finite conditional integrated squared score error makes it integrable. A random grid shift hits any null set of exceptional times with probability zero; the countable union over lattice indices is still null. By (133), the learned-score grid contribution is at most

$$\mathcal{E}_h^{\mathrm{grid}}(U_{\mathrm{grid}}) := h_{\mathrm{grid}}\sum_{j=0}^{J_{\mathrm{grid}}}\sigma^2_{v_j}e_h(v_j).$$

Averaging the shifted lattice over U_{grid} and applying the change of variables $x = \mathsf{A}_{t_0}(t)$ gives exactly

$$\mathbb{E}_{U_{\mathrm{grid}}}\mathcal{E}_h^{\mathrm{grid}}(U_{\mathrm{grid}}) = \int_0^{A_{t_0,T}} \sigma^2_{\mathsf{A}^{-1}_{t_0}(x)}e_h\{\mathsf{A}^{-1}_{t_0}(x)\}\,\mathrm{d}x = \int_{t_0}^T e_h(t)\,\mathrm{d}t. \tag{135}$$

The two integrated squared-error bounds now control the difference between the exact and frozen reverse drifts. For almost every grid shift u, let $\mathsf{P}^\star_{h,u}$ be the exact reverse path law initialized from $P_{\widetilde{T}_{\mathrm{grid}}(u)\mid h}$, and let $\mathsf{Q}_{h,u}$ be the equally initialized recursively frozen path law. Represent both paths in the decreasing forward-time coordinate $t \in [t_0, \widetilde{T}_{\mathrm{grid}}(u)]$, write $Z_t^{\leftarrow}$ for the coordinate process under $\mathsf{P}^\star_{h,u}$, and let $\kappa_u(t) = v_j$ for almost every $t \in [v_{j-1}, v_j]$. Define the integrated squared score error along the full reverse path

$$\mathcal{I}_h(u) := \int_{t_0}^{\widetilde{T}_{\mathrm{grid}}(u)} \mathbb{E}_{\mathsf{P}^\star_{h,u}}\left[\left\|s^\star_t(Z_t^{\leftarrow} \mid h) - \widehat{s}(Z^{\leftarrow}_{\kappa_u(t)}, \kappa_u(t), h)\right\|_2^2\right]\,\mathrm{d}t.$$

The reference process is defined interval by interval by the explicit exponential–Euler formula, so finite-valued Borel evaluation of $\widehat{s}$ suffices to construct it with the score frozen on each interval. The reverse-drift difference is twice the score difference defining $\mathcal{I}_h(u)$, so the factor $1/4$ in (127) gives the first inequality below. Minkowski's inequality with respect to time and the exact reverse-path law, followed by (133), gives the remaining inequalities:

$$\begin{aligned}
\mathrm{KL}\big(\mathsf{P}^\star_{h,u}\|\mathsf{Q}_{h,u}\big) &\le \mathcal{I}_h(u),\\
\mathcal{I}_h(u)^{1/2} &\le \mathcal{F}_h(u)^{1/2} + \left\{\sum_{j=0}^{J_{\mathrm{grid}}}\Delta_j e_h(v_j)\right\}^{1/2}\\
&\le \mathcal{F}_h(u)^{1/2} + \mathcal{E}_h^{\mathrm{grid}}(u)^{1/2}.
\end{aligned}$$

Pinsker's inequality, Jensen's inequality in U_{grid}, (134), and (135) yield the last two terms in (131) after averaging over h.

It remains to replace the exact terminal distribution by the sampler's Gaussian initialization. Since $A_{t_0,T} - \mathsf{A}_{t_0}(\widetilde{T}_{\mathrm{grid}}) < h_{\mathrm{grid}}$ and $\mathrm{d}t/\mathrm{d}\mathsf{A} = \sigma_t^2 \le 1$, we have $\widetilde{T}_{\mathrm{grid}} \in [T - h_{\mathrm{grid}}, T]$. Invariance of $\mathcal{N}(0, I_{d_z})$ under the OU semigroup and TV contraction give the terminal term $\varepsilon_{\mathrm{init}}(T - h_{\mathrm{grid}})$. Convexity of TV in the first argument permits averaging the output distribution over U_{grid}. The update count follows because, with $x = A_{t_0,T}/h_{\mathrm{grid}}$ and $y = U_{\mathrm{grid}}/h_{\mathrm{grid}} \in (0,1)$, $J_{\mathrm{grid}} + 1 = \lfloor x - y \rfloor + 1 \le \lceil x \rceil$ almost surely. □

G.5.2 Generation error with a prescribed number of updates

The proof combines the grid-spacing bound in Proposition G.5 with the conditional stopping bound in Corollary G.2. The main theorem uses the following equivalent form, which is indexed by the number of steps.

For $0 < t_0 < T$ with $t_0 \le 1$, write $A = A_{t_0,T}$ and choose an integer

$$M \ge \max\{\lceil 2A \rceil, \lceil A/(T - t_0) \rceil\}. \tag{136}$$

Corollary G.6 (Randomized time-grid bound for at most M updates)**.** *Fix $\beta > 0$ and $L_\beta \ge 1$, and consider either the sample-only or the input–output prompt model in (19) and (20) with any Borel task prior Π. Suppose the task laws admit jointly Borel densities $p_\theta(z)$ in the sample-only model or $f_\theta(x, z)$ in the input–output model, with $p_\Theta \in \mathcal{F}_\beta^{d_z}(L_\beta)$ or $f_\Theta \in \mathcal{F}_\beta^{d_x+d_z}(L_\beta)$, respectively, Π-almost surely. Let $T > t_0 > 0$ with $t_0 \le 1$, and write $\beta' = \beta \wedge 1$. Suppose the learned score is independent of the test episode and, conditionally on the completed training run, finite-valued and jointly Borel with finite integrated squared score error. Let $M \in \mathbb{N}$ satisfy*

$$M \ge \max\left\{\lceil 2A_{t_0,T} \rceil, \left\lceil \frac{A_{t_0,T}}{T - t_0} \right\rceil\right\}, \qquad h_{\mathrm{grid}} := \frac{A_{t_0,T}}{M}.$$

When $T - t_0 \ge 1/2$, as under the parameter choices below for all sufficiently large prompt lengths, the simpler condition $M \ge \lceil 2A_{t_0,T} \rceil$ implies both requirements. Then the sampler uses at most M reverse updates almost surely, and its output distribution, averaged over the random grid shift, satisfies

$$\begin{aligned}\mathbb{E}\mathsf{TV}\left(\widehat{P}^{(M)}_{0\,|\,H_n}, P_{0\,|\,H_n}\right) \le{}& \varepsilon_{\mathrm{init}}(T - A_{t_0,T}/M) + \sqrt{\frac{\mathcal{E}_{\mathrm{score}}(\widehat{s})}{2}} + C_{\mathrm{stop}} t_0^{\beta'/2} \\ &+ \frac{A_{t_0,T}}{M}\sqrt{\frac{d_z(T - t_0)}{6}} + e^{A_{t_0,T}/M}\frac{A_{t_0,T}}{\sqrt{M}}\sqrt{d_z(d_z + 3)}.\end{aligned} \tag{137}$$

Here $\widehat{P}^{(M)}_{0\,|\,h} := \widehat{P}^{\mathrm{grid},A_{t_0,T}/M}_{0\,|\,h}$, so the parenthesized argument denotes a bound on the number of updates rather than a grid spacing. Moreover, there is a constant $0 < C_{\mathrm{disc}} = C_{\mathrm{disc}}(d_z) < \infty$ such that, for every $0 < \varepsilon_{\mathrm{disc}} \le 1$, the displayed conditions on M and $M \ge C_{\mathrm{disc}}\varepsilon_{\mathrm{disc}}^{-2}(A_{t_0,T} + 1)^2$ make the sum of the two discretization terms at most $\varepsilon_{\mathrm{disc}}$. Under the parameter choices used in (10), $A_{t_0,T} \lesssim \log(e/\varepsilon)$ and $\log(e/\varepsilon) \ge 1$, so $A_{t_0,T} + 1 \lesssim \log(e/\varepsilon) \lesssim \log \varepsilon^{-1}$ for $\varepsilon \le 1/2$. Thus, for a sufficiently large implicit constant, the condition on M in (10) implies both inequalities in (136) and the preceding discretization requirement with $\varepsilon_{\mathrm{disc}} = \varepsilon$. This is a sufficient number of sampling steps, not a minimax lower bound for the number of reverse updates.

Proof of Corollary G.6. The conditions on M give $A_{t_0,T}/M \le \min\{1/2, A_{t_0,T}, T - t_0\}$, so Proposition G.5 applies with $h_{\mathrm{grid}} = A_{t_0,T}/M$. Projection contracts TV, and the triangle inequality adds the conditional stopping error from Corollary G.2. With $h_{\mathrm{grid}} = A_{t_0,T}/M$, subadditivity of the square root in (132) gives

$$\varepsilon_{\mathrm{grid}}(A_{t_0,T}/M) \le \frac{A_{t_0,T}}{M}\sqrt{\frac{d_z(T - t_0)}{6}} + e^{A_{t_0,T}/M}\frac{A_{t_0,T}}{\sqrt{M}}\sqrt{d_z(d_z + 3)}.$$

The update count is $\lceil A_{t_0,T}/(A_{t_0,T}/M) \rceil = M$. Substitution proves the claim. Since $\sigma_t^2 \le 1$, one has $T - t_0 \le A_{t_0,T}$, and the condition $M \ge 2A_{t_0,T}$ gives $e^{A_{t_0,T}/M} \le e^{1/2}$. If $M \ge C_{\mathrm{disc}}\varepsilon_{\mathrm{disc}}^{-2}(A_{t_0,T} + 1)^2$, the sum of the two discretization terms is at most

$$\frac{\sqrt{d_z/6}}{C_{\mathrm{disc}}}\frac{A_{t_0,T}^{3/2}}{(A_{t_0,T} + 1)^2}\varepsilon_{\mathrm{disc}}^2 + \frac{e^{1/2}\sqrt{d_z(d_z + 3)}}{\sqrt{C_{\mathrm{disc}}}}\frac{A_{t_0,T}}{A_{t_0,T} + 1}\varepsilon_{\mathrm{disc}}.$$

Choosing $C_{\mathrm{disc}}(d_z)$ sufficiently large makes this expression at most $\varepsilon_{\mathrm{disc}}$ for every $0 < \varepsilon_{\mathrm{disc}} \le 1$. □

This bound for at most M updates applies to any learned score with finite integrated squared score error, regardless of the training objective. In the finite-sample results, the score-error bound comes from variance-balanced DSM, and (137) converts it into an explicit almost-sure bound on the number of reverse updates.

The Hölder condition is needed only to pass from the positive-time distribution to the data distribution; the positive-time discretization bound requires only the standing cube support and its stated score assumptions.

Under the same conditions, the preceding corollary gives the simpler form

$$\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0\,|\,H_n}, P(\cdot\,|\,H_n)\big) \lesssim \underbrace{\varepsilon_{\mathrm{init}}(T - A_{t_0,T}/M)}_{\text{Gaussian initialization}} + \underbrace{\sqrt{\mathcal{E}_{\mathrm{score}}(\widehat{s})}}_{\text{learned score}} + \underbrace{t_0^{(\beta\wedge 1)/2}}_{\text{early stopping}} + \underbrace{A_{t_0,T}/\sqrt{M}}_{\text{discretization}}. \quad (138)$$

The implicit constant depends only on the fixed dimension and regularity bounds. Indeed, $T - t_0 \le A_{t_0,T} \le M/2$, so both discretization terms in (137) are $\mathcal{O}(A_{t_0,T}/\sqrt{M})$. The square root converts integrated squared-score error $\mathcal{O}(\varepsilon^2)$ into generation error $\mathcal{O}(\varepsilon)$.

The formal posterior-predictive guarantee is stated in Theorem 3.1. The next three subsections complete its proof by choosing the common score class and diffusion parameters, verifying the required tail bounds, and combining the approximation, learning, and sampling results.

G.6 End-to-End Parameter Choices for Target Accuracy

Fix the common parameters in Assumption 3.1, and put $\beta' = \beta \wedge 1$. All constants below depend only on these parameters and dimensions. Choose $0 < \varepsilon \le \varepsilon_0$, where a sufficiently small fixed ε_0 ensures $0 < t_0 < T$, $t_0 \le 1$, and the radius inequalities below. For every integer $n \ge 1$, set

$$t_0 = \left(\frac{\varepsilon}{4C_{\mathrm{stop}}}\right)^{2/\beta'}, \qquad T = \max\left\{1, 1 + \log\frac{4\sqrt{d_z/2}}{\varepsilon}\right\}, \qquad A = A_{t_0,T}. \quad (139)$$

Here $C_{\mathrm{stop}} > 0$ is any fixed valid constant in the early-stopping TV bound. Choose fixed constants satisfying

$$c_{\mathrm{state}} \ge 2\sqrt{d_z}, \qquad c_{\mathrm{state}}^2/16 > 2 + 2/\beta', \qquad c_{\mathrm{lab}}^2/2 > 2,$$

and choose

$$R_{\mathrm{state},n} = \sqrt{d_z} + c_{\mathrm{state}}\sqrt{\log(e/\varepsilon)}, \qquad R_{\mathrm{out},n} = R_{\mathrm{lab},n} = \sqrt{d_z} + c_{\mathrm{lab}}\sqrt{\log(e/\varepsilon)},$$
$$R_{\mathrm{bal},n} = \frac{R_{\mathrm{state},n} + \sqrt{d_z}}{\sigma_{t_0}}. \quad (140)$$

The dependence on ε in all quantities carrying a subscript n is suppressed only in this construction for a fixed accuracy. We denote the resulting class and trainable count explicitly by $\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}$ and $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$. The symbols $\mathcal{S}^{\mathrm{lat}}_{n}$, $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n}$, and $B_{\mathrm{lat},n}$ in the supporting lemmas suppress this fixed accuracy; in the main text they are written $\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}$, $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$, and $B_{\mathrm{lat},n,\varepsilon}$. When the theorem is applied at the prompt rate, these symbols always mean $\varepsilon = \sqrt{\log(en)/n}$.

For a fixed sufficiently small $c_{\mathrm{acc}} > 0$, choose the grid by

$$m_{\theta,n} = \left\lceil \sqrt{d_\theta}\left(1 + \frac{(n+1)H_\theta^2}{c_{\mathrm{acc}}\underline{p}\,\varepsilon^2}\right)^{1/(2\beta_\theta)} \right\rceil, \qquad J_{\theta,n} = m_{\theta,n}^{d_\theta},$$
$$\eta_{\theta,n} = \sqrt{d_\theta}/m_{\theta,n}, \qquad \varepsilon_{\mathrm{floor},n} = c_{\mathrm{acc}}\varepsilon^2. \quad (141)$$

Use the regular cells and representatives of Section D.3. Their maximum distance is at most $\eta_{\theta,n}$. This gives

$$D_n(\eta_{\theta,n}, \varepsilon_{\mathrm{floor},n}) \le 2c_{\mathrm{acc}}\varepsilon^2, \qquad J_{\theta,n} \lesssim \left(1 + \frac{\sqrt{n+1}}{\varepsilon}\right)^{d_\theta/\beta_\theta}. \quad (142)$$

The extra one in (141) makes this choice valid also when $H_\theta = 0$. The prior-logit radius is bounded by $\log(2J_{\theta,n}/\varepsilon_{\mathrm{floor},n}) = \mathcal{O}(\log(en/\varepsilon))$.

Set $\varepsilon_{\mathrm{tar},n} = c_{\mathrm{acc}}\varepsilon/\sqrt{A}$ and choose the neural tolerances

$$\varepsilon_{\ell,n} = \frac{\varepsilon_{\mathrm{tar},n}}{18n\mathsf{C}_{\mathrm{net},n}}, \qquad \varepsilon_{\mathrm{qry},n} = \frac{\varepsilon_{\mathrm{tar},n}}{18\mathsf{C}_{\mathrm{net},n}},$$
$$\varepsilon_{a,n} = \frac{\varepsilon_{\mathrm{tar},n}}{18(1+R^{\mathrm{lat}}_{\mathrm{num},n}\underline{q}^{-2}_{\mathrm{out}})}, \qquad \varepsilon_{b,n} = \frac{\varepsilon_{\mathrm{tar},n}}{18(1+\underline{q}^{-1}_{\mathrm{out}})},$$
$$\varepsilon_{\mathrm{rat},n} = \varepsilon_{\mathrm{tar},n}/18, \qquad \varepsilon_{\mathrm{cov},n} = \frac{c_{\mathrm{acc}}\varepsilon^2}{A(R_{\mathrm{out},n}+R_{\mathrm{lab},n})}. \tag{143}$$

For sample-only prompts set $\varepsilon_{\mathrm{qry},n} = 0$ and omit that module. Decreasing ε_0 makes all nonzero component tolerances at most one and $R_{\mathrm{out},n} \le R_{\mathrm{bal},n}$. These choices are deterministic, independent of the unknown family and prior, and fixed before observing training data. They do not depend on N or the requested confidence level.

G.7 Tail Bounds for the Selected Radii

Lemma G.7 (Gaussian-tail bounds for the selected radii)**.** *Use the diffusion times and radii in (139) and (140), with $\beta' = \beta \wedge 1$. Let $\delta_{\mathrm{tail},n}$ and $\varepsilon^2_{\mathrm{tail},n}$ be the probability and score-energy terms in (39) and (40). There is $\varepsilon_1 \in (0,1]$, depending only on d_z, β', C_{stop}, c_{state}, and c_{lab}, such that, for $0 < \varepsilon \le \varepsilon_1$ and $c_{\mathrm{state}} \ge 2\sqrt{d_z}$, the following bounds hold with implicit constants depending only on the same quantities:*

$$\sigma_{t_0}^{-1} \lesssim \varepsilon^{-1/\beta'}, \qquad A \lesssim \log(e/\varepsilon), \tag{144}$$
$$R^2_{\mathrm{bal},n} \lesssim \log(e/\varepsilon)\varepsilon^{-2/\beta'}, \qquad \delta_{\mathrm{tail},n} \le \varepsilon^{c^2_{\mathrm{state}}/8}, \tag{145}$$
$$\varepsilon^2_{\mathrm{tail},n} \lesssim \log(e/\varepsilon)\varepsilon^{c^2_{\mathrm{state}}/16-2/\beta'}, \tag{146}$$
$$AR^2_{\mathrm{bal},n}\delta_{\mathrm{tail},n} \lesssim \log^2(e/\varepsilon)\varepsilon^{c^2_{\mathrm{state}}/8-2/\beta'}, \tag{147}$$
$$\varepsilon^2_{\mathrm{lab,clip}}\{R_{\mathrm{lab},n}\} \lesssim \log(e/\varepsilon)\varepsilon^{c^2_{\mathrm{lab}}/2}. \tag{148}$$

Consequently, if

$$\frac{c^2_{\mathrm{state}}}{16} > 2 + \frac{2}{\beta'}, \qquad \frac{c^2_{\mathrm{lab}}}{2} > 2, \tag{149}$$

then

$$4AR^2_{\mathrm{bal},n}\delta_{\mathrm{tail},n} + 4\varepsilon^2_{\mathrm{tail},n} + \varepsilon^2_{\mathrm{lab,clip}}\{R_{\mathrm{lab},n}\} = o(\varepsilon^2). \tag{150}$$

In particular, the left-hand side is at most $c_1\varepsilon^2$ for every fixed $c_1 > 0$ and all sufficiently small ε.

Proof. For $0 < t \le 1$, concavity of $t \mapsto 1 - e^{-2t}$ gives

$$\sigma_t^2 = 1 - e^{-2t} \ge (1 - e^{-2})t.$$

After decreasing ε_1 so that $t_0 \le 1$, this proves the first bound in (144). Moreover, $e^{2t_0} - 1 \ge 2t_0$ and $e^{2T} - 1 \le e^{2T}$ imply

$$A \le T + \frac{1}{2}\log\{1/(2t_0)\} \lesssim \log(e/\varepsilon).$$

The definition of $R_{\mathrm{bal},n}$ now gives its bound in (145).

Let $X := \|\xi\|_2$ for $\xi \sim \mathcal{N}(0, I_{d_z})$ and put $a := c_{\mathrm{state}}\sqrt{\log(e/\varepsilon)}$. The Gaussian norm concentration inequality gives

$$\Pr(X > a) \le \exp\{-\tfrac{1}{2}(a - \sqrt{d_z})^2\} \le e^{-a^2/8} \le \varepsilon^{c^2_{\mathrm{state}}/8},$$

where the second inequality uses $a \ge 2\sqrt{d_z}$. This and (39) prove the probability bound. Cauchy–Schwarz and the finite fourth Gaussian moment give

$$\mathbb{E}\Big[(X + 2\sqrt{d_z})^2\,\mathbb{1}\{X > a\}\Big] \le \Big\{\mathbb{E}(X + 2\sqrt{d_z})^4\Big\}^{1/2}\Pr(X > a)^{1/2} \lesssim \varepsilon^{c^2_{\mathrm{state}}/16}.$$

Substitution into (41), together with (144), proves (146). Multiplying the bounds for A, $R^2_{\mathrm{bal},n}$, and $\delta_{\mathrm{tail},n}$ proves (147). Finally, apply (106) with $x = c_{\mathrm{lab}}\sqrt{\log(e/\varepsilon)}$ to obtain (148). Under (149), each exponent remaining after division by ε^2 is strictly positive. Since $\varepsilon^a\{1 + \log(\varepsilon^{-1})\}^k \to 0$ for every $a > 0$ and fixed k, (150) follows. □

G.8 Proof of the Posterior-Predictive Generation Guarantee

For the score class in Theorem G.8, combine its logarithmic covering number with the confidence term:

$$\Gamma_{n,\varepsilon,\delta} := \log\left(\frac{2N_{\mathrm{cov}}(\varepsilon_{\mathrm{cov},n}, \mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}, d_{\mathrm{bal},\infty})}{\delta}\right) = \log N_{\mathrm{cov}}\big(\varepsilon_{\mathrm{cov},n}, \mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}, d_{\mathrm{bal},\infty}\big) + \log(2/\delta).$$

The cover centers belong to the class, and its radius is fixed by (143). The following calculation establishes the approximation and complexity claims used in Theorem G.8. It combines Proposition E.6 with the covering-number bound for the entire score class in Proposition E.7. In particular, the approximation and parameter-count bounds hold for the same deterministic class used in training.

The chosen scaled-output and label radii $R_{\mathrm{out},n}$ and $R_{\mathrm{lab},n}$ are both $\mathcal{O}(\sqrt{\log(e/\varepsilon)})$, and $A \lesssim \log(e/\varepsilon)$. On the training event in Theorem G.8, the more detailed score and predictive bounds are

$$\mathcal{E}_{\mathrm{score}}(\widehat{s}_{N,n,\varepsilon}) \lesssim \varepsilon^2 + \varepsilon_{\mathrm{opt}} + \frac{A(R_{\mathrm{out},n} + R_{\mathrm{lab},n})^2 \Gamma_{n,\varepsilon,\delta}}{N}, \tag{151}$$

$$\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0\,|\,H_n}, P(\cdot \mid H_n)\big) \lesssim \varepsilon + \sqrt{\varepsilon_{\mathrm{opt}}} + (R_{\mathrm{out},n} + R_{\mathrm{lab},n})\sqrt{\frac{A\Gamma_{n,\varepsilon,\delta}}{N}} + \frac{A}{\sqrt{M}}.$$

Using the definition of $\Gamma_{n,\varepsilon,\delta}$, the radius bounds, and (153), and imposing $\varepsilon_{\mathrm{opt}} \lesssim \varepsilon^2$, gives (9) and the sufficient choices in (10).

Theorem G.8 (Posterior-predictive generation; formal version of Theorem 3.1)**.** *Under Assumption 3.1, for each $n \geq 2$ and sufficiently small $\varepsilon > 0$, there exists a common DiT score class $\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}$ and diffusion parameters that depend on (n, ε) and the fixed dimensions and regularity bounds. Its number of trainable parameters $\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}$, stored nonzero coefficients, and arithmetic operations per forward pass are polynomial in (n, ε^{-1}). Use the estimator and reverse sampler of Section 2.4, with $N \geq 1$ independent pretraining episodes, empirical optimization gap $\varepsilon_{\mathrm{opt}} \lesssim \varepsilon^2$, and an integer M satisfying (136). For every fixed $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over the training episodes and their diffusion augmentations,*

$$\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0|H_n}, P(\cdot \mid H_n)\big) \lesssim \varepsilon + \log(e/\varepsilon)\sqrt{\frac{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}\log(en/\varepsilon) + \log(2/\delta)}{N}} + \frac{\log(e/\varepsilon)}{\sqrt{M}}. \tag{152}$$

The expectation averages a fresh prompt conditional on the completed training run.

Proof of Theorem G.8. We first verify approximation and covering for the deterministic class. We then apply the score-learning inequality and transfer its error to the randomized sampler on the same training event.

The grid and floor choices give (142). Substitution of (143) into the definition of $\varepsilon_{\mathrm{net},n}$ gives $\varepsilon_{\mathrm{net},n} \leq \varepsilon_{\mathrm{tar},n}/3$. Lemmas E.5 and G.7 give

$$AR^2_{\mathrm{bal},n}\delta_{\mathrm{tail},n} + \varepsilon^2_{\mathrm{tail},n} + \varepsilon^2_{\mathrm{out,clip},n} + \varepsilon^2_{\mathrm{lab,clip}}(R_{\mathrm{lab},n}) \lesssim \varepsilon^2.$$

In particular the output-clipping contribution is at most $2A\exp(-c^2_{\mathrm{lab}}\log(e/\varepsilon)/2) = o(\varepsilon^2)$. Applying Proposition E.6 proves

$$\mathcal{E}_{\mathrm{app}}(\mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}) \lesssim \varepsilon^2.$$

The quantitative size and sensitivity bounds in Proposition E.7 hold for arbitrary grids and component tolerances. They therefore apply directly to the parameter choices made above for the integrated error. Here t_0^{-1}, T, $R_{\mathrm{bal},n}$, the grid size, all inverse nonzero tolerances, and the prior-logit radius are polynomially bounded in n and ε^{-1}. The local OU moduli and the numbers of components are likewise polynomial. The fixed-depth attention sensitivity recursion and the clipped readout give polynomial parameter radius and parameter sensitivity over the entire class. Consequently, using the bounds $B_{\mathrm{lat},n}$ and $\mathsf{Lip}^{\mathrm{par}}_{\mathrm{lat},n}$ from Proposition E.7, evaluated at the choices above,

$$\log N_{\mathrm{cov}}(u, \mathcal{S}^{\mathrm{lat}}_{n,\varepsilon}, d_{\mathrm{bal},\infty}) \leq \mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}\log\left(1 + 2B_{\mathrm{lat},n}\mathsf{Lip}^{\mathrm{par}}_{\mathrm{lat},n}/u\right),$$
$$\Gamma_{n,\varepsilon,\delta} \lesssim \mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\varepsilon}\log(en/\varepsilon) + \log(2/\delta). \tag{153}$$

This establishes the approximation and complexity claims used in Theorem G.8. The exact empirical minimizer exists measurably by the compactness and continuity argument in Proposition E.7; the theorem permits any measurable approximate minimizer with the specified scaled-objective gap.

Apply Theorem F.2 with its predictor radius $R_{\mathrm{pred}} = R_{\mathrm{out},n}$, the cover radius in (143), and optimization tolerance $\varepsilon_{\mathrm{opt}}$. With probability at least $1-\delta$, this gives (151). The uniform output bound and $t \geq t_0 > 0$ ensure that the learned score is finite-valued and Borel on the full state domain, as required by Corollary G.6. In that corollary, $A/M \leq 1/2$ makes the shifted terminal-time error $\lesssim \varepsilon$ by (139); the early-stopping error is at most $\varepsilon/4$. Also $A \geq T - t_0$ and $M \geq 2A$, so

$$\frac{A\sqrt{T-t_0}}{M} \leq \frac{A}{\sqrt{M}}.$$

Taking square roots of (151) proves the main predictive bound. The task-conditional consequence is derived separately in Corollary H.4 by combining this predictive bound with (3). Finally $A \lesssim \log(e/\varepsilon)$ and $(R_{\mathrm{out},n} + R_{\mathrm{lab},n})^2 \lesssim \log(e/\varepsilon)$ give the displayed logarithmic form. The probability statement is for each model separately, with uniformly bounded failure probability and common deterministic constants. □

The sufficient choices in (10) make the predictive error $\mathcal{O}(\varepsilon)$. Choose their structural constant so that both inequalities required of the integer-valued choices hold. They are statistical and real-arithmetic bounds; they do not guarantee efficient optimization of the empirical loss.

G.9 Scalar Sample-Only Complexity Specialization

For sample-only prompts with $d_z = d_\theta = \beta = \beta_\theta = 1$, $\delta = n^{-1}$, and $\varepsilon = \sqrt{\log(en)/n}$, the parameter choices based on the episode-level KL bound give the following sufficient bounds under exact real arithmetic:

$$J_{\theta,n} = \widetilde{\mathcal{O}}(n), \qquad \mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n} = \widetilde{\mathcal{O}}(n^5), \qquad \mathsf{N}^{\mathrm{lat}}_{\mathrm{store},n} = \widetilde{\mathcal{O}}(n^5), \qquad \mathsf{N}^{\mathrm{lat}}_{\mathrm{op},n} = \widetilde{\mathcal{O}}(n^5),$$
$$N = \widetilde{\mathcal{O}}(n^6), \qquad M = \widetilde{\mathcal{O}}(n). \tag{154}$$

Here $\widetilde{\mathcal{O}}$ suppresses powers of $\log(en)$.

Derivation of (154). Let $d_z = d_\theta = 1$, $\beta = \beta' = \beta_\theta = 1$, with sample-only prompts and $\delta = n^{-1}$. In this paragraph $\widetilde{\mathcal{O}}$ hides powers of $\log(en)$ only. Let $t_{0,n}$ and T_n denote the choices in (139) evaluated at $\varepsilon = \sqrt{\log(en)/n}$. With $\varepsilon = \sqrt{\log(en)/n}$, the diffusion and radius choices give

$$t_{0,n} \asymp \log(en)/n, \qquad \sigma^{-1}_{t_{0,n}}, R_{\mathrm{bal},n}, R^{\mathrm{lat}}_{\mathrm{num},n} = \widetilde{\mathcal{O}}(n^{1/2}), \qquad J_{\theta,n} = \widetilde{\mathcal{O}}(n).$$

The state, final-output, and label radii, T_n, and $A_{t_{0,n},T_n}$ are logarithmic in n. Formula (143) gives

$$\varepsilon^{-1}_{\ell,n} = \widetilde{\mathcal{O}}(n^2), \quad \varepsilon^{-1}_{a,n} = \widetilde{\mathcal{O}}(n), \quad \varepsilon^{-1}_{b,n} = \widetilde{\mathcal{O}}(n^{1/2}), \quad \varepsilon^{-1}_{\mathrm{rat},n} = \widetilde{\mathcal{O}}(n^{1/2}).$$

There are $\mathcal{O}(\log n)$ overlapping time patches. On patches with $a_j \leq 1$, $h^{\mathrm{loc}}_j \leq 7a_j$ and $\sigma^2_{a_j} \asymp a_j$; on later patches, $\sigma^{-1}_{a_j} = \mathcal{O}(1)$ and $h^{\mathrm{loc}}_j \leq T_n$. Since $H_{q,x} = 0$, (74) and (75) give

$$\max_j H_{a,j,n} = \widetilde{\mathcal{O}}(n), \qquad \max_j H_{b,j,n} = \widetilde{\mathcal{O}}(n^{3/2}).$$

The two-dimensional state–time basis size is therefore

$$\mathsf{N}^{\mathrm{lat}}_{\mathrm{basis},n} = \widetilde{\mathcal{O}}\Big(n\{(n\,n)^2 + (n^{3/2}n^{1/2})^2\}\Big) = \widetilde{\mathcal{O}}(n^5).$$

The likelihood and fixed ratio module sizes are $\widetilde{\mathcal{O}}(n^3)$ and $\widetilde{\mathcal{O}}(n^2)$, respectively. The token width is $\mathcal{O}(J_{\theta,n})$; the attention projection count is $\widetilde{\mathcal{O}}(n^2)$ and its forward cost is $\widetilde{\mathcal{O}}((n + J_{\theta,n})J^2_{\theta,n}) = \widetilde{\mathcal{O}}(n^3)$. Repeated likelihood evaluation costs $\widetilde{\mathcal{O}}(n^4)$. Thus the basis term dominates the numbers of trainable parameters, stored coefficients, and forward-pass operations, each bounded by $\widetilde{\mathcal{O}}(n^5)$. Substitution into (10) with $\varepsilon = \sqrt{\log(en)/n}$ gives $N = \widetilde{\mathcal{O}}(n^6)$ and $M = \widetilde{\mathcal{O}}(n)$. □

The direct implementation uses at most $M\mathsf{N}^{\mathrm{lat}}_{\mathrm{op},n} = \widetilde{\mathcal{O}}(n^6)$ arithmetic operations for a full sample in this scalar example. The training sample contains $N(n+1) = \widetilde{\mathcal{O}}(n^7)$ observed outputs; these outputs are not generally independent within episodes. The large exponents make the sufficient nature of the construction explicit. They are not empirical scaling laws.

H SHARP FINITE-PROMPT TASK UNCERTAINTY

This appendix first states the finite-prompt uncertainty theorem, then proves it in three steps: posterior averaging compares task uncertainty with oracle risk, predictive information gives the upper bound, and an Assouad construction gives the matching lower bound. The final subsection combines this result with Theorem 3.1 to obtain the task-conditional guarantee. The input–output argument keeps the fresh query in the conditioning variables, and the family used for the lower bound varies along only one latent coordinate while the class bounds remain fixed.

H.1 FINITE-PROMPT THEOREM AND ORACLE RISK

Even exact posterior prediction can differ from the realized task distribution. We therefore quantify this difference over the model class $\mathfrak{M}_{\mathrm{lat}}$ defined after Assumption 3.1, whose members share the same dimensions and regularity bounds. For each $\mathcal{M}$, the symbols $\Pr_{\mathcal{M}}$, $\mathbb{E}_{\mathcal{M}}$, and $\mathsf{Law}_{\mathcal{M}}$ refer to its induced joint law; the dependence of the kernels P on $\mathcal{M}$ is suppressed. For a known model $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$, define the oracle task risk

$$\mathcal{R}_n^{\mathrm{or}}(\mathcal{M}) := \inf_{Q^{\mathcal{M}}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(Q^{\mathcal{M}}(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big), \tag{155}$$

where the infimum is over Borel probability kernels from $\mathcal{H}_n$ to $\mathcal{Z}$ that may use the entire task family and task prior. Generation randomness is included in the output distribution. An independent random seed cannot improve the infimum by convexity of TV. Lemma H.2 and the Borel posterior-predictive kernel fixed in (16) imply

$$\mathcal{R}_n^{\mathrm{or}}(\mathcal{M}) \le \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \le 2\mathcal{R}_n^{\mathrm{or}}(\mathcal{M}). \tag{156}$$

A lower bound on this oracle risk therefore limits every generator, including one with unlimited independent pretraining.

Theorem H.1 (Uniform upper and matching lower bounds for finite-prompt task uncertainty)**.** Uniform upper bound. *For every $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ and $n \ge 1$, the following bound holds with an implicit constant depending only on the common class parameters:*

$$\mathbb{E}_{\mathcal{M}} \mathrm{KL}\big(P(\cdot \mid H_n, \Theta) \,\|\, P(\cdot \mid H_n)\big) \lesssim \frac{\log(en)}{n}. \tag{157}$$

Consequently,

$$\mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \lesssim \sqrt{\frac{\log(en)}{n}}$$

uniformly over $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$.

Matching lower bound. *If $H_\theta > 0$, $\underline{p} < 1$, and $L_\beta > 1$, then, for all sufficiently large n,*

$$\begin{aligned} \sqrt{\frac{\log(en)}{n}} &\lesssim \sup_{\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}} \mathcal{R}_n^{\mathrm{or}}(\mathcal{M}) \\ &\le \sup_{\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \lesssim \sqrt{\frac{\log(en)}{n}}. \end{aligned} \tag{158}$$

The lower bound allows any number of independent training episodes, including a countably infinite sequence, and unrestricted computation at generation time. The implicit constants and the sufficiently-large-n threshold depend only on the common class parameters. The worst-case model and prior may depend on n, while all class parameters remain fixed.

Asymptotic interpretation of the upper bound. The uniform upper bound gives

$$\mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \lesssim \sqrt{\log(en)/n}.$$

Thus the expected TV error tends to zero. In particular, for every fixed $\rho_{\mathrm{TV}} > 0$, Markov's inequality gives

$$\begin{aligned} &\Pr_{\mathcal{M}}\big(\mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) > \rho_{\mathrm{TV}}\big) \\ &\quad \le \frac{1}{\rho_{\mathrm{TV}}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \lesssim \frac{1}{\rho_{\mathrm{TV}}} \sqrt{\frac{\log(en)}{n}}. \end{aligned}$$

Hence the TV error converges to zero in probability under the joint law of the realized task and prompt for each fixed $\mathcal{M}$. This is convergence of the posterior-predictive output distribution to the realized task-conditional output distribution. It does not establish convergence for every fixed θ, almost-sure convergence, or contraction of $\Pi(\cdot \mid H_n)$ to δ_Θ. Recovery of a unique latent parameter would additionally require identifiability; predictive convergence itself does not, since different parameters may induce the same distribution.

The upper bound applies to every $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$ without the additional nondegeneracy conditions. The matching lower bound shows that the upper bound's dependence on n cannot be improved uniformly over the stated nondegenerate class, even when the family and prior are known; it does not assert that every fixed family or prior attains this rate. The predictive KL upper bound is proved in Section H.3, and the matching lower bound is proved in Section H.4.

H.2 Posterior Averaging and Oracle Comparison

The posterior-predictive distribution averages the task law conditional on the prompt. The following lemma bounds its expected task-conditional TV error by twice that of any prompt-based probability kernel.

Lemma H.2 (Posterior averaging bound)**.** *Suppose the posterior-mixture identity in (2) holds, and let $h \mapsto Q(\cdot \mid h)$ be any measurable probability kernel from $\mathcal{H}_n$ to $\mathcal{Z}$. Then*

$$\mathbb{E}\left[\mathsf{TV}(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta))\right] \le 2\mathbb{E}\left[\mathsf{TV}(Q(\cdot \mid H_n), P(\cdot \mid H_n, \Theta))\right].$$

Proof. Write $Q_h = Q(\cdot \mid h)$ and $R_{h,\theta} = P(\cdot \mid h, \theta)$. On the common full-measure set where the posterior-mixture identity holds, convexity of TV gives

$$\mathsf{TV}(Q_h, P(\cdot \mid h)) \le \int \mathsf{TV}(Q_h, R_{h,\theta})\Pi(\mathrm{d}\theta \mid h).$$

These TV distances are measurable and bounded by one: on a standard Borel space, the defining supremum can be taken over a countable generating algebra. Taking expectations and applying the tower property yields

$$\mathbb{E}\mathsf{TV}(Q(\cdot \mid H_n), P(\cdot \mid H_n)) \le \mathbb{E}\mathsf{TV}(Q(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)).$$

The triangle inequality through $Q(\cdot \mid H_n)$ proves the claim. □

The lemma uses only the posterior-mixture identity in (2) and the measurability of $Q(\cdot \mid H_n)$ as a function of the prompt. The TV bound requires no density, smoothness, conditional i.i.d., or diffusion assumption. Taking the infimum over Q gives the factor-two oracle comparison in (156); the reverse inequality follows by choosing the posterior-predictive kernel.

Under TV loss, the posterior-predictive distribution need not minimize posterior expected TV loss: with posterior weights 0.9 and 0.1 on δ_0 and δ_1, its posterior expected loss is 0.18, whereas δ_0 has loss 0.1. Thus, Lemma H.2 is a factor-two posterior-averaging bound, not an exact Bayes-optimality statement.

H.3 Upper Bound via Predictive Information

Fix either the sample-only or the input–output prompt model and the common dimensions and regularity bounds in Assumption 3.1, and recall the corresponding model class $\mathfrak{M}_{\mathrm{lat}}$ defined after that assumption.

Under the positive latent-family assumption, the information supplied by the next output is bounded in terms of the cumulative information in the prompt. This comparison concerns the realized task and its exact posterior predictive. The episode-KL argument in Proposition D.4 instead compares true and discretized models to control score approximation.

For standard-Borel random variables, write

$$I(A; B \mid C) := \mathbb{E}\mathrm{KL}\big(\mathsf{Law}(B \mid A, C) \,\|\, \mathsf{Law}(B \mid C)\big)$$

for conditional mutual information, and omit C for unconditional mutual information. The reference-mixture bound below ensures finiteness of the information quantities used in the proof, whose identities follow from the KL chain rule.

Lemma H.3 (From a latent KL modulus to one-step prediction). *Fix $d_\theta \in \mathbb{N}$ and a probability kernel $\theta \mapsto G_\theta$ on a standard Borel observation space. Let Θ have any Borel prior Π on $[0,1]^{d_\theta}$, and, conditionally on $\Theta = \theta$, let $U_1, U_2, \dots$ be i.i.d. with law G_θ. Suppose that, for some $K_{\mathrm{KL}} \in [0,\infty)$ and $\beta_\theta \in (0,1]$,*

$$\mathrm{KL}(G_\theta \| G_{\theta'}) \le K_{\mathrm{KL}} \|\theta - \theta'\|_2^{2\beta_\theta} \quad \text{for all } \theta, \theta'. \tag{159}$$

There is a constant C_0, depending only on $(d_\theta, \beta_\theta, K_{\mathrm{KL}})$, such that for every integer $m \ge 1$,

$$I(\Theta; U_{1:m}) \le C_0 + \frac{d_\theta}{2\beta_\theta} \log m. \tag{160}$$

Moreover, for every $n \ge 0$,

$$\mathbb{E}\mathrm{KL}\big(G_\Theta \,\|\, \mathsf{Law}(U_{n+1} \mid U_{1:n})\big) \le \frac{C_0 + \frac{d_\theta}{2\beta_\theta}\log(n+1)}{n+1}. \tag{161}$$

Proof. Fix $\varepsilon \in (0,1]$. The latent cube admits a Euclidean ε-net $\{\theta_1, \dots, \theta_{J_\varepsilon}\}$ with $J_\varepsilon \le C_{d_\theta} \varepsilon^{-d_\theta}$. For example, subdivide each coordinate into intervals of length at most $\varepsilon/\sqrt{d_\theta}$ and use their centers. Define the reference law

$$Q_{m,\varepsilon} := J_\varepsilon^{-1} \sum_{j=1}^{J_\varepsilon} G_{\theta_j}^{\otimes m}.$$

For each θ, choose a nearest net point. The mixture dominates $J_\varepsilon^{-1} G_{\theta_j}^{\otimes m}$, so product additivity of KL and (159) give

$$\mathrm{KL}(G_\theta^{\otimes m} \| Q_{m,\varepsilon}) \le \log J_\varepsilon + m K_{\mathrm{KL}} \varepsilon^{2\beta_\theta}. \tag{162}$$

Let $\overline{G}_m := \int G_\theta^{\otimes m} \Pi(\mathrm{d}\theta)$ be the actual marginal law of the examples. Splitting the log density ratio through $\overline{G}_m$ and integrating gives

$$\int \mathrm{KL}(G_\theta^{\otimes m} \| Q_{m,\varepsilon}) \Pi(\mathrm{d}\theta) = I(\Theta; U_{1:m}) + \mathrm{KL}(\overline{G}_m \| Q_{m,\varepsilon}).$$

The nonnegative final term can be dropped. Choose $\varepsilon = m^{-1/(2\beta_\theta)}$ in (162). This proves (160) after absorbing $\log C_{d_\theta} + K_{\mathrm{KL}}$ into C_0. The equal weights in $Q_{m,\varepsilon}$ are used only to obtain the bound; the actual prior Π has not been replaced or restricted.

Put $\mathcal{I}_k = I(\Theta; U_{k+1} \mid U_{1:k})$. The chain rule gives

$$\sum_{k=0}^{m-1} \mathcal{I}_k = I(\Theta; U_{1:m}).$$

To prove monotonicity, set $F = U_{1:k}$, $A = U_{k+1}$, and $B = U_{k+2}$. Applying the mutual-information chain rule to $I((\Theta, A); B \mid F)$ in the two possible orders gives

$$I(\Theta; B \mid F) + I(A; B \mid \Theta, F) = I(A; B \mid F) + I(\Theta; B \mid A, F).$$

Conditional independence makes $I(A; B \mid \Theta, F) = 0$. Conditional exchangeability identifies the first term with $\mathcal{I}_k$ and the final term with $\mathcal{I}_{k+1}$. Hence

$$\mathcal{I}_k - \mathcal{I}_{k+1} = I(U_{k+1}; U_{k+2} \mid U_{1:k}) \ge 0.$$

Therefore

$$(n+1)\mathcal{I}_n \le \sum_{k=0}^{n} \mathcal{I}_k = I(\Theta; U_{1:n+1}).$$

Finally, conditional independence gives

$$\mathcal{I}_n = \mathbb{E}\mathrm{KL}\big(G_\Theta \,\|\, \mathsf{Law}(U_{n+1} \mid U_{1:n})\big).$$

Combining the last two displays with (160) proves (161). The preceding finite bound also justifies every mutual-information chain-rule identity used here. □

Proof of the upper bounds in Theorem H.1. Fix $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$. Let $U_i = Z_i$ in the sample-only model and $U_i = (X_i, Z_i)$ in the input–output model. Let G_θ denote its one-example law, with density g_θ on the unit-volume cube. Using $\log x \le x - 1$ and the normalization of the two densities,

$$\mathrm{KL}(G_\theta \| G_{\theta'}) \le \int \frac{(g_\theta - g_{\theta'})^2}{g_{\theta'}} \le \frac{1}{\underline{p}} \|g_\theta - g_{\theta'}\|_\infty^2 \le \frac{H_\theta^2}{\underline{p}} \|\theta - \theta'\|_2^{2\beta_\theta}.$$

Thus Lemma H.3 applies uniformly over all allowed task maps and all Borel priors.

In the sample-only model, the one-step predictive law is $P(\cdot \mid C_n)$, and the oracle is P_Θ. (161) proves the asserted KL bound.

In the input–output model, put $\mathcal{I}_n = I(\Theta; U_{n+1} \mid U_{1:n})$. Conditional independence and the KL chain rule give

$$\begin{aligned} &\mathbb{E}_{\mathcal{M}} \mathrm{KL}\big(P_\Theta^{Z|X}(\cdot \mid X_\star) \,\|\, P(\cdot \mid C_n, X_\star)\big) \\ &\quad = I(\Theta; Z_\star \mid C_n, X_\star) \\ &\quad \le I(\Theta; (X_\star, Z_\star) \mid C_n) = \mathcal{I}_n. \end{aligned}$$

The inequality follows by dropping the nonnegative term $I(\Theta; X_\star \mid C_n)$. In particular, the argument does not require a task-independent input marginal and still conditions the predictive distribution on the query $X_\star$. This proves (157) in the second model.

Pinsker's inequality $\mathsf{TV}(P, Q)^2 \le \mathrm{KL}(P\|Q)/2$ and Jensen's inequality yield

$$\begin{aligned} &\mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(P(\cdot \mid H_n), P(\cdot \mid H_n, \Theta)\big) \\ &\quad \le \sqrt{\frac{1}{2} \mathbb{E}_{\mathcal{M}} \mathrm{KL}\big(P(\cdot \mid H_n, \Theta) \,\|\, P(\cdot \mid H_n)\big)} \lesssim \sqrt{\frac{\log(en)}{n}}. \end{aligned}$$

□

The lower bound in Theorem H.1 is proved in Section H.4. Together with the preceding upper bounds and the posterior-averaging inequality, it gives the stated oracle-risk comparison. The constants depend only on the common class parameters. When $H_\theta = 0$ or $\underline{p} = 1$, all tasks within a model have the same law and both predictive KL and task uncertainty are zero.

H.4 Lower Bound via a Smooth Latent Assouad Family

Proof of the lower bound in Theorem H.1. For the lower bound, we construct a hypercube of smooth densities, apply an Assouad testing argument (Tsybakov, 2009), and embed the entire family into a one-dimensional latent path.

Choose a nonzero $\psi \in C_c^\infty((0,1))$ with $\int \psi = 0$ and $\|\psi\|_\infty \le 1$, extended by zero outside $(0,1)$. Write $b_1 = \int_0^1 |\psi| > 0$ and $b_2 = \int_0^1 \psi^2 > 0$. For sufficiently large n, set

$$m = \lfloor \tfrac{1}{4} \log_2 n \rfloor \ge 1, \qquad a = \kappa \sqrt{m/n}, \qquad K = 2^m,$$

where $\kappa > 0$ is a fixed small constant chosen below. For $v \in \{-1, 1\}^m$, define

$$p_v(z) = 1 + a \sum_{j=1}^m v_j \psi(m z_1 - j + 1), \qquad z \in [0,1]^{d_z}.$$

The perturbations have disjoint supports and zero integral, so p_v is a density with $1 - a \le p_v \le 1 + a$.

We check spatial regularity under (24), including integer β. Recall $m_\beta = \lceil \beta \rceil - 1$ and $\vartheta_\beta = \beta - m_\beta \in (0, 1]$ from Appendix B.4, and write $\widetilde{p}_v = (p_v - 1)/a$. Disjointness gives $\|\partial_1^s \widetilde{p}_v\|_\infty \le C_s m^s$ for every fixed integer $s \ge 0$; derivatives involving other coordinates vanish. The m_βth derivative has both a uniform bound $C m^{m_\beta}$ and a Lipschitz bound $C m^{m_\beta + 1}$. Thus

$$|\partial_1^{m_\beta} \widetilde{p}_v(z) - \partial_1^{m_\beta} \widetilde{p}_v(z')| \lesssim \min\{m^{m_\beta}, m^{m_\beta+1} \|z - z'\|_2\} \le C m^\beta \|z - z'\|_2^{\vartheta_\beta}.$$

Consequently $\|p_v\|_{C^\beta} \le 1 + C_{\beta, d_z, \psi} a m^\beta$. Since $a m^\beta \to 0$, all these densities satisfy the fixed spatial norm bound $L_\beta > 1$ and lower density bound $\underline{p} < 1$ for large n.

Let P_v have density p_v, let $\mathbb{E}_v := \mathbb{E}_{P_v^{\otimes n}}$, and let d_{H} denote Hamming distance. Direct integration gives

$$\mathsf{TV}(P_v, P_w) = \frac{ab_1}{m} d_{\mathrm{H}}(v, w), \qquad \mathrm{KL}(P_v^{\otimes n} \| P_w^{\otimes n}) \le \frac{8na^2 b_2}{m} = 8\kappa^2 b_2 \quad \text{if } d_{\mathrm{H}}(v, w) = 1.$$

For the KL bound, take $a \le 1/2$ and use $\mathrm{KL}(P_v \| P_w) \le \int (p_v - p_w)^2 / p_w$ on the single differing bump. Choose $0 < \kappa \le \min\{1, (8\sqrt{b_2})^{-1}\}$. Pinsker's inequality then gives $\mathsf{TV}(P_v^{\otimes n}, P_w^{\otimes n}) \le 1/2$ for neighboring vertices.

For any Borel probability-valued estimator Q_D, where $D = Z_{1:n}$, let $\widehat{v}(Q_D)$ be the lexicographically first minimizer of $w \mapsto \mathsf{TV}(Q_D, P_w)$. This is measurable: TV from a fixed probability measure on the cube is a supremum over a countable generating algebra. The definition of $\widehat{v}(Q_D)$ and the triangle inequality imply

$$\mathsf{TV}(Q_D, P_v) \ge \tfrac{1}{2} \mathsf{TV}(P_{\widehat{v}(Q_D)}, P_v) = \frac{ab_1}{2m} d_{\mathrm{H}}(\widehat{v}(Q_D), v).$$

For each bit, pair the vertices that differ only in that bit. The sum of the two testing-error probabilities in each pair is at least $1 - \mathsf{TV}(P_v^{\otimes n}, P_w^{\otimes n}) \ge 1/2$. Averaging uniformly over all vertices therefore gives error probability at least $1/4$ for each bit. Summing over the m bits yields

$$2^{-m} \sum_v \mathbb{E}_v \mathsf{TV}(Q_D, P_v) \ge \frac{ab_1}{8} \gtrsim \sqrt{\frac{\log(en)}{n}},$$

because $m \asymp \log(en)$. The estimator need not have a density. Randomized probability-valued estimators satisfy the same bound by averaging their output measures over the auxiliary randomization and using convexity of TV.

We next embed all $K = 2^m$ densities into the fixed latent class. Enumerate the vertices as $v^{(1)}, \dots, v^{(K)}$, put $\lambda_j = (j-1)/(K-1)$, and let $\bar{p}_\lambda$ interpolate linearly between $p_{v^{(j)}}$ and $p_{v^{(j+1)}}$ on each interval $[\lambda_j, \lambda_{j+1}]$. Convexity preserves normalization, positivity, and the spatial norm. Since $\|p_v - p_w\|_\infty \le 2a$, summing the intervalwise Lipschitz bounds gives

$$\|\bar{p}_\lambda - \bar{p}_{\lambda'}\|_\infty \le 2a(K-1)|\lambda - \lambda'|, \qquad 2aK \le Cn^{-1/4}\sqrt{\log n} \longrightarrow 0.$$

For sufficiently large n, this Lipschitz constant is at most the fixed $H_\theta > 0$. Set $g_\theta(z) = \bar{p}_{\theta_1}(z)$ for $\theta \in [0,1]^{d_\theta}$. This map is jointly Borel and satisfies

$$\|g_\theta - g_{\theta'}\|_\infty \le H_\theta |\theta_1 - \theta_1'| \le H_\theta |\theta_1 - \theta_1'|^{\beta_\theta} \le H_\theta \|\theta - \theta'\|_2^{\beta_\theta},$$

using $|\theta_1 - \theta_1'| \le 1$. Choose $\Pi_n = K^{-1} \sum_{j=1}^K \delta_{(\lambda_j, 0, \dots, 0)}$. The resulting environment belongs to $\mathfrak{M}_{\mathrm{lat}}$ with all class bounds fixed, and its oracle risk is at least $c\sqrt{\log(en)/n}$ by the testing bound obtained above under the uniform prior on the vertices. Although the hypercube has m bits, its embedding uses only one latent coordinate.

For input–output prompts, set $g_\theta(x, z) = \bar{p}_{\theta_1}(z)$ on the unit input–output cube. The same spatial and latent regularity checks apply. The inputs, including $X_\star$, are uniform and independent of the task and output context. Conditional on all these inputs, any kernel $Q(\cdot \mid C_n, X_\star)$ is an estimator based on the output context to which the preceding bound applies. Averaging over the inputs gives the same lower bound.

Finally, fix the constructed environment $\mathcal{M}_n$. For any rule using independent training episodes, write $\widehat{Q}_N^{\mathcal{M}_n}(\cdot \mid \mathsf{d}, h)$ for its output kernel given a training-data realization d and prompt h, and define the Borel kernel

$$\overline{Q}_h(A) := \int \widehat{Q}_N^{\mathcal{M}_n}(A \mid \mathsf{d}, h) \mathsf{Law}_{\mathcal{M}_n}(\mathcal{D}_N^{\mathrm{epi}})(\mathrm{d}\mathsf{d}).$$

Independence of training from the fresh task and prompt, together with convexity of TV, shows that $\overline{Q}$ has no greater expected loss than the original rule. The same construction applies to a countably infinite training sequence on its standard-Borel product space. Hence independent pretraining cannot reduce the lower bound even when the entire environment is known. □

The lower bound is a uniform statement over a fixed class. It does not assert the rate $\sqrt{\log(en)/n}$ for every fixed latent family or fixed prior. Additional identifiability or local regularity may yield a log-free parametric rate for particular models. No positive lower bound in (158) is asserted when $H_\theta = 0$ or $\underline{p} = 1$.

H.5 Task-Conditional Consequence and Minimaxity

Combining the predictive guarantee with the finite-prompt uncertainty bound gives a bound on the learned generator's error relative to the task distribution. To compare it with a single learning rule that does not know the family or prior, define, for finite $N \geq 0$, the unknown-model minimax task risk

$$\mathcal{R}^{\mathrm{lat}}_{n,N} := \inf_{Q} \sup_{\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(Q(\cdot \mid \mathcal{D}^{\mathrm{epi}}_N, H_n), P(\cdot \mid H_n, \Theta)\big). \tag{163}$$

Here $\mathbb{E}_{\mathcal{M}}$ averages training and the fresh episode. The infimum ranges over Borel learning rules, including randomized rules, that know the common class parameters but are not given the family or prior. In $\mathbb{E}_{\mathrm{train},\mathcal{M}}\mathbb{E}_{\mathcal{M}}$, the inner expectation averages a fresh task and prompt conditional on training, and the outer expectation averages training, augmentations, and fitting randomness under $\mathcal{M}$. The learning rule does not depend on which model in the class generates the data.

Corollary H.4 (Task-conditional generation at the sharp prompt rate)**.** *For all sufficiently large n, fix $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$, $N \geq 1$, $\delta \in (0,1)$, and an integer M satisfying (136). Set $\varepsilon = \sqrt{\log(en)/n}$, and use the score class, estimator, and sampler of Theorem 3.1 with $\varepsilon_{\mathrm{opt}} \leq c_{\mathrm{opt}} \log(en)/n$. With probability at least $1-\delta$ over the training episodes and their diffusion augmentations,*

$$\begin{aligned} &\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0 \mid H_n}, P(\cdot \mid H_n, \Theta)\big) \\ &\quad \lesssim \sqrt{\frac{\log(en)}{n}} + \log(en) \sqrt{\frac{\mathsf{N}^{\mathrm{lat}}_{\mathrm{param},n,\sqrt{\log(en)/n}} \log(en) + \log(2/\delta)}{N}} + \frac{\log(en)}{\sqrt{M}}. \end{aligned} \tag{164}$$

On the same event, the displayed bound also holds for $\mathbb{E}\mathsf{TV}\big(\widehat{P}^{(M)}_{0 \mid H_n}, P(\cdot \mid H_n)\big)$. If N and M satisfy (10) with $\varepsilon = \sqrt{\log(en)/n}$, then both TV errors are $\lesssim \sqrt{\log(en)/n}$ and the integrated score error in (151) is $\lesssim \log(en)/n$. The implicit constants may also depend on c_{opt}. Under the nondegeneracy conditions of Theorem H.1, take $\delta = n^{-1}$ and these sufficient choices of N and M. Then

$$\begin{aligned} \sqrt{\frac{\log(en)}{n}} &\lesssim \mathcal{R}^{\mathrm{lat}}_{n,N} \\ &\leq \sup_{\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}} \mathbb{E}_{\mathrm{train},\mathcal{M}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(\widehat{P}^{(M)}_{0 \mid H_n}, P(\cdot \mid H_n, \Theta)\big) \lesssim \sqrt{\frac{\log(en)}{n}}. \end{aligned} \tag{165}$$

The lower bound holds for every number of independent training episodes, even if the model is known. The upper bound requires the stated choices of N, M, and the optimization tolerance.

The proof below combines posterior-predictive error with finite-prompt task uncertainty. Since $\log(en) \asymp \log n$ for sufficiently large n, this is the same rate stated in the main text and gives the same orders of N and M. A smaller target ε can reduce posterior-predictive error at fixed n, but it does not reduce the uncertainty caused by observing only n examples from the fresh task; individual task families may nevertheless have faster rates or zero uncertainty.

H.5.1 Proof of the task-conditional prompt-rate corollary

Proof of Corollary H.4. Set $\varepsilon = \sqrt{\log(en)/n}$ and restrict to n large enough that $\varepsilon \leq \varepsilon_0$. Then $\log(e/\varepsilon) \lesssim \log(en)$ and $\log(en/\varepsilon) \lesssim \log(en)$. Applying Theorem 3.1 and the covering bound in (153) gives the posterior-predictive part of (164); (151) gives the stated score bound. The uniform upper bound in Theorem H.1 is $\lesssim \sqrt{\log(en)/n}$, so the triangle inequality gives the task-conditional bound on the same training event.

For the upper minimax bound, the score class chosen above and the fitting rule are common to all $\mathcal{M} \in \mathfrak{M}_{\mathrm{lat}}$. Under each model, the conditional task risk is $\lesssim \sqrt{\log(en)/n}$ on an event of probability at least $1 - n^{-1}$ and at most one otherwise. Hence

$$\mathbb{E}_{\mathrm{train},\mathcal{M}} \mathbb{E}_{\mathcal{M}} \mathsf{TV}\big(\widehat{P}^{(M)}_{0 \mid H_n}, P(\cdot \mid H_n, \Theta)\big) \lesssim \sqrt{\frac{\log(en)}{n}} + n^{-1} \lesssim \sqrt{\frac{\log(en)}{n}}.$$

The constants and sufficient choices are common, so taking the supremum over models preserves this bound. The minimax infimum is no larger than the risk of this particular common rule. Existence

is ensured by the Borel exact ERM in Proposition E.7; any common Borel fitting rule meeting the prescribed approximate-ERM tolerance satisfies the same upper bound.

For the lower bound, the construction in Section H.4 gives, for each sufficiently large n, one model in $\mathfrak{M}_{\mathrm{lat}}$ on which every probability-valued estimator has expected task risk at least $c\sqrt{\log(en)/n}$, even if given the model and independent training data. This applies to every unknown-model learning rule in (163). Taking its supremum risk and then the infimum over rules proves the first inequality in (165). Averaging the predictive bound in the same way gives expected posterior-predictive TV $\lesssim \sqrt{\log(en)/n}$, without a matching $\sqrt{\log(en)/n}$ lower claim for that risk. □

I Examples of Smooth Latent Task Families

This section illustrates the scope of Assumption 3.1 in two layers. We first verify one sample-only and one input–output family directly, and then give three reusable construction templates for generating further admissible families.

I.1 Direct Sample-Only and Input–Output Families

The following family illustrates the common-support and latent-regularity conditions in Assumption 3.1. Its posterior predictive is examined in Section J.2.

Example I.1 (Smoothly varying sample densities)**.** *In the sample-only model, let $d_z = d_\theta = 1$ and*

$$g_\theta(z) = 1 + \tfrac{1}{4}\cos\big(2\pi(z-\theta)\big), \qquad (\theta, z) \in [0,1]^2.$$

The latent coordinate shifts the density's peak without changing its support. Integration over one period gives normalization, and

$$\tfrac{3}{4} \le g_\theta \le \tfrac{5}{4}, \qquad \|g_\theta\|_{C^1} \le \tfrac{5}{4} + \tfrac{\pi}{2}, \qquad \sup_z |g_\theta(z) - g_{\theta'}(z)| \le \tfrac{\pi}{2}|\theta - \theta'|,$$

since both $|\partial_z g_\theta|$ and $|\partial_\theta g_\theta|$ are at most $\pi/2$. Joint continuity and these uniform bounds verify Assumption 3.1 with $\beta = \beta_\theta = 1$, $\underline{p} = 3/4$, $L_1 = 5/4 + \pi/2$, and $H_\theta = \pi/2$, for any Borel prior. The prompt consists of samples from one task; its posterior predictive averages over possible peak locations.

This family also has a simple coefficient-space geometry. Expanding the shifted cosine gives

$$g_\theta(z) = 1 + \tfrac{1}{4}\cos(2\pi\theta)\cos(2\pi z) + \tfrac{1}{4}\sin(2\pi\theta)\sin(2\pi z).$$

Thus, the coefficient vector $\gamma(\theta) = \frac{1}{4}\big(\cos(2\pi\theta), \sin(2\pi\theta)\big)$ traces a circle of radius $1/4$. Given H_n, the posterior-predictive density is

$$\bar{g}_n(z) := \int g_\theta(z)\,\Pi(\mathrm{d}\theta \mid H_n) = 1 + \bar{a}_n\cos(2\pi z) + \bar{b}_n\sin(2\pi z),$$
$$(\bar{a}_n, \bar{b}_n) := \int \gamma(\theta)\,\Pi(\mathrm{d}\theta \mid H_n).$$

Consequently, $\sqrt{\bar{a}_n^2 + \bar{b}_n^2} \le 1/4$, with strict inequality unless the posterior distribution of $\gamma(\Theta)$ is a point mass. The task densities therefore lie on the circle, while their posterior predictive can lie in its interior.

Example I.2 (Smoothly varying conditional output densities)**.** *Let $d_x = d_z = 1$, let $\psi(x,z) = (2x-1)(2z-1)$, and set*

$$g_\theta(x,z) = 1 + \rho(2\theta_1 - 1)\psi(x,z), \qquad \theta \in [0,1]^{d_\theta}, \quad (x,z) \in [0,1]^2, \quad 0 < \rho < 1.$$

These are jointly continuous probability densities with $g_\theta \ge 1 - \rho$ and a common finite spatial C^β bound for every fixed $\beta > 0$. Moreover,

$$\sup_{x,z} |g_\theta(x,z) - g_{\theta'}(x,z)| \le 2\rho\|\psi\|_\infty\|\theta - \theta'\|_2.$$

Thus, for sufficiently small ρ, this family satisfies Assumption 3.1 with density floor $1 - \rho$ and $\beta_\theta = 1$, for any Borel prior. Because $\int_0^1 \psi(x,z)\,\mathrm{d}z = 0$ for every x, the input marginal is uniform, while the conditional output density varies with the task.

I.2 General Construction Templates

Example I.3 (Smooth perturbations of the uniform density)**.** *Let $\psi \in C^\infty([0,1]^{d_{\mathrm{stat}}})$ satisfy $\int_{[0,1]^{d_{\mathrm{stat}}}} \psi(u)\,\mathrm{d}u = 0$ and $\|\psi\|_\infty \le 1$, and define*

$$g_\theta(u) = 1 + \rho(2\theta_1 - 1)\psi(u), \qquad \theta \in [0,1]^{d_\theta}.$$

For sufficiently small $\rho > 0$, this family satisfies Assumption 3.1 with $\beta_\theta = 1$.

Example I.4 (Smoothly varying multimodal mixtures)**.** *Let $\varphi_1, \dots, \varphi_K$ be fixed smooth densities with separated modes, let $\Delta^{K-1} := \{w \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$, and let $w : [0,1]^{d_\theta} \to \Delta^{K-1}$ be Hölder continuous. Then*

$$g_\theta(u) = (1 - \gamma_{\mathrm{mix}}) + \gamma_{\mathrm{mix}} \sum_{k=1}^{K} w_k(\theta)\varphi_k(u)$$

satisfies Assumption 3.1 for sufficiently small $\gamma_{\mathrm{mix}} > 0$. Thus, the relative weights of the modes may vary across tasks.

Example I.5 (Normalized exponential families)**.** *Fix $r \in \mathbb{N}$, let $\phi : [0,1]^{d_{\mathrm{stat}}} \to \mathbb{R}^r$, and let $\eta_{\mathrm{nat}} : [0,1]^{d_\theta} \to \mathbb{R}^r$. The following is a normalized exponential family:*

$$g_\theta(u) = \frac{\exp\{\eta_{\mathrm{nat}}(\theta)^\top \phi(u)\}}{\int_{[0,1]^{d_{\mathrm{stat}}}} \exp\{\eta_{\mathrm{nat}}(\theta)^\top \phi(v)\}\, \mathrm{d}v}.$$

It satisfies Assumption 3.1 when the coordinates of ϕ have bounded C^β norms, $\eta_{\mathrm{nat}}([0,1]^{d_\theta})$ is compact, and η_{nat} is β_θ-Hölder continuous. Hence, the class includes exponential families with bounded parameters, as well as truncated Gaussian families.

J Numerical Illustrations

J.1 Posterior-Predictive Generation in a Synthetic Image Task

To illustrate the distributional behavior motivated by our theory, we construct a synthetic episodic image-transformation task with an exact, generally bimodal posterior-predictive distribution. The task is designed so that a short context leaves the sign of a latent transformation rule ambiguous, whereas a longer context identifies the sign and contracts the query distribution. We use the experiment to ask whether a diffusion Transformer can represent both modes under an ambiguous context, contract toward the appropriate mode as more demonstrations are observed, and learn this prompt-to-distribution map from increasing numbers of pretraining episodes while preserving the query content.

Following controlled synthetic experiments in previous ICL studies (e.g., Panwar et al., 2024; Wakayama & Suzuki, 2026; Ma et al., 2025), we use a simple task family in which the latent rule and its posterior predictive can be calculated exactly.

J.1.1 Experimental Setup

Each episode has a shared binary latent variable

$$G \in \{-1, +1\}, \qquad \Pr(G = -1) = \Pr(G = +1) = \frac{1}{2}. \tag{166}$$

Conditioned on G, the episode-specific position rule is

$$c_i = a + b r_i + \varepsilon_i, \quad a \sim \mathcal{N}(0.5, 0.04^2), \quad b \mid G \sim \mathcal{N}(0.30G, 0.06^2), \quad \varepsilon_i \sim \mathcal{N}(0, 0.05^2), \tag{167}$$

where r_i is a normalized descriptor of object size and c_i is the latent horizontal-position coordinate of the output. The corresponding horizontal translation in pixels is

$$\Delta x_i = 18(c_i - 0.5). \tag{168}$$

No vertical translation is applied. The variables a, b, and G are shared within an episode, whereas the observation noises are independent.

For the context examples, we draw $r_i \sim \mathrm{Unif}[-0.2, 0.2]$, while the query descriptor is fixed at $r_q = 0.9$. Thus, the context observes the rule only near its intercept, whereas the query requires extrapolation away from the observed range. With few examples, the context may be compatible with both signs of b, which produces left and right predictive modes at the query. As the context grows, the posterior probability of one sign concentrates and the posterior-predictive distribution contracts toward the corresponding mode.

Images are 32×32 RGB and are generated from six shape families: circle, triangle, star, square, blob, and glyph. The query shape family is excluded from its context, so the task cannot be solved by memorizing a transformation associated with a particular family. An input X_i contains a dark object on a bright background, and the corresponding output Y_i contains the same object in blue after applying the horizontal translation in Eq. (168). We write $C_n = \{(X_i, Y_i)\}_{i=1}^n$ for the rendered image context. We do not use shadows, textures, or vertical transformations.

The exact posterior predictive is available in closed form (Gelman et al., 2013). Let $\theta = (a, b)^\top$, $\phi(r) = (1, r)^\top$, $m_g = (0.5, 0.30g)^\top$, $S_0 = \mathrm{diag}(0.04^2, 0.06^2)$, and $\sigma_\varepsilon = 0.05$. For computing the exact reference law for C_n, let $r_{1:n}$ and $c_{1:n}$ denote the simulator-side descriptors and latent output positions used to render its context pairs, and write $\Phi_n \in \mathbb{R}^{n \times 2}$ for the matrix with rows $\phi(r_i)^\top$ and $c_{1:n} = (c_1, \ldots, c_n)^\top$. Conditioned on $G = g$, Gaussian conjugacy gives

$$S_{g,n} = \left(S_0^{-1} + \sigma_\varepsilon^{-2} \Phi_n^\top \Phi_n\right)^{-1}, \quad m_{g,n} = S_{g,n}\left(S_0^{-1} m_g + \sigma_\varepsilon^{-2} \Phi_n^\top c_{1:n}\right). \tag{169}$$

The posterior sign probability is obtained from the Gaussian marginal likelihood,

$$\pi_{g,n} := \Pr(G = g \mid C_n) = \frac{\frac{1}{2}\mathcal{N}\left(c_{1:n}; \Phi_n m_g, \sigma_\varepsilon^2 I_n + \Phi_n S_0 \Phi_n^\top\right)}{\sum_{g' \in \{-1,+1\}} \frac{1}{2}\mathcal{N}\left(c_{1:n}; \Phi_n m_{g'}, \sigma_\varepsilon^2 I_n + \Phi_n S_0 \Phi_n^\top\right)}. \tag{170}$$

Hence, with

$$\mu_{g,n} = \phi(r_q)^\top m_{g,n}, \qquad \sigma^2_{g,n} = \sigma^2_\varepsilon + \phi(r_q)^\top S_{g,n}\phi(r_q), \tag{171}$$

the exact query distribution is the two-component Gaussian mixture

$$p(c_q \mid C_n) = \sum_{g\in\{-1,+1\}} \pi_{g,n}\,\mathcal{N}(c_q; \mu_{g,n}, \sigma^2_{g,n}). \tag{172}$$

This distribution is used as ground truth in all distributional evaluations. Although the simulator retains $r_{1:n}$ and $c_{1:n}$ to compute (172), the learned model receives only the image context C_n and the diffusion-conditioning variables described below.

We train a cross-attention DiT (Peebles & Xie, 2023) on this episodic task. Each context pair (X_i, Y_i) is concatenated along the channel dimension and mapped by a convolutional pair encoder to a 2×2 grid of four 512-dimensional tokens. The model accepts at most 64 pairs, or 256 context tokens in total. No positional embedding identifies the order of the context pairs. The query input and the noisy target are also concatenated channel-wise and embedded by a convolution with patch size 2, producing a 16×16 grid of 256 query tokens. The network contains 12 DiT blocks, each consisting of self-attention among the query tokens, cross-attention from the query tokens to the context tokens, and an MLP. The hidden dimension is 512, the number of attention heads is 8, the feed-forward dimension is 2,048, and the model has 80,271,180 trainable parameters. Diffusion-time and context-length embeddings condition the blocks through adaptive layer normalization (AdaLN), with the latter computed from the number of valid context pairs; the implementation follows the public DiT codebase.[2] The model is given the context input–output images, the query input image, the noisy target, the diffusion time, and a valid-token mask, but not G, a, b, r_i, c_i, or the analytical posterior.

Training uses freshly generated episodes at every optimizer update, similar to Garg et al. (2022) and Wakayama & Suzuki (2026). We sample one context length uniformly from $n \in \{1, 2, 4, 8, 16, 32, 64\}$ and process 64 independent episodes with that same length. The microbatch size is 8 and gradients are accumulated for 8 microbatches, giving an effective batch size of 64. We train for 40,000 optimizer updates, corresponding to 2.56 million episodes. The clean query output x_0 is scaled to $[-1, 1]$. A diffusion time is sampled from 1,000 steps with a cosine noise schedule (Nichol & Dhariwal, 2021), Gaussian noise is added to obtain x_t, and the model directly predicts the clean target rather than the noise:

$$\mathcal{L} = \mathbb{E}\left[\|\widehat{x}_0(x_t, t, X_q, C_n) - x_0\|_2^2\right]. \tag{173}$$

We use AdamW (Kingma & Ba, 2015; Paszke et al., 2019) with learning rate 2×10^{-4}, weight decay 0.01, a 1,000-step linear warmup followed by cosine decay, and gradient clipping at 1.0. At evaluation time, we draw images with deterministic DDIM sampling (Song et al., 2021a) using 200 reverse steps and $\eta = 0$. All reported results use the raw model weights.

We evaluate generated images by recovering the object center. For test task $b \in \{1, \ldots, B\}$, context length n, and generated sample k, let $\widehat{c}_{b,k,n}$ denote the recovered latent position and define

$$\widehat{P}^c_{b,n} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{c}_{b,k,n}}, \qquad K = 256. \tag{174}$$

Let $P^{c,\mathrm{post}}_{b,n}$ denote the exact mixture in (172). We report three quantitative metrics.

- *Right-probability error.* The boundary $c = 0.5$ separates the left and right predictive modes. We compare the generated right-mode frequency with its exact posterior-predictive probability:

$$E_{\mathrm{right}}(b, n) = \left| \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[\widehat{c}_{b,k,n} > 0.5] - \Pr(c_q > 0.5 \mid C_{b,n}) \right|. \tag{175}$$

  This metric isolates whether the model assigns the correct probability mass to the two alternatives.

[2] https://github.com/facebookresearch/DiT/blob/main/models.py

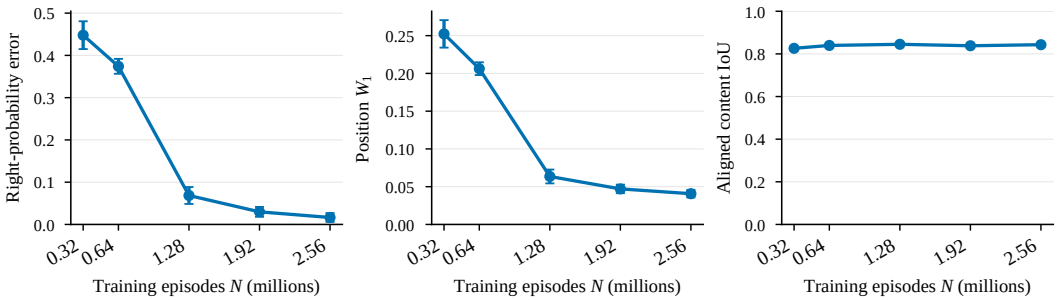

Figure 3: **Scaling with the number of training episodes.** Performance at 0.32, 0.64, 1.28, 1.92, and 2.56 million episodes. From left to right, the panels show the right-probability error, the 1-Wasserstein distance between the recovered generated-position distribution and the exact two-component Gaussian posterior predictive, and aligned content IoU. Evaluation uses the same 64 fixed test tasks, $n = 64$, 256 generations per task, and the raw model weights. Error bars indicate the standard error across tasks.

- *Posterior-distribution accuracy.* We measure the discrepancy in position space by

$$E_{W_1}(b, n) = W_1\left(\widehat{P}^{c}_{b,n}, P^{c,\mathrm{post}}_{b,n}\right). \tag{176}$$

  Unlike the right-probability error, this quantity jointly reflects mixture weights, mode locations, and within-mode dispersion.
- *Aligned content preservation.* To separate shape preservation from position inference, we undo the horizontal translation estimated from each generated image before comparing object masks. Let M^{query}_b be the query mask and $\widetilde{M}^{\mathrm{gen}}_{b,k,n}$ the generated mask after alignment. We report

$$\mathrm{IoU}_{\mathrm{align}}(n) = \frac{1}{BK}\sum_{b=1}^{B}\sum_{k=1}^{K}\frac{\left|\widetilde{M}^{\mathrm{gen}}_{b,k,n} \cap M^{\mathrm{query}}_b\right|}{\left|\widetilde{M}^{\mathrm{gen}}_{b,k,n} \cup M^{\mathrm{query}}_b\right|}. \tag{177}$$

To visualize posterior contraction directly, we additionally compare the full empirical distribution of recovered generated positions with the exact posterior-predictive distribution for a fixed task and query across nested context lengths. We also inspect independent generated images and their pixel-wise average $\overline{Y}_K = K^{-1}\sum_{k=1}^{K} Y^{(k)}$. For a multimodal position distribution, this average can be spatially blurred even when individual samples are sharp and valid.

J.1.2 Results

We first examine the effect of the number of training episodes. Fig. 3 evaluates checkpoints at 5,000, 10,000, 20,000, 30,000, and 40,000 optimizer updates, corresponding to $N \in \{0.32, 0.64, 1.28, 1.92, 2.56\}$ million episodes. The same 64 fixed test tasks are evaluated at context length $n = 64$ with 256 generations per task using the raw weights. As N increases, the mean right-probability error decreases from 0.448 to 0.016, and position W_1 decreases from 0.252 to 0.041. Both distributional metrics improve most sharply between 0.64 and 1.28 million episodes. In contrast, aligned content IoU remains approximately 0.83–0.85 throughout training. Thus, the model learns to render and preserve the query object early, while additional episodes primarily improve inference of the latent rule and calibration of the posterior-predictive position distribution.

We next study test-time scaling with the context length. Fig. 4 shows the recovered position distribution for a representative fixed task (task 12) under nested contexts of lengths $n \in \{1, 2, 4, 8, 16, 32\}$. At $n = 1$, the sign of the latent slope is not identified, and both the exact posterior predictive and the model distribution have left and right modes. As more examples are observed, the posterior probability over the slope sign concentrates and the exact distribution contracts toward the left mode. The model reproduces this overall transition, but contraction is delayed in the intermediate regime. In particular, at $n = 8$, the exact posterior predictive has already placed almost all of its mass on the left mode, whereas the model still assigns visible probability mass to the right mode. By $n = 16$,

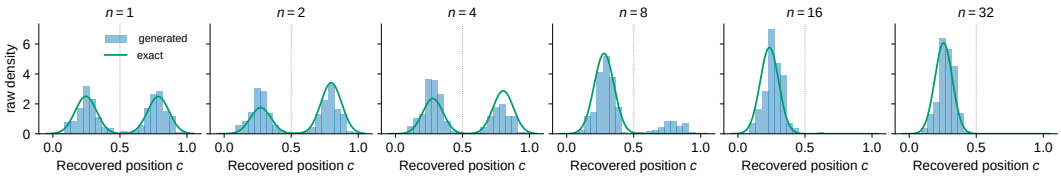

Figure 4: **Posterior-predictive contraction with context length.** Each column corresponds to $n \in \{1, 2, 4, 8, 16, 32\}$ for the same fixed task and query. Histograms show positions recovered from 256 generated images using the raw model weights, and green curves show the analytically exact posterior-predictive distribution. The dotted line at $c = 0.5$ separates the left and right modes. The model captures the transition from a broad bimodal predictive law at short contexts to a narrow, essentially unimodal law at long contexts, with visible excess right-mode mass in the intermediate regime.

the generated distribution is narrow and essentially unimodal, closely matching the exact posterior predictive, and the agreement persists at larger context lengths. Because the panels follow one realized episode, newly observed noise can transiently change the posterior mode probabilities, so the degree of concentration need not evolve monotonically at every intermediate context length.

Finally, Fig. 5 gives qualitative examples of posterior-predictive generation for a circle query and a triangle query, comparing $n = 1$ and $n = 8$ with the raw weights. For the circle task, the exact right-mode probability changes from 0.550 to 0.999, while the generated frequency changes from 0.562 to 0.965. For the triangle task, the exact right-mode probability instead changes from 0.527 to 0.009, while the generated frequency changes from 0.410 to 0.055. Thus, a short context leaves both horizontal placements plausible, whereas the longer context selects the right mode for the circle and the left mode for the triangle. The individual generations preserve the query shape in both cases. The pixel averages at $n = 1$ are blurred along the horizontal direction because they combine samples from both modes, whereas the averages at $n = 8$ are sharp after the posterior predictive has concentrated. A single averaged image obscures the multimodal uncertainty represented by the generated sample distribution.

J.2 Oracle Diagnostics for Predictive and Task-Conditional Error

The image experiment above evaluates posterior-predictive learning in a trained diffusion model. We next complement it with a separate oracle diagnostic in a simple known-density family, isolating task uncertainty from predictive approximation without neural training or reverse sampling.

J.2.1 Task uncertainty and predictive approximation in a known family

The two error targets can be seen in a sample-only family with known densities and prior: $g_\theta(z) = 1 + \frac{1}{4}\cos(2\pi(z - \theta))$ for $z, \theta \in [0, 1]$, with $\Pi = \mathrm{Unif}[0, 1]$. This positive sinusoidal family satisfies the task assumptions (Example I.1). We compute its exact predictive density and compare it with both the realized task density and a predictive approximation on a finite latent grid.

In Figure 6, longer prompts reduce the observed mean task uncertainty. At fixed prompt length, a finer grid instead reduces error against the posterior predictive. These are distinct effects: the former uses more examples from the task, while the latter refines the calculation on the same prompt. The calculations below give the analytical TV formulas, quadrature checks, seeds, and recorded per-episode results.

J.2.2 Analytical TV calculation and numerical protocol

We examine how task uncertainty decreases with the prompt length and how a finite latent grid approximates the posterior predictive distribution in the sample-only family considered here. These are oracle calculations with a known family and prior. They do not train a network, estimate an episode-complexity rate, or evaluate a reverse sampler.

Use the family in Example I.1,

$$g_\theta(z) = 1 + a\cos(2\pi(z - \theta)), \qquad a = 1/4, \quad z, \theta \in [0, 1], \qquad \Pi = \mathrm{Unif}[0, 1].$$

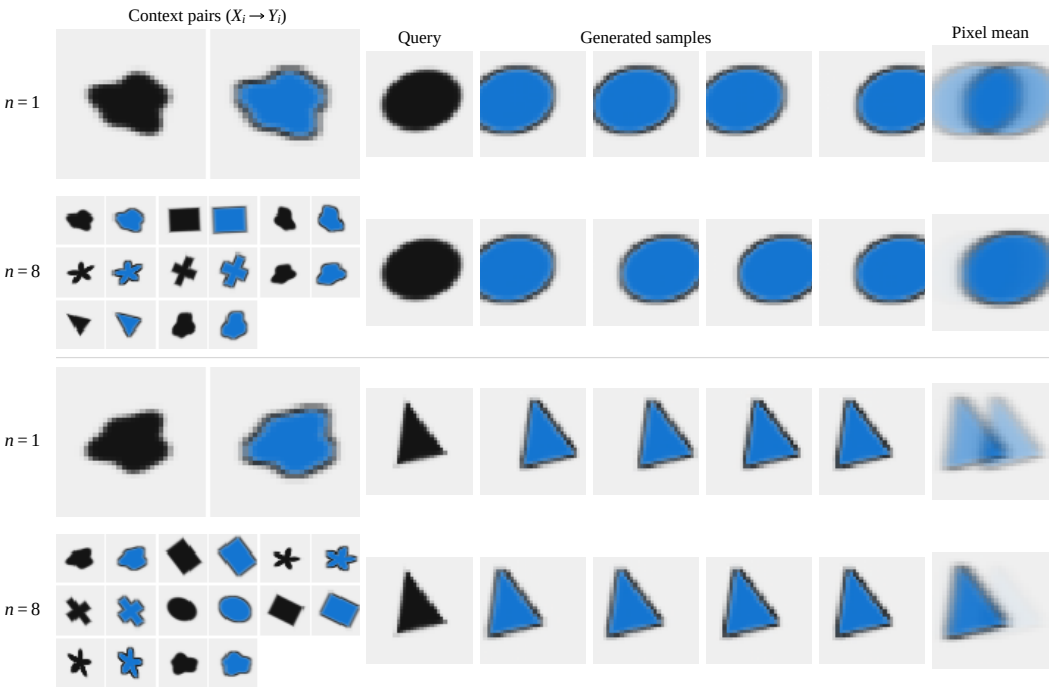

Figure 5: **Qualitative posterior-predictive generation.** The top half shows a task with a circle query, and the bottom half shows a task with a triangle query; each task is displayed at $n = 1$ and $n = 8$. From left to right, each row contains the available context input–output pairs $X_i \to Y_i$, the untransformed query input, four independent generations for that query, and the pixel-wise average of 256 generations. All samples use the raw model weights. For the circle task, the exact/generated right-mode probabilities are $0.550/0.562$ at $n = 1$ and $0.999/0.965$ at $n = 8$. For the other task, they are $0.527/0.410$ at $n = 1$ and $0.009/0.055$ at $n = 8$. The horizontally blurred averages at $n = 1$ reflect bimodality, while the sharp averages at $n = 8$ reflect posterior-predictive contraction.

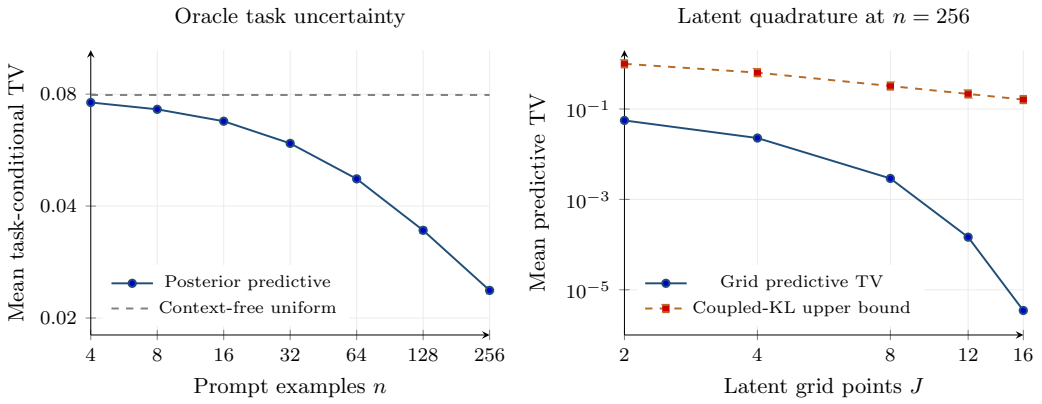

Figure 6: Oracle illustration for the positive sinusoidal family. Left: task uncertainty with nested prompts, compared with the context-free uniform predictor. Right: predictive TV from latent quadrature at $n = 256$, compared with the coupled episode-KL upper bound. Means use 5000 independent tasks; bars show 1.96 standard errors across tasks. TV between densities is calculated analytically. This fixed-family diagnostic does not test the worst-case rate, neural training, or reverse sampling.

The fixed positivity and Lipschitz bounds are verified in that example. Write $v(\theta) = (\cos(2\pi\theta), \sin(2\pi\theta))$ and $m(C_n) = \mathbb{E}[v(\Theta) \mid C_n]$. The posterior has density proportional to

$\prod_{i=1}^{n} g_\theta(z_i)$ relative to Π; its normalizer lies in $[(3/4)^n, (5/4)^n]$, so it is positive for every context. The predictive density equals $1 + a\{m_1 \cos(2\pi z) + m_2 \sin(2\pi z)\}$. The moment vector has norm at most one, so posterior averaging may flatten the density rather than produce the density at a posterior-mean coordinate. For any two such densities with coefficient vectors v, w, their continuous TV distance is exactly

$$\mathsf{TV}(p_v, p_w) = \frac{a}{\pi}\|v - w\|_2. \tag{178}$$

Indeed, the difference is a sinusoid of amplitude $a\|v - w\|_2$, and $\int_0^1 |\cos(2\pi z)|\,\mathrm{d}z = 2/\pi$. Thus all per-episode TV values below are density-to-density integrals computed analytically, with no histogram or empirical-measure TV surrogate.

For each of five independent seeds, draw 1000 tasks and 256 context examples per task by rejection sampling from the displayed density, using the uniform proposal and envelope $1 + a$. Use nested prefixes of lengths $n \in \{4, 8, 16, 32, 64, 128, 256\}$. At each length, evaluate the posterior moments with the midpoint periodic trapezoidal rule using $J_{\mathrm{ref}} = 2^{\lceil \log_2(n+2) \rceil}$ equally weighted latent points. The likelihood is a trigonometric polynomial of degree at most n, and multiplication by either first moment raises that degree to at most $n + 1$. Since $J_{\mathrm{ref}} > n + 1$, this quadrature integrates the numerator and denominator exactly in exact arithmetic. The computation uses log likelihoods with a common normalization to avoid underflow. Doubling the reference grid for the first 32 episodes at every seed and length changed predictive TV by at most 2.3×10^{-16}.

The left panel of Figure 6 evaluates expected task uncertainty by averaging (178) between the predictive moment and the true task moment over 5000 independent tasks. Its mean decreases from 0.07598 at $n = 4$ to 0.02372 at $n = 256$; the corresponding standard errors are 0.00018 and 0.00024. The context-free uniform density has task TV exactly $a/\pi = 0.07958$. The error bars are the mean plus or minus 1.96 times the empirical standard error across tasks. They measure simulation uncertainty, not variability across neural-network training runs. The same tasks are used across prompt lengths. This fixed-family curve does not test the uniform lower bound, whose hard families may change with n.

At $n = 256$, evaluate grid predictive moments using $J \in \{2, 4, 8, 12, 16\}$ midpoint cells and the exact cell weights $1/J$. The mean predictive TV errors are 0.05582, 0.02280, 0.002908, 0.0001460, and 3.482×10^{-6}, respectively. The right panel compares these errors with the Pinsker bound obtained from the coupled episode KL:

$$\mathbb{E}\mathsf{TV}(P_{0|C_n}, Q^J_{0|C_n}) \le \min\left\{1, \sqrt{\frac{n+1}{2}\mathbb{E}_\Theta \mathrm{KL}(g_\Theta \| g_{\pi_J(\Theta)})}\right\}.$$

This follows from the conditional KL chain rule in Proposition D.4, Pinsker's inequality, and Jensen's inequality. The plotted KL uses the actual single-task divergence, instead of the coarser common Hölder bound. Periodicity reduces its prior integral to a displacement uniform on $[-1/(2J), 1/(2J)]$. We compute the displacement and output integrals using 128-node Gauss–Legendre rules and check them against 256 nodes; the coupled episode values agree within 10^{-10}. The resulting upper bounds range from 1 to 0.16177 and are conservative for this family. The prior is uniform, so adding a uniform floor would not change these particular grid weights; this diagnostic does not test prior smoothing for a singular prior.

K Qualitative Diagnostics and Resource Allocation

The decomposition in Section 2.2 separates posterior-predictive generation error from finite-prompt task uncertainty. Table 2 translates this decomposition into qualitative diagnostics for experiments and resource allocation. The entries are not mutually exclusive and should be interpreted as diagnostic hypotheses rather than as additional theorem statements.

Table 2: Qualitative diagnostics and resource allocation for in-context generation.

Observed pattern	**Possible bottleneck**	**Suggested intervention**
The task posterior remains spread over tasks with different output laws.	Finite-prompt task uncertainty	Request additional demonstrations that distinguish the plausible task-specific output laws.
The oracle is concentrated, but the model remains diffuse after checking sampler convergence.	Amortization or learning error	Increase independent pretraining episodes; improve model capacity, optimization, or use of context.
Error relative to the oracle varies substantially with sampler steps or schedule.	Sampling error or score–sampler interaction	Refine the sampler steps and schedule; check convergence with the learned score held fixed.
Individual samples are plausible, but their mode frequencies disagree with the oracle.	Probability calibration error	Evaluate mode probabilities, improve episodic distributional training, and check sensitivity to the sampler.
Learning and sampling errors are small relative to finite-prompt task uncertainty.	Prompt information	Prioritize more informative demonstrations over further reductions in learning or sampling error.
A large token budget contains few independent episodes.	Limited independent training data	Track and increase independent episode count; account for within-episode dependence when allocating pretraining data.