The Design Space of Tri-Modal Masked Diffusion Models

Louis Bethune Apple
Victor Turrisi Apple
Bruno Kacper Mlodozeniec University of Cambridge (Apple internship)
Pau Rodriguez Lopez Apple
Lokesh Boominathan Apple
Nikhil Bhendawade Apple
Amitis Shidani Apple
Joris Pelemans Apple
Theo X. Olausson MIT (Apple internship)
Devon Hjelm Apple
Paul Dixon Apple
João Monteiro Apple
Pierre Ablin Apple
Vishnu Banna Apple
Arno Blaas Apple
Nick Henderson Apple
Kari Noriy Apple
Dan Busbridge Apple
Josh Susskind Apple
Marco Cuturi Apple
Irina Belousova Apple
Luca Zappella Apple
Russ Webb Apple
Jason Ramapuram Google Deepmind (work done at Apple)

Abstract

Discrete diffusion models have emerged as strong alternatives to autoregressive language models, with recent multimodal work either finetuning unimodal diffusion bases or distilling autoregressive backbones for bi-modal generation. Diverging from these approaches, we introduce the first tri-modal Masked Diffusion Models (MDM) *pretrained from scratch* on text, image-text, and audio-text data, where all generation tasks are learned simultaneously within a single model. We systematically study the design space governing stability and efficiency at scale: we conduct the first empirical characterization of the critical batch size B_{crit} under the SDE reparameterization of AdamW, and introduce a drift–horizon interpolation parameter γ that balances gradient noise reduction and optimization horizon when scaling the token budget. We derive multimodal scaling laws and ablate modality mixing ratios, noise schedules, and anti-masking. Lastly, we pretrain a 3B model on 6.4T tokens, achieving competitive results in text generation, text-to-image, and text-to-speech.

1 Introduction

Bidirectional models consistently outperform unidirectional ones when full context is available - in representation learning, classification, and comprehension tasks - from early RNNs [Schuster and Paliwal, 1997] and LSTMs [Graves and Schmidhuber, 2005] to modern transformers [Vaswani

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) (b) (c) (d)

Figure 1: **High-fidelity generation.** The pretrain-only 3B MDM demonstrates strong prompt adherence alongside high-quality visual rendering of *texture*, *lighting*, and *composition*. Samples show: **(a)** natural daylight and depth of field ("egg in a field of crocuses"); **(b)** fine-grained fur texture in B&W ("lion's face"); **(c)** soft, warm lighting with vintage color tones ("preparing bread dough"); and **(d)** complex multi-object arrangement ("noodle soup with toppings").

et al., 2017]. For unimodal discrete generation, on the other hand, the strongest language models [DeepSeek-AI, 2024, Gemini Team, 2023, Singh et al., 2025, Anthropic, 2024] are trained autoregressively. One natural way to recover generation with bidirectional context for discrete sequences is iterative denoising: corrupt a token sequence and train the model to restore it.

Recent work like MMaDA [You et al., 2025] shows MDMs can jointly generate across modalities, something autoregressive models struggle with architecturally. However, these approaches adapt pretrained models (like LLaDA, Nie et al. [2025]) rather than training from scratch, leaving open which design decisions matter at pretraining time. There is no equivalent unified AR baseline, and existing approaches instead rely on distilling existing AR backbones into diffusion models [Hui et al., 2024, Gong et al., 2025, Zheng et al., 2025, Zhang et al., 2025, Bie et al., 2025].

Here, we explore training an MDM from scratch, emphasizing practical implementation choices and empirical analysis. Much of the dominance of AR models stems not from higher intrinsic expressiveness, but from the accumulated engineering effort and convenience of the causal modeling: simple likelihood factorization, efficient per-token signal, fast KV-cache decoding. MDMs lack equivalent recipes, which may explain their remaining quality gap on unimodaltasks [Llama Team, 2024] and latency issues [Jazbec et al., 2025, Wu et al., 2025]. The multimodal setting compounds this: it introduces a large and underexplored design space where choices that may appear secondary in isolation can dominate stability and compute efficiency. We train from scratch out of this necessity, which enables controlled scaling-law analysis and isolates the effect of architectural and optimization choices without inherited biases from pretrained backbones.

In this work, we extend MDM to a tri-modal setting (text, image, audio) via a unified discrete token space with modality-specific boundary and mask tokens. Our contributions are:

1. **Unified Multimodal MDM.** We introduce the first tri-modal MDM pretrained from scratch, supporting text, image, and audio generation via a single transformer backbone and unified vocabulary, without modality-specific heads, adapters or unimodal conditioning [Liu et al., 2023]. This unlocks multimodal generation from multimodal conditioning.

2. **Batch-Size Invariance and Drift–Horizon Tradeoff.** Using the SDE reparametrization of Mlodozeniec et al. [2026], we render training loss invariant to batch size below a critical threshold B_{crit}, eliminating B_{opt} [Bergsma et al., 2025]. We conduct the first empirical study to date of critical batch size B_{crit} for models trained under SDE parameterization. We further introduce a novel interpolation parameter $\gamma \in (0, 1)$ that balances gradient noise reduction and optimization horizon when scaling token budget (Section 4.1,Appendix E).

3. **Multimodal Scaling Laws.** We derive empirical scaling laws $L(N, D)$ for tri-modal MDMs, finding $D^\star(N) \approx 7754 \cdot N^{0.84}$: at the FLOP-optimal frontier, parameter count grows faster than token count, suggesting higher efficiency in data-scarce regimes. When data scarcity forces multi-epoch training, anti-masking (Section 5) reduces gradient variance and improves results at equal FLOP budget.

2 Background and Related Work

2.1 Masked Diffusion Models

Although diffusion models first gained prominence in continuous settings [Song et al., 2021], the original formulation by Sohl-Dickstein et al. [2015] encompasses both continuous and discrete domains. MDMs were proposed by Austin et al. [2021] and generalized by Hoogeboom et al. [2021]: the forward process $q(x_t|x_{t-1})$ progressively replaces tokens with a special MASK token, and a learned reverse process $p_\theta(x_{t-1}|x_t)$ reconstructs the data. Campbell et al. [2022] extended this to continuous time via CTMCs; Dieleman et al. [2022] proposed embedding discrete inputs into continuous space.

Text. LLaDA [Nie et al., 2025] trained an 8B-parameter MDM on 2.3T tokens. Dream [Ye et al., 2025] instead finetuned a pretrained AR model (Qwen2.5-7B) on 580B tokens [Yang et al., 2024a], without accounting for the initial pretraining budget. Both employ the same variational evidence lower bound (ELBO) objective for the continuous-time masked diffusion process:

$$-\mathbb{E}_{\substack{x_0\sim p_{data},\ t\sim U(0,1)\\ x_t\sim q(x_t|x_0)}}\left[w(t)\sum_{\ell=1}^{L}\mathbf{1}_{x_t^\ell=\text{MASK}}\ \log p_\theta(x_0^\ell|x_t)\right]. \tag{1}$$

Here L is the sequence length and x^ℓ the ℓ-th token. The indicator $\mathbf{1}_{x_t^\ell=\text{MASK}}$ restricts the loss to masked positions. The weight $w(t)$ depends on the forward noise; LLaDA uses $w(t) = 1/t$, while Dream uses a more complex schedule. Follow-up works improve efficiency via variance reduction [Zhu et al., 2025a] and mixture-of-experts [Zhu et al., 2025b]. We provide more details on weighting in Appendix B.

Image. Austin et al. [2021] and Shi et al. [2024] apply masked diffusion directly to pixel values, but are limited to low-resolution datasets [Deng et al., 2009, van den Oord et al., 2016]. MaskGIT and VQ-Diffusion [Chang et al., 2022, Gu et al., 2022] use pretrained discrete tokenizers (VQ-GAN [Esser et al., 2021, ai-forever, 2023, Zheng et al., 2022], VQ-VAE [van den Oord et al., 2017]) to enable high-resolution discrete diffusion.

Audio. Yang et al. [2023] apply discrete diffusion to audio. Borsos et al. [2023] combine the SoundStream tokenizer [Zeghidour et al., 2022] with a MaskGIT-style approach [Chang et al., 2022].

2.2 Multimodal Masked Diffusion Models

Early multimodal MDM work [Gu et al., 2022, You et al., 2025] is restricted to generating a single modality. MMaDA [Yang et al., 2025], initialized from LLaDA weights, unifies language and image generation by training on a joint token sequence. Swerdlow et al. [2025] and Hu et al. [2023] train unified text–image diffusion models at smaller scale. We move beyond these bi-modal MDMs by introducing audio as a third modality. Unlike Yang et al. [2025], we *pretrain from scratch*, properly account for the total token budget, and jointly optimize across all modalities. Additionally, we encourage the community to jointly report model and data size, (N, D_{total}), for fair comparison.

3 Method

We consider three modalities $m \in \mathcal{M} := \{\text{text}, \text{audio}, \text{image}\}$, each represented as a sequence of discrete tokens $x_m = (x_m^1, \ldots, x_m^{L_m})$ with $x_m^i \in \mathcal{V}_m$. We construct a shared vocabulary $\mathcal{V} = \mathcal{V}_{\text{text}} \sqcup \mathcal{V}_{\text{audio}} \sqcup \mathcal{V}_{\text{image}}$ (disjoint union), augmented with modality-specific boundary and mask tokens $\{\text{BOS}_m, \text{EOS}_m, \text{MASK}_m\}$, task tokens $\{\text{TASK}_{\text{text}}, \text{TASK}_{\text{image-text}}, \text{TASK}_{\text{audio-text}}\}$, and a padding token PAD. Throughout, superscripts denote token positions and subscripts denote diffusion time or modality.

Training. Training sequences are constructed by wrapping modality tokens with their boundary tokens. For example, an audio-text sample is represented as:

$$s = [\text{TASK}_{\text{audio-text}}, \text{BOS}_{\text{audio}}, x_{\text{audio}}^1, \ldots, x_{\text{audio}}^{L_{\text{audio}}}, \text{EOS}_{\text{audio}}, \text{BOS}_{\text{text}}, x_{\text{text}}^1, \ldots, x_{\text{text}}^{L_{\text{text}}}, \text{EOS}_{\text{text}}],$$

with an image-text sample following the same format. Text-only sequences are packed to fill the maximum sequence length $L^\star$; mixed-modality sequences shorter than $L^\star$ are right-padded with PAD. See Figure 2 for a demonstration.

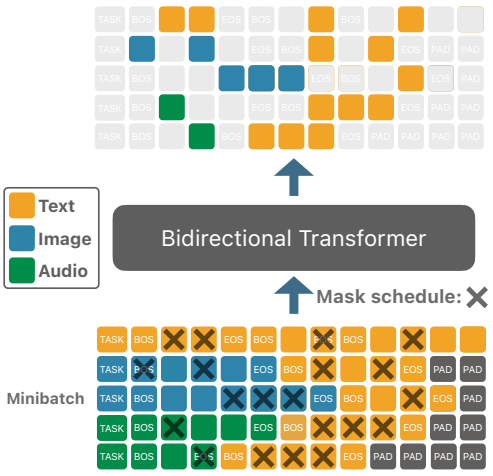


Figure 2: **Tri-Modal masked diffusion model architecture**. Pure text is packed. Image-caption and audio-transcription pairs are padded to maximum length. Padding is ignored by attention and loss computation.

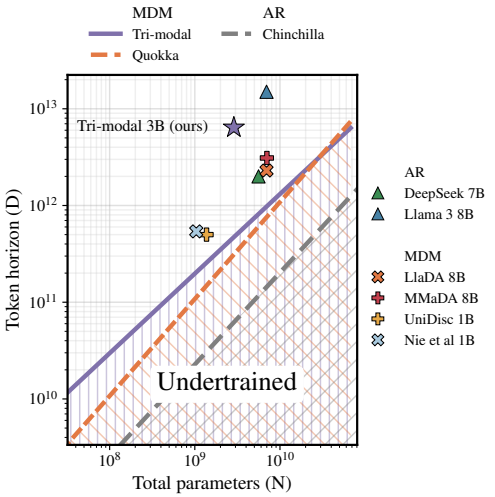


Figure 3: **Token-optimal curve $D^\star(N)$ for different model families**. In tri-modal MDM, token count grows sub-linearly with model size, suggesting diminishing returns of additional data. Identical methodology to report all curves.

Given a sequence $s = (s^1, \ldots, s^{L^\star}) \in \mathcal{V}^{L^\star}$, we define a continuous-time forward masking process indexed by $t \in [0, 1]$. Each position i is independently replaced by its modality-specific mask token $\text{MASK}_{m(i)}$ with probability β_t, where $\beta_0 = 0$ and $\beta_1 = 1$. Task tokens are never masked. The resulting corrupted sequence s_t interpolates between the original data (t=0) and a fully masked sequence (t=1). Once masked, a token stays masked. The full derivation of the forward kernel and its factorization is given in Appendix B.

Denoising model. A denoising model $f_\theta : \mathcal{V}^{L^\star} \to \mathbb{R}^{L^\star \times V}$ predicts logits over the vocabulary at each position, with the per-token cross-entropy loss defined as $\ell_i(\theta, s) = -\log p_\theta(s^i \mid s_t)$, implemented via cut-cross-entropy [Wijmans et al., 2025] for memory efficiency. Let $\mathcal{I}_t := \{\, i \mid s_t^i = \text{MASK}_{m(i)} \,\}$ denote the masked, non-padding positions. The training objective is then defined as the re-weighted cross-entropy loss

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{s\sim\mathcal{D},\\ t\sim\text{U}(\epsilon,1-\epsilon)}} \left[\frac{w(t)}{|\mathcal{I}_t|} \sum_{i\in\mathcal{I}_t} \ell_i(\theta, s) \right] + \lambda \mathcal{L}_z,$$

where $\epsilon, \lambda > 0$ are small constants for numerical stability, and $\mathcal{L}_z$ denotes the z-loss regularizer [de Brébisson and Vincent, 2016]. We follow prior work [Nie et al., 2025, Yang et al., 2025] and choose $w(t) = 1/t$ which yields an unbiased estimator of the ELBO under Bernoulli masking. See Appendix B for the effect of this weighting scheme and general formulation.

Inference. At generation time, we iteratively unmask tokens according to a predefined linear schedule. Note that alternative schedules can improve quality for some modalities (see Section L). Multimodal generation is conditioned on a prompt, e.g., text, and a target modality $m \in \mathcal{M}$. We start the process from a fully masked sequence of the form

$$s_K = [\text{TASK}_{\text{task}}, \text{BOS}_m, \text{MASK}_m, \ldots, \text{MASK}_m, \text{EOS}_m, \text{BOS}_{\text{text}}, x^1_{\text{text}}, \ldots, x^{L_{\text{text}}}_{\text{text}}, \text{EOS}_{\text{text}}].$$

At each step $k \in [K]$, candidate tokens are sampled from the modality-constrained distribution $s^i_k \sim p_\theta(\cdot \mid s_{k-1}) \propto \exp(h^i_{k,v})$, $v \in \mathcal{V}_{m(i)}$, and a subset of masked positions is revealed. The process repeats until no mask token remains. Text-only generation follows the same procedure with a text-only initial sequence (see Appendix A).

3.1 Architecture

For all experiments presented in this paper, we rely on a standard bi-directional transformer architecture with pre-normalization RMSNorm [Zhang and Sennrich, 2019], SwiGLU MLPs [Shazeer,

2020], rotary positional embeddings (RoPE) [Su et al., 2021] and QK-norm [Dehghani et al., 2023, Wortsman et al., 2024, Chameleon Team, 2024, Gemma Team, 2025]. Our Tri-modal 3B model is pretrained from scratch for 1M steps with batch size of 3072 and sequence length of 3256. We tokenize modalities with specialized encoders: SBER-MoVQGAN [ai-forever, 2023] for images, Higgs Audio v2 [Boson AI, 2025] for audio, and Tiktoken [OpenAI, 2023] for text. To manage the large vocabulary efficiently, we employ cut-cross-entropy [Wijmans et al., 2025] and apply a z-loss regularizer [de Brébisson and Vincent, 2016] to stabilize logits amplitudes. See Table 4 for full hyperparameter details.

4 Training at Scale with SDE Reparametrization

Grid search at large scale is infeasible so we rely on *hyperparameter transfer rules* from small-scale sweeps. We use CompleteP [Dey et al., 2025], a width-and-depth transfer rule that extends μP [Yang et al., 2021, 2024b, Blake et al., 2025], with the per-module extensions of Mlodozeniec et al. [2026]. In the main paper we do not apply per-module hyperparameters, but these results are present in Appendix E.

4.1 Eliminating B_{opt} with SDE Parametrization

AdamW operates like the discretization of an SDE [Malladi et al., 2022, Mlodozeniec et al., 2026] whose drift and noise depend on its hyperparameters. Since these are redundant with batch size (e.g., larger batches reduce gradient noise, as does lower momentum), we reparametrize them to decouple batch size from training dynamics [Mlodozeniec et al., 2026]. This guarantees similar performance across all compute budgets, *as long as the batch size is not larger than* B_{crit}. Unlike AR models, the MDM cross-entropy loss is not a likelihood but an ELBO [Turok et al., 2026]; its exponential is analogous to perplexity (lower is better), so we report $\exp(\text{ELBO})$ in the following figures. Results are in Figure 4: SDE parametrization guarantees constant loss across batch sizes, unlike the U-curve of non-SDE parametrization [Bergsma et al., 2025].

4.2 Scaling Optimizer Hyperparameters with Token Horizon

When scaling the token horizon D, extra tokens can reduce gradient noise (the SDE *drift*) or extend the optimization trajectory (the SDE *horizon*). Mlodozeniec et al. [2026] introduce a scaling factor κ that rescales AdamW's hyperparameters, but fix γ=1 (pure isohorizon). We generalize to $\gamma \in [0, 1]$, interpolating between isonoise (γ=0) and isohorizon (γ=1) via $\kappa = (D^{\text{base}}/D)^{\gamma}(B/B^{\text{base}})$:

$$\text{lr} = \text{lr}^{\text{base}}\sqrt{\kappa}, \quad \beta_1 = (\beta_1^{\text{base}})^{\kappa}, \quad \beta_2 = (\beta_2^{\text{base}})^{\kappa}, \quad \epsilon = \frac{\epsilon^{\text{base}}}{\sqrt{\kappa}}, \quad \text{wd} = \text{wd}^{\text{base}}\sqrt{\kappa}. \tag{2}$$

We conduct the base grid-search ($\approx$ 3k runs) with N=320M, D^{base}=13B tokens and B^{base}=256.

All experiments presented use a cosine learning rate schedule with 1k steps of linear warmup (never exceeding 25% of the total). Following Busbridge et al. [2025], we set width proportional to depth, fixing $\rho = d_{\text{head}} = d_{\text{emb}}/n_{\text{layers}} = 128$, i.e. using a single degree of freedom to scale model size N.

4.3 Scaling Rules for the Critical Thresholds

In this section, we conduct the first empirical study of the critical batch size B_{crit} and critical iteration count S_{crit} for models trained under SDE parameterization.

Scaling in conventional regime. Without SDE parametrization, batch size affects both gradient variance and iteration count, so beyond B_{crit} the marginal utility of each additional token decreases. Previous work [Kaplan et al., 2020, DeepSeek-AI, 2024, Bergsma et al., 2025] suggests a B_{opt} minimizing loss at fixed D. Under SDE parametrization, that notion disappears (Figure 4): all batch sizes under B_{crit} yield identical results. See Section D.2 for details.

Critical integration steps for SDE. S_{crit} is the minimum number of optimization steps for FLOP-efficient training. When $S < S_{\text{crit}}$ (i.e. the batch size too large), the SDE approximation breaks and performance at constant D degrades. When $S \geq S_{\text{crit}}$, all runs are ***FLOP-efficient***; when $S < S_{\text{crit}}$, runs are *wall-clock faster* but waste FLOPs. We measure S_{crit} empirically as a function of D and N and define the critical batch size as $B_{\text{crit}} = D/(L\, S_{\text{crit}})$. See subsection F.2 for the full definition.

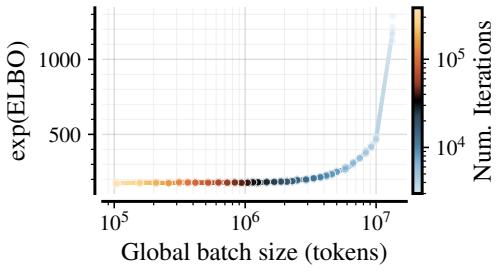


Figure 4: **Below the critical batch size B_{crit} the SDE parametrization guarantees constant loss**. In that regime, larger batch sizes allow fewer iterations. Above it, SDE discretization breaks and training ceases to be FLOP-efficient.

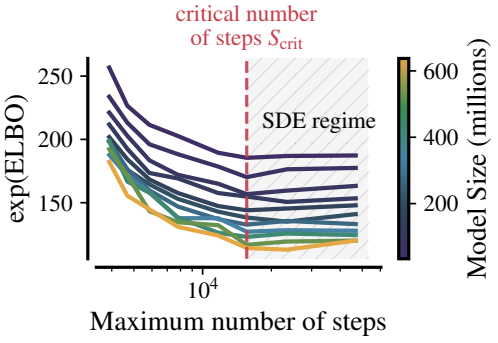


Figure 5: **Critical iteration count S_{crit} is constant w.r.t. model size** under the SDE regime. This is compatible with the findings of Bergsma et al. [2025], but their study was done *outside* the SDE regime.

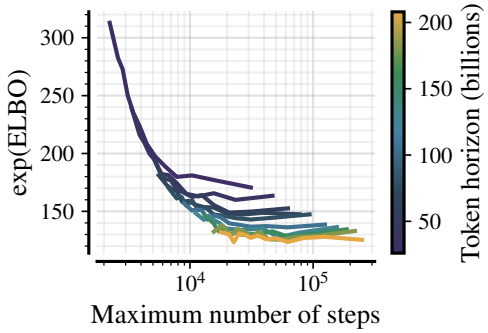


Figure 6: **Critical iteration count S_{crit} increases with token horizon D**. The increase is sub-linear, meaning that the critical batch size B_{crit} also increases with horizon D.

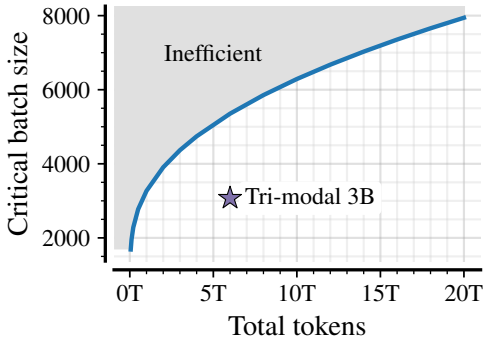


Figure 7: **Critical batch-size B_{crit} as a function of the token horizon**. There is an intrinsic tension between wall-clock time and FLOP-efficiency.

Critical batch size as function of model size. Figure 5 shows that S_{crit} is **independent** of model size at constant D=13B, consistent with Bergsma et al. [2025] in the non-SDE case.

Critical batch-size as a function of the token horizon. Figure 6 shows that S_{crit} grows sub-linearly with D, so B_{crit} also grows sub-linearly. We extrapolate its value via two power laws (Figure 7).

Takeaways

Under the SDE parameterization, the global batch size can be chosen to saturate GPU capacity without compromising final loss, *as long as it is less than B_{crit}*. B_{crit} grows sub-linearly with D, so the GPU count cannot scale proportionally with token horizon forever. Long horizon training cannot be both FLOP-efficient and fast in wall-clock time.

4.4 Balancing Gradient Noise and Optimization Horizon

The SDE reparametrization maps the physical batch size B to a *virtual batch size* $\tilde{B}$ and *virtual iteration count* $\tilde{S}$ with $D = \tilde{B}\tilde{S}L$. The physical B is chosen from compute constraints; the virtual $\tilde{B}$ controls the noise–horizon tradeoff. We parameterize their joint scaling as:

$$\tilde{S} = G(D/L)^{1-\gamma} \quad \text{and} \quad \tilde{B} = G^{-1}(D/L)^{\gamma}, \quad \text{with} \quad \gamma = \alpha/(\alpha+\beta) \text{ and } G = \left(\frac{\alpha C_1}{\beta C_2}\right)^{1/(\alpha+\beta)}.$$

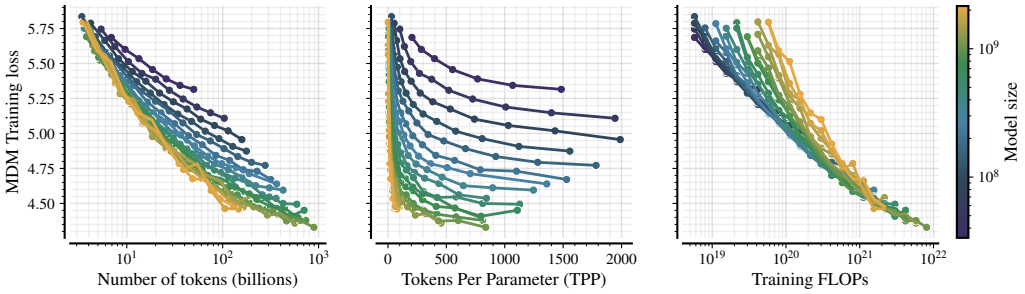


Figure 8: **Training curves for tri-modal MDM.** We select 24 log-distributed isoFLOPS between 5e18 and 1e22 to cover the (N, D) grid. The performance is dominated by the total compute budget C, following formula in appendix **H.1** of Busbridge et al. [2025]. The loss is computed on an independent validation set of 50k samples with identical mixture weights.

Here γ=0 corresponds to the standard LLM setup (more tokens = more iterations) [Kaplan et al., 2020], while γ=1 corresponds to the work of Mlodozeniec et al. [2026]. By sweeping $\gamma \in [0, 1]$ (Figure 9), we find that neither extreme is optimal. Fitting $C_0 + C_1\tilde{S}^{-\alpha} + C_2\tilde{B}^{-\beta}$ yields α=0.18, β=0.23 and γ^*=0.44 $\approx 1/2$. See also Figure 18.

Takeaways

When doubling the token budget, roughly half of the extra tokens should lengthen the optimization trajectory (more virtual iterations $\tilde{S}$) and half should reduce gradient noise (larger virtual batch size $\tilde{B}$). Geometrically, extra tokens are split between traveling further through the loss landscape and descending deeper into its valleys.

Link with learning rate schedule. γ plays a role akin to a learning rate scheduler: for $\gamma < 1$, the effective LR rises with batch size; for $\gamma > 0$, longer horizons reduce it. The optimal γ may shift under different schedules, e.g., warmup-stable-decay [Hägele et al., 2024, Schaipp et al., 2025].

4.5 Scaling Laws for Tri-modal MDM

We train *262 tri-modal MDM models* under CompleteP + SDE, with tokens-per-parameter (TPP) ratios between 1 and 2000 along 24 isoFLOPs from $5{\times}10^{18}$ to $1{\times}10^{22}$ (Figure 8). Model size N excludes embedding parameters [Pearce and Song, 2024] (Figure 17). FLOPs follow Busbridge et al. [2025]. We fit scaling laws following Shukor et al. [2025a] with bootstrap cross-validation, using the form of Kaplan et al. [2020] since the additive form of Hoffmann et al. [2022] was insufficient:

$$L = C_0 + \left(\frac{C_1}{N^{a/b}} + \frac{C_2}{D}\right)^b. \tag{3}$$

We report scaling coefficients of $a \approx 0.14$ and $b \approx 0.17$, with R^2=99.3% and MRE of 0.5% (Figure 10). We report isoloss contours in Figure 11 and isoFLOP curves in Figure 12. The compute-optimal token count is:

$$D^\star(N) = 7754 \cdot N^\alpha \quad \text{with} \quad \alpha = a/b = 0.84. \tag{4}$$

We compare $D^\star(N)$ against other model families in Figure 3, using approach 3 of Chinchilla with corrected coefficients from Besiroglu et al. [2024]. A 3B tri-modal MDM requires ∼480B tokens, in sharp contrast with ∼60B from Chinchilla [Hoffmann et al., 2022] for AR models. The crossing with Quokka [Ni et al., 2025] occurs around 20B parameters: below, tri-modal MDMs should be *smaller* and *trained longer*; above, the trend reverses.

Analysis. As N grows, the TPP ratio $D^\star/N \propto N^{-0.16}$ decreases: MDMs become asymptotically more data-efficient per parameter, in contrast to the AR rule of thumb $D \propto 20N$ [Hoffmann et al., 2022]. The optimal compute allocation is:

$$N^\star(C) \propto C^{0.55} \quad \text{and} \quad D^\star(C) \propto C^{0.45}. \tag{5}$$

These exponents are *not* sufficient to conclude superiority across model families [Lin et al., 2024]: the ELBO is informative within a diffusion family but misleading across families with different

forward processes [Sahoo et al., 2026], and b is sensitive to data composition [Muennighoff et al., 2023].

Takeaways

The sub-linear growth $D^*(N) \propto N^{0.84}$ confirms increasing data-efficiency within this model family as scale grows. At current practical scales, a 3B model still requires $\sim$480B tokens at the compute-optimal frontier.

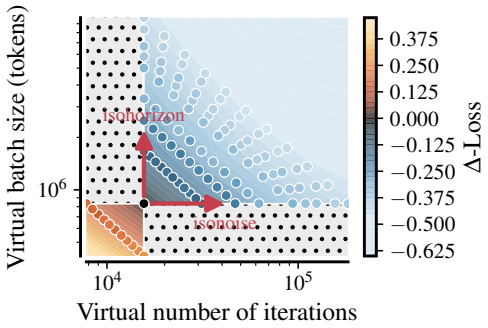


Figure 9: **Drift–horizon tradeoff** γ. The optimal choice lies between reducing the drift (bigger virtual batch size) and increasing the SDE horizon (bigger virtual iteration count).

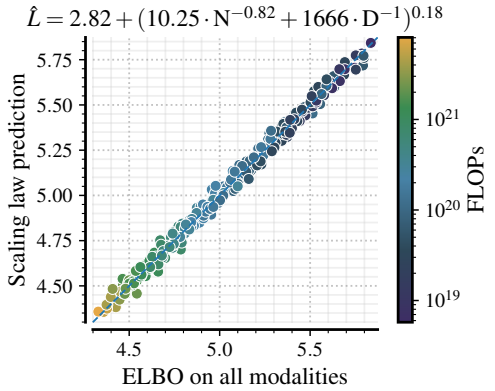


Figure 10: **Scaling law fit for tri-modal MDMs** using Equation 3, with R^2=99.3%, MRE of 0.5%. Model size N and token horizon D are expressed in billions.

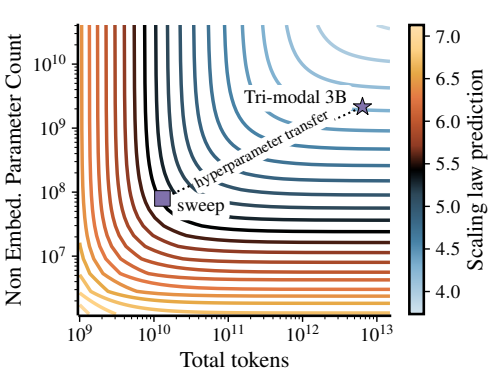


Figure 11: **Isoloss contours for tri-modal MDM**. Dashed line indicates direction in which the 0-shot hyperparameter transfer is done using CompleteP + SDE parameterization.

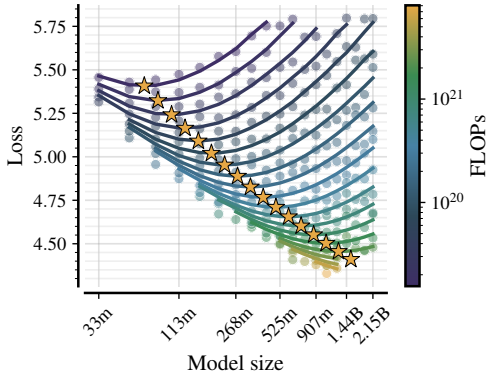


Figure 12: **IsoFLOPs for tri-modal MDMs.** Solid lines indicate scaling law predictions. Points represent measurements and $\star$ indicates the lowest loss achievable at each isoFLOP.

5 Experiments

Unless specified otherwise, we conduct all experiments using a 33% (pure text), 33% (image-text), 33% (audio-text) mixing ratio. The token horizon D is smaller than the total dataset size, ensuring single-epoch training; some high-quality sub-mixtures are repeated up to 4 times, within estimated repetition tolerance [Prabhudesai et al., 2025]. Our text corpus aggregates Nemotron-CC [Su et al., 2025], DCLM [Li et al., 2024], subsets of The Pile [Gao et al., 2020], and synthetic reasoning traces. Audio data consists of 2M hours transcribed by Whisper [Radford et al., 2023] and quality-filtered [Tjandra et al., 2025]. Image data aggregates CC3M [Sharma et al., 2018], CC12M [Changpinyo et al., 2021] and datasets from Li et al. [2025].

Table 1: **3B tri-modal MDM results.** *Left:* generation metrics on a held-out test set and separate datasets. FID uses DINOv2-L [Oquab et al., 2024] and Inception-v3 [Szegedy et al., 2016]; FAD [Kilgour et al., 2019, Gui et al., 2024]. *Right:* LM Harness [Gao et al., 2023]. *Bottom:* GenEval [Ghosh et al., 2023] results. Full results in Figure 21 and hyperparameters in Table 4.

Metric	Test set	Separate dataset
FID-Inception ↓	10.41	10.06 (CC12M)
FID-DINOv2 ↓	112.12	107.61 (CC12M)
FAD ↓	0.218	0.368 (Libri-PC)
WER ↓	0.124	0.164 (Libri-PC)

Benchmark	Acc.	Benchmark	Acc.
MMLU	41.57	HellaSwag	65.88
ARC-Easy	72.52	ARC-Chall.	43.09
WinoGrande	64.69	PIQA	71.55
BBH	24.97	TruthfulQA	40.76
OpenBookQA	37.40	LogiQA2	30.66

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
GenEval	93.12	63.38	33.44	64.89	11.50	27.00	48.89

Unified 3B Tri-modal MDM. We evaluate the 3B model on each modality. Image generation is assessed with FID and GenEval [Ghosh et al., 2023] for compositionality. Audio is evaluated using FAD, WER, and Audiobox Aesthetics (PQ, CU, CE, PC). Key results are in Table 1; the full LM Harness breakdown and radar visualization are in the appendix (Figure 21).

Inference-time ablations. We explore the effect of inference hyperparameters for image and audio generation: FID improves with higher CFG scales, whereas audio generation saturates much earlier; optimal top-p values are lower for audio generation than image generation. For all modalities, optimal temperature is around $T \approx 1$ which is expected for a calibrated pretrained model. Detailed sweep curves are reported in Figure 23 and Figure 24.

Modality mixing. We vary the global mixing ratios $\{w_{\text{text}}, w_{\text{image}}, w_{\text{audio}}\}$ across 15 runs (320M model, 13B tokens, minimum 20% per modality). Results are in Figure 13: the loss on each modality decreases as its own weight increases, but is independent of the relative ratio of the other two. We observe no cross-modal synergies at this scale; each modality competes for capacity and tokens. Token cost of multimodal models has also been documented in the past [Shukor et al., 2025b,a].

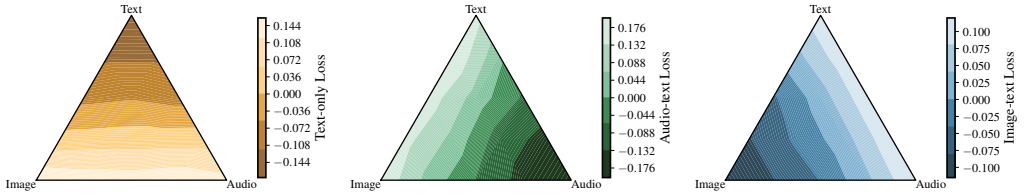


Figure 13: **Loss contours for tri-modal mixture coefficients**, taking $[1/3, 1/3, 1/3]$ as the reference point for the 0-level contour. We do not observe synergies between modalities at this scale.

Anti-masking. The stochastic masking in MDM training induces high gradient variance [von Rütte et al., 2025]. Anti-masking [Jia et al., 2025, Gong et al., 2025] mitigates this by pairing each sample with its complementary mask: a batch $[1, 1^*, 2, 2^*]$ instead of $[1, 2, 3, 4]$, where $*$ denotes the negated mask. This doubles the cost per unique sample, so we compare under fixed token budget D: anti-masking trains for one epoch while the baseline trains for two epochs on the same data. Under this compute-matched comparison, anti-masking improves FID and FAD on all datasets (1.3B multimodal), and accuracy on 9/10 LM Harness benchmarks (7B text-only) (Table 6, Table 7).

6 Conclusion

Limitations. Tri-modal MDMs have few direct competitors: most multimodal diffusion work is bi-modal (text–image) and often initialized from pretrained AR weights, making controlled comparisons difficult. Our scaling laws optimize the ELBO, not downstream metrics (FID, WER, accuracy); loss improvements do not translate monotonically into generation quality gains, so the $D^*(N)$ curve is prescriptive for loss only. The scaling coefficients (a, b) are fitted on a single uniform mixture;

their sensitivity to data composition is unexplored. Generation quality is bounded by the reconstruction fidelity of the tokenizers: their artifacts propagate regardless of model capacity. Finally, our largest model is 3B parameters; whether the sub-linear $D^*(N)$ trend and the absence of cross-modal synergies persist beyond 10B remains unknown.

Future work. In this work, we demonstrate feasibility of large-scale multi-modal training with masked discrete diffusion models. Supervised finetuning and preference optimization would likely improve generation quality but are unexplored in this work. Finally, closing the inference latency gap with AR models is a priority; for example relying on recent advances in parallel decoding [Jazbec et al., 2025, Wu et al., 2025].

References

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093. URL https://doi.org/10.1109/78.650093.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005. doi: 10.1016/J.NEUNET.2005.06.042. URL https://doi.org/10.1016/j.neunet.2005.06.042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

Gemini Team. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. doi: 10.48550/ARXIV.2312.11805. URL https://doi.org/10.48550/arXiv.2312.11805.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Anthropic. Introducing the next generation of claude, 2024. URL https://www.anthropic.com/news/claude-3-family.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-v: Large language diffusion models with visual instruction tuning. *CoRR*, abs/2505.16933, 2025. doi: 10.48550/ARXIV.2505.16933. URL https://doi.org/10.48550/arXiv.2505.16933.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *CoRR*, abs/2502.09992, 2025. doi: 10.48550/ARXIV.2502.09992. URL https://doi.org/10.48550/arXiv.2502.09992.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, An Yang, Rui Men, Fei Huang, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report. *CoRR*, abs/2409.12186, 2024. doi: 10.48550/ARXIV.2409.12186. URL https://doi.org/10.48550/arXiv.2409.12186.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. Diffucoder: Understanding and improving masked diffusion models for code generation. *CoRR*, abs/2506.20639, 2025. doi: 10.48550/ARXIV.2506.20639. URL https://doi.org/10.48550/arXiv.2506.20639.

Huangjie Zheng, Shansan Gong, Ruixiang Zhang, Tianrong Chen, Jiatao Gu, Mingyuan Zhou, Navdeep Jaitly, and Yizhe Zhang. Continuously augmented discrete diffusion model for categorical generative modeling. *CoRR*, abs/2510.01329, 2025. doi: 10.48550/ARXIV.2510.01329. URL https://doi.org/10.48550/arXiv.2510.01329.

Ruixiang Zhang, Shuangfei Zhai, Yizhe Zhang, James Thornton, Zijing Ou, Joshua M. Susskind, and Navdeep Jaitly. Target concrete score matching: A holistic framework for discrete diffusion. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ZMrdvSm7xi.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. LLaDA2.0: Scaling up diffusion language models to 100b. *CoRR*, abs/2512.15745, 2025. doi: 10.48550/ARXIV.2512.15745. URL https://doi.org/10.48550/arXiv.2512.15745.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Metod Jazbec, Theo X Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, Joao Monterio, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models. *arXiv preprint arXiv:2512.09106*, 2025.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. *CoRR*, abs/2505.22618, 2025. doi: 10.48550/ARXIV.2505.22618. URL https://doi.org/10.48550/arXiv.2505.22618.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Bruno Kacper Mlodozeniec, Pierre Ablin, Louis Bethune, Dan Busbridge, Michal Klein, Jason Ramapuram, and marco cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=elB9k4nTL1.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in LLM pre-training. *CoRR*, abs/2505.13738, 2025. doi: 10.48550/ARXIV.2505.13738. URL https://doi.org/10.48550/arXiv.2505.13738.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In Marc'Aurelio Ranzato,

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12454–12465, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract-Conference.html.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *CoRR*, abs/2211.15089, 2022. doi: 10.48550/ARXIV.2211.15089. URL https://doi.org/10.48550/arXiv.2211.15089.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *CoRR*, abs/2508.15487, 2025. doi: 10.48550/ARXIV.2508.15487. URL https://doi.org/10.48550/arXiv.2508.15487.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. *CoRR*, abs/2505.19223, 2025a. doi: 10.48550/ARXIV.2505.19223. URL https://doi.org/10.48550/arXiv.2505.19223.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, Hongrui Guo, Jiaqi Hu, Wentao Ye, Tieyuan Chen, Chenchen Li, Chengfu Tang, Haibo Feng, Jun Hu, Jun Zhou, Xiaolu Zhang, Zhenzhong Lan, Junbo Zhao, Da Zheng, Chongxuan Li, Jianguo Li, and Ji-Rong Wen. LLaDA-moe: A sparse moe diffusion language model. *CoRR*, abs/2509.24389, 2025b. doi: 10.48550/ARXIV.2509.24389. URL https://doi.org/10.48550/arXiv.2509.24389.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1747–1756. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/oord16.html.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11305–11315. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01103. URL https://doi.org/10.1109/CVPR52688.2022.01103.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10686–10696. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01043. URL https://doi.org/10.1109/CVPR52688.2022.01043.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

ai-forever. SBER-MoVQGAN. https://github.com/ai-forever/MoVQGAN, 2023.

Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation, 2022. URL https://arxiv.org/abs/2209.09002.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1720–1733, 2023. doi: 10.1109/TASLP.2023.3268730. URL https://doi.org/10.1109/TASLP.2023.3268730.

Zalán Borsos, Matthew Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *CoRR*, abs/2305.09636, 2023. doi: 10.48550/ARXIV.2305.09636. URL https://doi.org/10.48550/arXiv.2305.09636.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. *CoRR*, abs/2505.15809, 2025. doi: 10.48550/ARXIV.2505.15809. URL https://doi.org/10.48550/arXiv.2505.15809.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *CoRR*, abs/2503.20853, 2025. doi: 10.48550/ARXIV.2503.20853. URL https://doi.org/10.48550/arXiv.2503.20853.

Minghui Hu, Chuanxia Zheng, Zuopeng Yang, Tat-Jen Cham, Heliang Zheng, Chaoyue Wang, Dacheng Tao, and Ponnuthurai N. Suganthan. Unified discrete diffusion for simultaneous vision-language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=8JqINxA-2a.

Erik Wijmans, Brody Huval, Alexander Hertzberg, Vladlen Koltun, and Philipp Krähenbühl. Cut your losses in large-vocabulary language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=E4Fk3YuG56.

Alexandre de Brébisson and Pascal Vincent. The z-loss: a shift and scale invariant classification loss belonging to the spherical family. *CoRR*, abs/1604.08859, 2016. URL http://arxiv.org/abs/1604.08859.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Noam Shazeer. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021. URL https://arxiv.org/abs/2104.09864.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7480–7512. PMLR, 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie E. Everett, Alexander A. Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *CoRR*, abs/2405.09818, 2024. doi: 10.48550/ARXIV.2405.09818. URL https://doi.org/10.48550/arXiv.2405.09818.

Gemma Team. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786.

Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio, 2025. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

OpenAI. tiktoken: a fast byte-pair encoding tokenizer for openai models. https://github.com/openai/tiktoken, 2023.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: CompleteP enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*, 2025.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite depth neural networks. In *The Twelfth International Conference on Learning Representations*, 2024b.

Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Yuri Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr. u-mu p: The unit-scaled maximal update parametrization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/32ac710102f0620d0f28d5d05a44fe08-Abstract-Conference.html.

Gilad Turok, Chris De Sa, and Volodymyr Kuleshov. Duel: Exact likelihood for masked diffusion via deterministic unmasking. *arXiv preprint arXiv:2603.01367*, 2026.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=1nEBAkpfb9.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8b970e15a89bf5d12542810df8eae8fc-Abstract-Conference.html.

Fabian Schaipp, Alexander Hägele, Adrien B. Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/schaipp25a.html.

Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=NLoaLyuUUF.

Mustafa Shukor, Louis Béthune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *CoRR*, abs/2507.09404, 2025a. doi: 10.48550/ARXIV.2507.09404. URL https://doi.org/10.48550/arXiv.2507.09404.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *CoRR*, abs/2404.10102, 2024. doi: 10.48550/ARXIV.2404.10102. URL https://doi.org/10.48550/arXiv.2404.10102.

Jinjie Ni, Qian Liu, Chao Du, Longxu Dou, Hang Yan, Zili Wang, Tianyu Pang, and Michael Qizhe Shieh. Training optimal large diffusion language models. *CoRR*, abs/2510.03280, 2025. doi: 10.48550/ARXIV.2510.03280. URL https://doi.org/10.48550/arXiv.2510.03280.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukic. Scaling beyond masked diffusion language models. *arXiv preprint arXiv:2602.15014*, 2026.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. Diffusion beats autoregressive in data-constrained settings. *CoRR*, abs/2507.15857, 2025. doi: 10.48550/ARXIV.2507.15857. URL https://doi.org/10.48550/arXiv.2507.15857.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, 2025.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *CoRR*, abs/2502.05139, 2025. doi: 10.48550/ARXIV.2502.05139. URL https://doi.org/10.48550/arXiv.2502.05139.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1238. URL https://aclanthology.org/P18-1238/.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00356. URL https://openaccess.thecvf.com/content/CVPR2021/html/Changpinyo_Conceptual_12M_Pushing_Web-Scale_Image-Text_Pre-Training_To_Recognize_Long-Tail_Visual_CVPR_2021_paper.html.

Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, Jialing Tong, Haoxuan You, Xianzhi Du, Zhe Gan, Hyunjik Kim, Chao Jia, Zhenbang Wang, Yinfei Yang, Mingfei Gao, Zi-Yi Dou, Wenze Hu, Chang Gao, Dongxu Li, Philipp Dufter, Zirui Wang, Guoli Yin, Zhengdong Zhang, Chen Chen, Yang Zhao, Ruoming Pang, and Zhifeng Chen. MANZANO: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *CoRR*, abs/2509.16197, 2025. doi: 10.48550/ARXIV.2509.16197. URL https://doi.org/10.48550/arXiv.2509.16197.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Gernot Kubin and Zdravko Kacic, editors, *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2350–2354. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-2219. URL https://doi.org/10.21437/Interspeech.2019-2219.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 1331–1335. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446663. URL https://doi.org/10.1109/ICASSP48485.2024.10446663.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–23, 2025b.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=rvZv7sDPV9.

Mengni Jia, Mengyu Zhou, Yihao Liu, Xiaoxi Jiang, and Guanjun Jiang. Bringing stability to diffusion: Decomposing and reducing variance of training masked diffusion models. *CoRR*, abs/2511.18159, 2025. doi: 10.48550/ARXIV.2511.18159. URL https://doi.org/10.48550/arXiv.2511.18159.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, pages 749–752. IEEE, 2001. doi: 10.1109/ICASSP.2001.941023. URL https://doi.org/10.1109/ICASSP.2001.941023.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Samuel Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Yen-Chen Lin, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zólkowski. Cosmos world foundation model platform for physical AI. *CoRR*, abs/2501.03575, 2025. doi: 10.48550/ARXIV.2501.03575. URL https://doi.org/10.48550/arXiv.2501.03575.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization, 2025. URL https://arxiv.org/abs/2412.02692.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation, 2025. URL https://arxiv.org/abs/2409.04410.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Mayee F Chen, Tyler Murray, David Heineman, Matt Jordan, Hannaneh Hajishirzi, Christopher Ré, Luca Soldaini, and Kyle Lo. Olmix: A framework for data mixing throughout lm development. *arXiv preprint arXiv:2602.12237*, 2026.

A Method Details

Vocabulary size. The unified vocabulary has size $|V| = \sum_{m\in\mathcal{M}}|V_m| + |V_{\text{spec}}| + |V_{\text{task}}| + 1$, where V_{spec} contains modality-specific boundary and mask tokens, V_{task} contains task tokens, and the additional token is PAD. Task tokens signal the generation task to the model, which is useful for adding new tasks through mid-training or supervised finetuning.

Per-token loss. The per-token cross-entropy loss at position i is:

$$\ell_i(\theta, s) = -\log \frac{\exp\big(h^i_{s^i}\big)}{\sum_{v\in\mathcal{V}}\exp\big(h^i_v\big)} := -\log p_\theta(s^i \mid s_t),$$

where $h = f_\theta(s_t)$ and h^i_v is the logit for token v at position i. We implement this using cut-cross-entropy (CCE) [Wijmans et al., 2025], which avoids materializing the full $L^\star \times V$ probability matrix, reducing memory usage proportionally to the vocabulary size.

Inference sequences. For multimodal generation conditioned on text, the initial fully masked sequence is:

$$s_K = [\text{TASK}_{\text{task}}, \text{BOS}_m, \text{MASK}_m, \ldots, \text{MASK}_m, \text{EOS}_m, \text{BOS}_{\text{text}}, x^1_{\text{text}}, \ldots, x^{L_{\text{text}}}_{\text{text}}, \text{EOS}_{\text{text}}].$$

For text-only generation:

$$s_K = [\text{TASK}_{\text{text}}, \text{BOS}_{\text{text}}, x^1_{\text{text}}, \ldots, x^{L_{\text{text}}}_{\text{text}}, \text{MASK}_{\text{text}}, \ldots, \text{MASK}_{\text{text}}, \text{EOS}_{\text{text}}].$$

At each step k, the denoising model produces logits $h_k = f_\theta(s_{k-1})$. Each masked position i samples a candidate from the modality-constrained distribution:

$$s^i_k \sim p_\theta(\cdot \mid s_{k-1}) \quad \propto \quad \exp\big(h^i_{k,v}\big), \quad v \in \mathcal{V}_{m(i)}.$$

B General formulation of weighting and the masking process

In this section, we provide the full derivation of the MDM forward process summarized in Section 3, following the notation of Shi et al. [2024].

In the general case of MDM, we progressively corrupt the original data $s_0 \in \mathcal{V}^{L^\star}$ into a masked version s_t over T discrete time steps. The forward process defines a Markov chain that transforms an original data s_0 into a corrupted version s_t at time step $t \in [T]$. This process is governed by a sequence of transition matrices for each position i denoted by $Q^i_t \in \mathbb{R}^{V\times V}$. At each step t, a token s^i_{t-1} is transformed into s^i_t according to the probability $q(s^i_t \mid s^i_{t-1})$. We define the masking probability at time t as $\beta_t \in [0, 1]$. Consequently, the probability of a token *not* being masked is $\alpha_t = 1 - \beta_t$.

The single-step transition matrix Q^i_t is typically defined as:

$$Q^i_t(v \mid r) = q(s^i_t = v \mid s^i_{t-1} = r) = \begin{cases} 1-\beta_t & \text{if } r \neq \text{MASK}_{m(i)} \text{ and } v = r \\ \beta_t & \text{if } r \neq \text{MASK}_{m(i)} \text{ and } v = \text{MASK}_{m(i)} \\ 1 & \text{if } r = v = \text{MASK}_{m(i)} \\ 0 & \text{otherwise} \end{cases}$$

This formulation implies that a token either remains unchanged or is replaced by the $\text{MASK}_{m(i)}$ token. More general formulations might allow transitions to any other token with a small probability.

The *cumulative* transition probability from s_0 to s_t is crucial for training and is given by the product of individual transition matrices: $\bar{Q}^i_t(v \mid r) = q(s^i_t = v \mid s^i_0 = r)$. This can be simplified by defining $\bar{\alpha}_t = \prod_{\tau=1}^t \alpha_\tau = \prod_{\tau=1}^t (1-\beta_\tau)$. The probability of a token s^i_0 remaining unchanged until time t is $\bar{\alpha}_t$. Conversely, the probability of it having been masked at least once and therefore being a $\text{MASK}_{m(i)}$ token at time t is $1 - \bar{\alpha}_t$. Therefore, the marginal distribution of s_t given s_0 for a single token i is:

$$q(s^i_t = v \mid s^i_0 = r) = \begin{cases} \bar{\alpha}_t & \text{if } v = r \\ 1-\bar{\alpha}_t & \text{if } v = \text{MASK}_{m(i)} \\ 0 & \text{otherwise} \end{cases}, \qquad q(s_t \mid s_0) = \prod_{i=1}^{L^\star} q(s^i_t \mid s^i_0).$$

This distribution $q(s_t \mid s_0)$ allows for direct sampling of s_t from s_0 at any time step t.

B.1 Connection of Loss Weighting and Cumulative Corruption

The weighting function $w(t)$ plays a critical role in balancing the contribution of different time steps to the total loss, with the choice of $w(t)$ being intimately connected to the masking schedule defined by α_t or equivalently β_t.

Recall that $\bar{\alpha}_t$ represents the cumulative probability that a token *has not been masked* up to time t. Conversely, $1 - \bar{\alpha}_t$ is the probability that a token *has been masked* by time t. This results in two opposite scenarios:

- **Early time steps** (small t, $\bar{\alpha}_t \approx 1$): Few tokens are masked. Predicting the original s_0 for these few masked tokens is relatively easy, as most of the context (unmasked tokens) is available.
- **Late time steps** (large t, $\bar{\alpha}_t \approx 0$): Most tokens are masked. Predicting the original s_0 becomes very challenging, requiring the model to infer from minimal context.

A common motivation for weighting comes from ELBO in continuous diffusion, which often leads to weighting terms that compensate for varying noise levels. Based on Section B, the forward marginal distribution is

$$q(s_t \mid s_0) = \prod_{i=1}^{L^\star} \bar{\alpha}_t \, \delta_{s_0^i}(s_t^i) + \big(1 - \bar{\alpha}_t\big) \, \delta_{\text{MASK}_{m(i)}}(s_t^i) \, .$$

Training maximizes the ELBO, which decomposes into timestep-wise Kullback-Leibler Divergence (KLD):

$$\mathcal{L} = \sum_t \mathbb{E}_q \Big[\mathrm{KL}\big(q(s_{t-1} \mid s_t, s_0) \parallel p_\theta(s_{t-1} \mid s_t)\big) \Big] \, .$$

In MDM, each KL term is non-zero only when $s_t^i = \text{MASK}_{m(i)}$. Conditioning on this event, the posterior $q(s_{t-1}^i \mid s_t^i = \text{MASK}_{m(i)}, s_0)$ is a categorical distribution whose parameters depend on the incremental masking rate. More precisely, let us denote π_t the probability that masking occurred at time t rather than earlier, so that

$$q(s_{t-1}^i \mid s_t^i = \text{MASK}_{m(i)}, s_0) = \pi_t \delta_{s_0^i}(s_{t-1}^i) + (1 - \pi_t) \delta_{\text{MASK}_{m(i)}}(s_{t-1}^i) \, .$$

Using Bayes' rule, we have:

$$\pi_t = \frac{\mathbb{P}(s_{t-1}^i = s_0^i) \mathbb{P}(s_t^i = \text{MASK}_{m(i)} \mid s_{t-1}^i = s_0^i)}{\mathbb{P}(s_t^i = \text{MASK}_{m(i)})} \, ,$$

where $\mathbb{P}(s_{t-1}^i = s_0^i) = \bar{\alpha}_{t-1}$, $\mathbb{P}(s_t^i = \text{MASK}_{m(i)} \mid s_{t-1}^i = s_0^i) = 1 - \alpha_t$, and $\mathbb{P}(s_t^i = \text{MASK}_{m(i)}) = 1 - \bar{\alpha}_t$. Therefore, we have:

$$\pi_t = \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)}{1 - \bar{\alpha}_t} = \frac{\bar{\alpha}_{t-1} - \bar{\alpha}_t}{1 - \bar{\alpha}_t} \, .$$

In the continuous-time limit, $\pi_t = \frac{-\bar{\alpha}_t'}{1 - \bar{\alpha}_t}$. As a result, the ELBO is equivalent up to constants to minimizing the objective

$$\mathbb{E}_{t \sim \mathrm{U}(0,1)} \left[w(t) \sum_{i=1}^{L^\star} \mathbb{E}_{q(s_t^i \mid s_0^i)} \big[\ell_i(\theta, s_0) \;\big|\; s_t^i = \text{MASK}_{m(i)} \big] \right] , \qquad w(t) = \frac{\bar{\alpha}_t'}{1 - \bar{\alpha}_t} .$$

This weighting ensures that each timestep contributes proportionally to the rate at which information about s_0 is destroyed by masking.

B.2 Unbiasedness of the Importance Weighting

At a fixed timestep t, each token position is independently masked with probability t, so that $\mathbf{1}\{i \in \mathcal{I}_t\} \sim \mathrm{Bernoulli}(t)$. Averaging the reconstruction loss only over masked positions therefore corresponds to random subsampling of tokens.

Without reweighting, the expected contribution of token i to the loss in linear scheduling is

$$\mathbb{E}[\mathbf{1}\{i \in \mathcal{I}_t\}\, \ell_i(\theta, s)] = t\, \ell_i(\theta, s)\,,$$

which underweights tokens at small t and biases the objective toward high-noise regimes. Multiplying by $1/t$ yields an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{t}\, \mathbf{1}\{i \in \mathcal{I}_t\}\, \ell_i(\theta, s)\right] = \ell_i(\theta, s)\,.$$

Therefore, the $1/t$ factor corrects for the subsampling induced by masking, ensuring that each token contributes equally in expectation across timesteps. This corresponds to inverse-probability weighting and is analogous to the time-dependent weighting used in denoising score matching objectives for diffusion models, where losses are rescaled to normalize signal-to-noise ratio across noise levels.

C Tokenizer Ablations

C.1 Audio Tokenizer Ablations

Model	# Codebooks	PESQ ↑	Content Enjoy ↑	Content Useful. ↑	Prod. Complex. ↑	Prod. Quality ↑	Down. Factor ↑
Higgs pretrained	8	3.168	5.670	6.192	1.559	6.510	960
	4	2.544	5.577	6.087	1.561	6.390	960
DAC pretrained	9	3.658	5.555	5.994	1.567	6.255	512
	4	2.396	5.069	5.415	1.603	5.597	512
DAC retrained	9	2.969	5.691	6.234	1.556	6.550	1024
	4	2.433	5.585	6.131	1.566	6.434	1024

Table 2: **Reconstruction metrics for different audio tokenizers.**

The audio tokenizer determines the audio token rate and, therefore, the sequence length corresponding to the audio stream. Since we use a fixed context length ($L^\star = 3256$) and clips with at most 30 seconds, we need a low-rate codec that still preserves perceptual quality. Here, we compare several RVQ-based codecs: a pretrained 24 kHz DAC [Kumar et al., 2023], a pretrained Higgs Audio v2 [Boson AI, 2025], and a DAC-style tokenizer trained on the same data as the main model. In Table 2, we present the reconstruction evaluation on LibriTTS-clean using PESQ [Rix et al., 2001] and Audiobox. Among the options, only two configurations fit our token budget for 30 s audio (shaded rows). We select Higgs Audio v2 with 4 codebooks for its superior rate–distortion trade-off. Based on these results, we chose the Higgs Audio v2 tokenizer with 4 codebooks decoding as the default audio tokenizer as it provides a strong and convenient rate–distortion trade-off. As expected, increasing the number of codebooks improves reconstruction but quickly becomes impractical under our sequence-length constraint. Consistent with prior observations for RVQ codecs [Kumar et al., 2023], we also find that training with more codebooks and decoding with fewer can preserve perceptual quality substantially better than training the low-rate model directly.

C.2 Image Tokenizer Ablations

The same sequence length constrains apply to the image tokenizer. We want a discrete image tokenizer that maps an image to as few tokens as possible while maintaining good image representation. Here, we compare discrete versions of Cosmos [Agarwal et al., 2025], IBQ [Shi et al., 2025], OpenMagVIT [Luo et al., 2025], Unitok [Ma et al., 2025], and MoVQGAN [ai-forever, 2023]. We evaluate the reconstruction FID on ImageNet at 256 and 512 resolution as well as CC12M [Changpinyo et al., 2021] and present these results in Table 3. Based on these results, we chose MoVQGAN as the default image tokenizer as it provided a good balance between high reconstruction performance, high compression and small vocabulary size.

D Training details

D.1 Optimal Global Hyper-Parameter Search

As highlighted in Section 4, most ablations in this work rely on optimal global hyperparameters scaled up with CompleteP [Dey et al., 2025]. In Figure 14 we present a Gaussian Process fit on

Model	Type	Latent / Code Dim	Vocab Size	Tokens Per Image	rFID ↓
CC12M					
cosmos-ci8x8 [Agarwal et al., 2025]	Continuous	16	-	1024	1.37
cosmos-di16x16 [Agarwal et al., 2025]	FSQ	6	65536	256	3.50
cosmos-di8x8-360p [Agarwal et al., 2025]	FSQ	6	65536	1024	1.80
ibq-262144 [Shi et al., 2025]	IBQ	256	262144	256	0.89
movqgan-270m [ai-forever, 2023]	**MoVQ**	**256**	**16384**	**1024**	**0.50**
openmagvitv2 [Luo et al., 2025]	LFQ	18	262144	256	0.82
unitok [Ma et al., 2025]	MCQ	64	4096 (x8)*	256 (x8)	0.54
ImageNet					
cosmos-ci8x8 [Agarwal et al., 2025]	Continuous	16	-	1024	1.02
cosmos-di16x16 [Agarwal et al., 2025]	FSQ	6	65536	256	4.38
cosmos-di8x8-360p [Agarwal et al., 2025]	FSQ	6	65536	1024	0.95
ibq-262144 [Shi et al., 2025]	IBQ	256	262144	256	1.55
movqgan-270m [ai-forever, 2023]	MoVQ	256	16384	1024	0.55
openmagvitv2 [Luo et al., 2025]	LFQ	18	262144	256	1.67
unitok [Ma et al., 2025]	**MCQ**	**64**	**4096 (x8)***	**256 (x8)**	**0.36**
ImageNet - 512×512					
cosmos-ci8x8 [Agarwal et al., 2025]	**Continuous**	**16**	-	**4096**	**0.07**
cosmos-di16x16 [Agarwal et al., 2025]	FSQ	6	65536	1024	1.33
cosmos-di8x8-360p [Agarwal et al., 2025]	FSQ	6	65536	4096	0.51
ibq-262144 [Shi et al., 2025]	IBQ	256	262144	1024	0.50
movqgan-270m [ai-forever, 2023]	**MoVQ**	**256**	**16384**	**4096**	**0.17**
openmagvitv2 [Luo et al., 2025]	LFQ	18	262144	1024	0.53
unitok [Ma et al., 2025]	MCQ	64	4096 (x8)*	1024 (x8)	0.23

Table 3: **Reconstruction FID for different image tokenizers.** ImageNet consists of 50k validation examples and CC12M [Changpinyo et al., 2021] consists of 50k samples from the full dataset. Best FID per section is shown in **bold**. When a continuous model achieves the best FID, the best discrete model is also bolded. *UniTok uses 8 categorical predictions of size 4096 (x8 underlying codes that are merged into a single token).

2900 trial runs at small scale (320M parameters in total, including 80M non-embedding ones, 13B tokens) to determine optimal global hyperparameters for tri-modal MDM. The optimal values are highlighted in red. We initialize the per-module multiplier search from this optimum to better seed the search process.

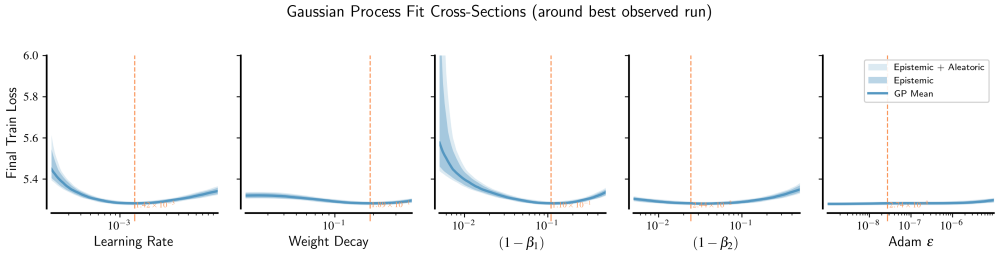


Figure 14: **Small-scale hyper-parameter search for 0-shot transfer**. Final average cross-entropy loss for a sweep of 5 AdamW hyperparameters (learning rate, β_1, β_2, weight-decay, ϵ) on a 350M model trained for 13B tokens. The plot shows cross-sections through the hyperparameter landscape through the best found hyperparameters.

D.2 Runtime as Function of Batch Size

As explored in Section 4.2, the SDE parametrization allows a wide range of batch sizes to be used. Typically, bigger batch sizes allow less iterations, which *can* reduce runtime. Two knobs are available to increase the batch size: increasing the number of nodes, or modifying the per-GPU batch size. In practice, overall throughput grows sub-linearly with number of nodes because of communications on the cluster. Furthermore, for smaller batch sizes, hosted on a single node, another phenomenon plays a role: once the batch size is too small, the GPU becomes idle because small batch sizes do not benefit as much from parallelism. This is measured in Figure 15. The runtime diminishes slowly as the per-GPU batch size increases, and diminishes sharply as the number of nodes increases. Highest node count reduces wall-clock time significantly, but also reduces FLOP-efficiency because of sub-linear scaling.

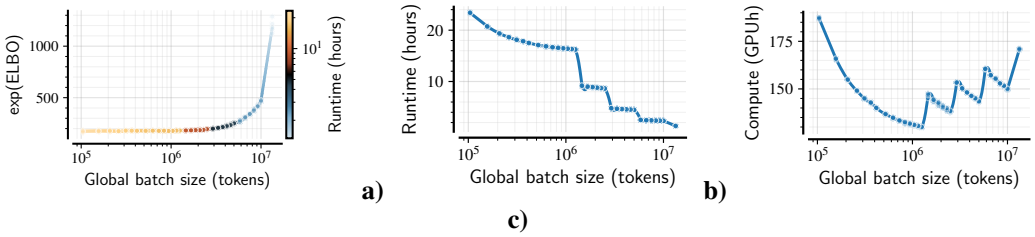


Figure 15: **Performance as function of physical batch size**. In **a)** we see impact of batch size on total runtime, with the effect of the node-count clearly visible: more nodes reduces wall-clock time faster than simply saturating GPUs. In **b)** we see the runtime with 1, 2, 4, 8 and 16 nodes respectively. Doubling node count approximately halves the runtime. Finally, in **c)**, we see that saturated GPUs have much more efficient FLOP usage than non-saturated ones. Doubling node count never allow to recover training efficiency of fewer nodes, because of communication overhead.

D.3 Hyperparameters for the Unified 3B Tri-modal MDM

Table 4 highlights all training details for the 3B model.

E MDM with Per-module Hyperparameters

To simplify and parallelize our analysis, we rely on CompleteP [Dey et al., 2025] for width and depth transfer using global hyperparameters (Appendix Section D.1) for all ablations done in this work, however, recent insights from Mlodozeniec et al. [2026] highlight that we can further improve performance by optimizing per-module hyper-parameter multipliers for AdamW (learning rate, weight decay, β_1, β_2, and ϵ).

Training Details. Each hyperparameter for each weight gets a unique multiplier. The multipliers for all depth-repeated blocks are parameterized as a product of a depth-dependent factor, and a module-type factor – all weights at the same depth get the same depth factor, and all weights of the same type (e.g., all QKV weights) get the same module-type factor. We initialize the local random search method from [Mlodozeniec et al., 2026] at the best global hyperparameters found using random search, shown in Figure 14. We conduct the search using a transformer with 8 blocks of width 1024, totaling 320M parameters (including 80M non-embed parameters) trained at a horizon of 13B multimodal tokens. The results of the search, and the speed-up obtained at this base model size, is reported in Figure 16.

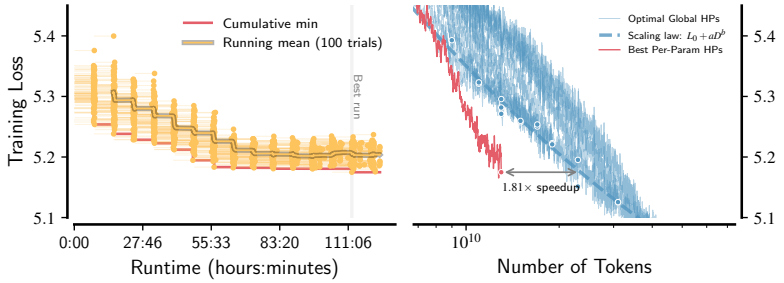


Figure 16: **Per-module hyperparameter search on a small scale model for 0-shot transfer**. Average final ELBO for a sweep of AdamW's per-module hyperparameters as well as initialization scales for a 350M model (including 80M non-embedding parameters) trained for 13B tokens, using a batch size of 256. Per-module tuning yields a **1.81**× reduction in token count to achieve an equivalent loss.

Multiplier analysis. In Table 5, we highlight the results of the search, listing the per module multipliers. Notably, the resulting multipliers are highly structured rather than uniform: embedding and

Model	
N blocks	24
Dimension	3072
N attention heads	24
QK Norm	Yes
Normalization	RMSNorm
Pre-norm	Yes
Post-norm	No
MLP style	SwiGLU
SwiGLU hidden dimension factor	2.75
Positional embedding	RoPE
Weight initialization (base_width)	trunc_normal(std=0.02)
Training parameters	
Batch size	3072
Sequence length	3256
Optimizer	AdamW
Base LR	$9e-4$
Base AdamW ϵ	$1e-8$
Base AdamW β_1	0.9
Base AdamW β_2	0.95
Base weight decay	0.1
LR warmup	$2,000$ steps
LR schedule	Cosine
Min LR	$1e-6$
Z-loss weight	$1e-5$
Training duration	$1,000,000$ steps
Hyperparameter Transfer Strategy	CompleteP [Dey et al., 2025]]
Modality sampling rate (text-only, image-text, audio-text)	$[0.33, 0.33, 0.33]$
Text tokens seen during training	3.4T
Image samples seen during training	1B
Audio samples seen during training	1B
Tokenizers	
Text	Tiktoken [OpenAI, 2023]
Image	SBER-MoVQGAN [ai-forever, 2023]
Audio	Higgs Audio Tokenizer v2 (4 codebooks) [Boson AI, 2025]
Vocabulary (incl. special tokens)	
Total	$117,698$
Text	$100,281$
Image	$16,387$
Audio	$1,027$
Text transformations	
P(Token packing)	0.95
P(Random sequence subsample)	0.05
Image transformations	
Target resolution	$256x256$
RandomResizedCrop	$[0.8, 1.0]$
Resize with white padding	-
P(RandomResizedCrop)	0.5
P(Resize with white padding)	0.5
Normalization μ and σ	$[0.5, 0.5]$
Audio transformations	
Max duration	30 seconds
Number of frames	25

Table 4: Model and training details for the 3B multimodal MDM.

Table 5: Optimal per-module hyperparameter multipliers found for the MDM Multimodal model. Depth factors apply to all layers where the $\frac{\texttt{block depth}}{\texttt{total depth}}$ falls within the highlighted fraction (as counted from the network input towards output).

Category	Module / Depth	LR	WD	α_1	α_2	ϵ	Init Scale
Standalone	Embedding weights (Audio)	2.192	1.009	0.962	1.493	1.494	1.826
	Embedding weights (Image)	1.013	0.864	2.108	0.685	0.734	0.554
	Embedding weight (Text)	3.937	1.593	1.421	1.791	0.317	0.379
	Unembedding weights (Audio)	1.633	1.510	3.442	0.594	0.742	3.422
	Unembedding weights (Image)	1.655	1.213	1.929	1.042	0.635	1.524
	Unembedding weights (Text)	3.008	0.737	1.346	0.955	1.206	0.341
	Unembedding norm weights	2.305	0.817	4.508	2.740	1.938	2.175
Blocks	`attn_qkv_weight`	1.714	0.821	0.173	0.557	0.391	2.498
	`attn_proj_weight`	0.630	0.354	0.256	0.339	1.627	4.732
	`attn_q_norm_weight`	0.535	0.731	1.530	0.902	0.848	1.344
	`attn_k_norm_weight`	0.754	0.497	1.074	0.822	0.368	0.436
	`mlp_gate_weight`	0.489	0.634	1.171	1.870	4.913	0.643
	`mlp_fc1_weight`	1.271	1.295	1.590	3.309	1.415	1.944
	`mlp_fc2_weight`	1.405	1.308	2.684	0.655	1.790	0.878
	`norm1_weight`	1.311	1.105	0.282	1.161	1.477	2.171
	`norm2_weight`	0.899	0.525	1.533	1.789	0.712	1.189
Depth Factors	$0 - 50\%$	1.102	0.725	1.030	3.053	0.663	0.997
	$50 - 100\%$	0.877	1.018	0.911	1.149	2.645	0.485

unembedding weights favor substantially larger effective learning rates (up to $\sim 4\times$), while attention projections and MLP gates are tuned more conservatively, often with increased ϵ for numerical damping. The learned depth factors further indicate smaller steps and stronger stabilization in later blocks, consistent with increasing sensitivity of deep representations to update noise.

F Extended Scaling Laws Results

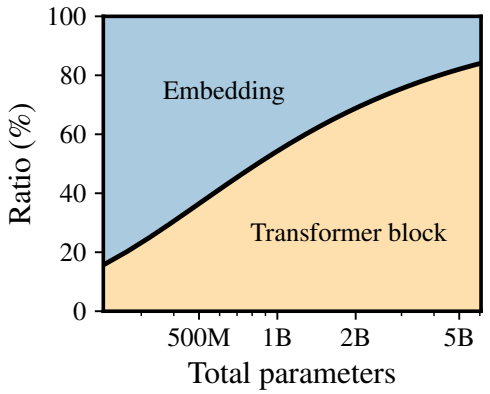


Figure 17: **Percentage of transformer-block parameters in the total parameter count**. Trimodality forces a larger vocabulary, yielding a ratio below 50% for small models. Other models with even larger vocabulary suffer form the same issue.

F.1 Isonoise and Isohorizon Intuition

The SDE is the continuous limit of the gradient flow elicited by AdamW. Intuitively, the *SDE horizon* corresponds to the trajectory length in parameter space (extending from the origin), while the *SDE drift* controls the scale of stochastic fluctuations induced by gradient noise. When increasing the token budget D, the extra tokens can be used in two ways: **(a)** conserving the drift while extending the horizon (*isonoise*, $\gamma{=}0$), which corresponds to the standard LLM setup where more tokens means more iterations; or **(b)** conserving the horizon while reducing the drift (*isohorizon*, $\gamma{=}1$), which modulates the optimizer hyperparameters instead. AdamW's learning rate scales as $\text{lr} = \text{lr}^{\text{base}}\sqrt{\kappa}$ because, in the SDE limit, the step size must compensate for changes in gradient noise magnitude; similarly, momenta scale as $\beta_i = (\beta_i^{\text{base}})^{\kappa}$ to preserve the effective averaging window.

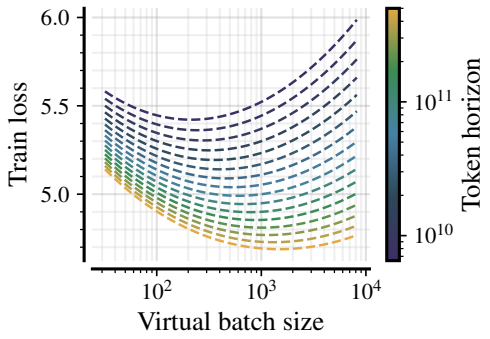


Figure 18: **Isotoken curves at various virtual batch sizes.** The optimal allocation between virtual batch size and virtual iteration count corresponds to a factor $\gamma^* \approx 0.44$. The bottom of the U-curve would be B_{opt} in non-SDE parametrization.

F.2 Computation and extrapolation for critical iteration count

Informally, S_{crit} is the iteration count at which the SDE approximation breaks. Formally, for a loss $\mathcal{L}$ depending on model size N, iteration count S' and batch size B at token horizon $D = BS'L$, we define it as:

$$S_{\text{crit}}(N, D) = \min\left\{ S \;\middle|\; S \in \arg\min_{S'} \mathcal{L}\left(N, S', \frac{D}{S'L}\right)\right\} \tag{6}$$

In the $S \geq S_{\text{crit}}$ region, the function $L(N, S, D/(SL))$ is constant since the SDE approximation emulates the optimal virtual batch iteration count $\tilde{S}$. From previous empirical observations, we know that the dependency in N can be removed. However the exact form of the dependency in D is unknown: it is a slow growing function, but we do not have enough points to identify the correct form among many candidates. Instead, we predict the trend of Figure Figure 7 by combining two power-laws.

Firstly, we build a model of the minimum iteration count S required to reach some target loss L:

$$S(L) = C_0 + C_1 L^{-C_2}, \tag{7}$$

with learnable coefficients C_0, C_1, C_2. This power law trend makes sense, as S decreases when L increases as one would expect. We fit it at constant N only, since S_{crit} is independent of N (Figure 5).

Secondly, we combine with a normal Chinchilla-like neural scaling law:

$$L(D) = C_0 + C_1 D^{-C_2}, \tag{8}$$

that are known to extrapolate well. Finally, we combine them to get the final estimate $S(D) = S(L(D))$. This indirect estimation is more robust than the direct one $S(D)$, since we have more points to fit $L(D)$. Finally, we report $B(D) = D/(S(D)L)$, i.e the final estimate of the critical batch size as function of the token horizon.

F.3 Computation of FLOPs for experiments

We rely on the formula in appendix **H.1** of Busbridge et al. [2025] to compute the FLOPs and the model size. In particular, we do not account for input/output embedding, as the size of our multi-modal vocabulary (117k) makes these embedding matrices bigger than the transformer backbone for small models. The ratio between transformer parameters and embedding parameters is illustrated in Figure 17. The value of N reported in scaling laws fit always use the non-embedding parameters only (which are up-to 5 times smaller than the total model's size). The FLOPs C reported in scaling law account for everything. We compute the optimal $N^\star(C)$ and $D^\star(C)$ by minimizing the parametric loss under the constraint $C = \text{FLOP per token}(N) \times D$. Since we do not account for embeddings parameters in total model size, the popular FLOP per token$(N) = 6N$ does not apply out-of-the-box. Instead, we minimize the parametric loss $L(N, D = C/\text{FLOP per token}(N))$ via linear-search over N to plot compute-optimal curves in Figure 3 and Figure 12. This effect is more striking for small models where the embedding size is significant. For larger models, we found that the approximation FLOP per token$(N) \approx 6$ holds. In that case, minimizing the parametric form

$$L(N, D) = C_0 + \left(C_1 N^{-a/b} + C_2 D^{-1}\right)^b,$$

under constraint $C = 6ND$ works well. By monotonicity, this is equivalent to minimizing $C_1 N^{-a/b} + C_2 D^{-1}$, which admits the following minimizers:

$$N^\star(C) = G^{-1}(C/6)^\tau \text{ and } D^\star(C) = G(C/6)^{1-\tau}, \quad \text{with } G = \left(\frac{bC_2}{aC_1}\right)^\tau \text{ and } \tau = \frac{b}{a+b}.$$

F.4 Scaling laws for Uni-Modal Text MDM

We also perform scaling laws run on uni-modal text models, using CompleteP parametrization, but *without* SDE scaling rules. Every model size relies on a different batch size to maximize GPU occupancy. The total sequence length is 4,096 with packing and truncation (no padding). Training curves as function of (N, D) FLOP budget are given in Figure 19 and the scaling laws predictions are reported in Figure 20. Finally, a full pretraining of a text-only 7B model at 4096 sequence-length and

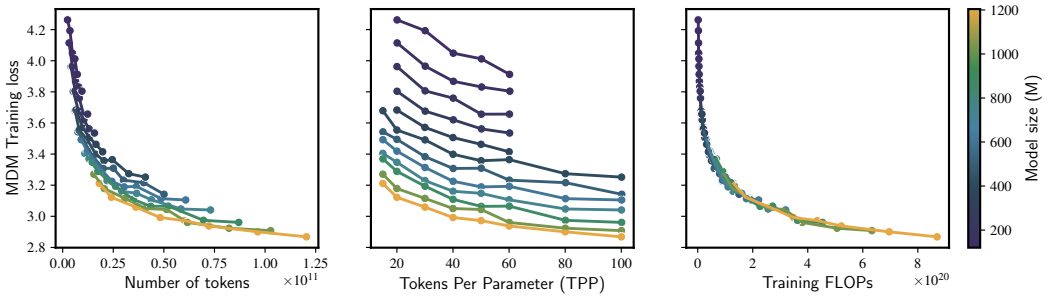


Figure 19: **Training curves for uni-modal text MDM models trained under CompleteP, without SDE re-parameterization.**

F.5 Additive Form and Cross-Modal Interaction

The scaling law of Equation 3 uses a non-additive form. To identify whether its cause is the MDM objective, the SDE parametrization, or multimodality, we conduct a text-only MDM scaling study with the CompleteP parametrization. Results are given in Figure 19 and Figure 20a), b) in subsection F.4. We see that under CompleteP, with MDM training on text only, the additive form of Chinchilla is sufficient. This is compatible with the observations of the Quokka scaling laws [Ni et al., 2025].

We also run text-only MDMs ($< 8\times10^{20}$ FLOPs scale) with CompleteP + SDE scaling (using the implicit form from Mlodozeniec et al. [2026]) and once again find a good fit (MRE $< 1\%$) using the additive Chinchilla form: $L = 2.07 + 6.07\, N^{-0.4} + 1.85\, D^{-0.3}$ (Figure 20c)).

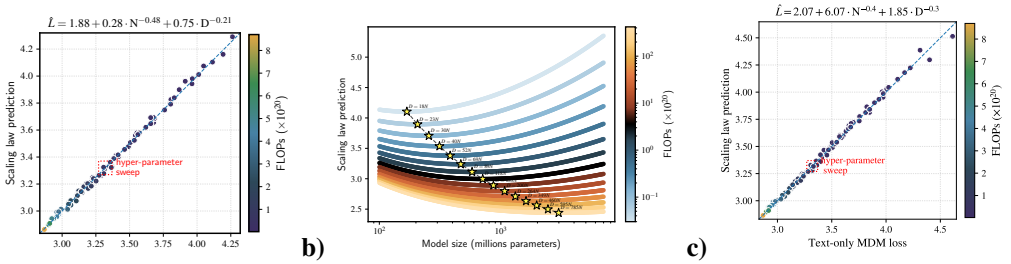


Figure 20: **Scaling laws for uni-modal text MDMs**. **a)** Scaling law predictions for text-only MDM models, using CompleteP. **b)** Iso-FLOP curves for text-only MDM models under CompleteP parameterization (no SDE scaling). **c)** Text-only MDMs under CompleteP + SDE scaling.

The key observation is that for outer exponent $b < 1$, dropping the interaction term systematically over-estimates the true ELBO at "small" scale, where N and D interact. This is exactly the regime we operate in. Otherwise, when $N \gg D$ or $D \gg N$, the two forms agree on the asymptotic exponents (differing only in the leading constant). So the non-additive form mostly corrects predictions in the intermediate regime $N \approx D$. Since we consider model sizes that are modest compared to Chinchilla (which went up to 16B dense models [Hoffmann et al., 2022]), we depend more on this error term. Input/output embedding sizes must be taken into account for proper FLOP accounting (see Figure 17 and subsection F.3 for their impact at small scale), and in our case the vocabulary split across modalities (85% for text, nearly 14% for image, and less than 1% for audio) is unbalanced compared to the data mixture (33% text, 33% image, 33% audio). This likely explains the non-additive form.

G Modality Mixing Ratios

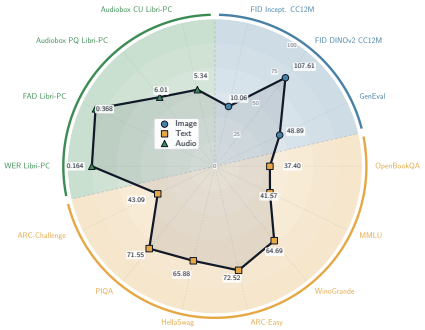
Understanding how to combine data is of critical importance for multimodal models. We take an empirical stride towards quantifying this by carefully constructing an experiment where we vary the global modality mixing ratios, $\{w_{\text{text}}, w_{\text{image}}, w_{\text{audio}}\}$, which respectively control the amount of text, image-text and audio-text data that is present within the pretraining mixture. We launch a total of 15 experiments with a model of size 320M (including. 80M non-embedding ones) and 13B tokens. We set a minimum of 20% per modality to avoid degeneracy with out-of-distribution modalities. Results are shown in Figure 13. Unsurprisingly, we find that the loss decreases as the corresponding mixture weight increases. However, we do not witness synergies between modalities *at these scales*: the average loss on a modality is independent on the relative ratio of the two other modalities. Therefore, the default choice of $[1/3, 1/3, 1/3]$ appears reasonable. While there exists prescriptive mechanisms to determine mixing ratios, they either fail to tackle multimodal models [Chen et al., 2026] or to account for data-repetition [Shukor et al., 2025a].

H Full 3B Model Results

Figure 21 provides the full radar visualization and metric breakdown for the 3B tri-modal model across all modalities and datasets.

I Token-Matched AR Baseline

We compare text-only AR and MDM models at matched token budgets. We train both at the Chinchilla optimum $D = 20N$ (1B parameters on 20B tokens, and 7B parameters on 140B tokens) and evaluate on LM Harness (Figure 22). At equal token budget, MDM trails AR on most benchmarks. Since the comparison is token-matched (equal D at the Chinchilla optimum $D = 20N$), the gap means that to reach AR's scores MDM must consume more tokens, i.e., AR is more token-efficient in this regime, consistent with [Sahoo et al., 2026].



Modality	Dataset and Metric
Image generation & composition	
Test set	FID-Inception 10.41, FID-DINOv2 112.12.
CC12M	FID-Inception 10.06, FID-DINOv2 107.61.
GenEval	Single Obj 93.12, Two Obj 63.38, Counting 33.44, Colors 64.89, Position 11.50, Color Attr. 27.00, Overall 48.89.
Language (LM Harness)	
Harness	OpenBookQA 37.40, TruthfulQA MC2 40.76, BBH 24.97, MMLU 41.57, WinoGrande 64.69, ARC-Easy (norm) 72.52, HellaSwag (norm) 65.88, LogiQA2 (norm) 30.66, PIQA (norm) 71.55, ARC-Challenge (norm) 43.09.
Audio generation	
Test set	FAD 0.218, WER 0.124, PQ 6.89, CU 6.20, CE 5.45, PC 1.89.
LibriSpeech-PC	FAD 0.368, WER 0.164, PQ 6.01, CU 5.34, CE 5.07, PC 3.01.

Figure 21: Tri-modal 3B overview across image, text, and audio. **Left:** radar summary (larger radius indicates better normalized score; vertex labels are raw values). Normalization is mixed by metric type: bounded metrics use natural bounds (GenEval/LM percentages to 100, WER to 1, Audiobox to 10), while unbounded Frechet metrics (FID-Incept, FID-DINOv2, FAD) use ECDF calibration from references, with lower-is-better inversion $s = 1 - F(x)$, where F is the ECDF. **Right:** raw unnormalized metrics.

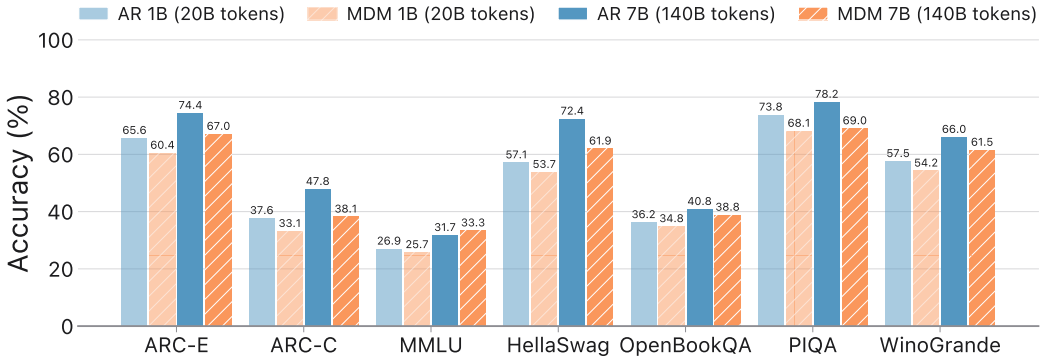


Figure 22: **Downstream accuracy of text-only AR and MDM models trained on the same number of tokens.** Both are at the Chinchilla optimum $D = 20N$.

J Anti-Masking Detailed Results

Table 6 reports multimodal generation metrics and Table 7 reports LM Harness results comparing standard training versus anti-masking.

Table 6: **Multimodal ablation results comparing standard MDM training versus anti-masking (1.3B multimodal).**

Model	FID (Inception) ↓		FID (DINOv2) ↓		FAD ↓	
	Train Data	CC12M	Train Data	CC12M	Train Data	LibriSpeech
Base	18.69	26.77	306.19	395.73	0.24	0.79
Anti-mask	**17.81**	**21.04**	**302.61**	**361.00**	**0.22**	**0.55**

Table 7: **Anti-masking results on the evaluation harness.** We report mean accuracy ± standard deviation (7B text).

	OpenBookQA	TruthfulQA	BBH	MMLU	Winogrande	ARC-Easy	HellaSwag	LogiQA2	PIQA	ARC-Challenge
Base	32.80 ± 2.10	**43.12** ± 1.45	21.52 ± 0.46	30.20 ± 0.39	52.01 ± 1.40	63.72 ± 0.99	52.20 ± 0.50	24.75 ± 1.09	67.19 ± 1.10	31.66 ± 1.36
Anti-mask	**33.00** ± 2.10	36.97 ± 1.41	**27.05** ± 0.50	**32.86** ± 0.39	**54.06** ± 1.40	**66.04** ± 0.97	**55.15** ± 0.50	**26.27** ± 1.11	**67.41** ± 1.09	**34.90** ± 1.39

K Inference Hyperparameter Sweeps

Figure 23 and Figure 24 show per-hyperparameter sweeps for text-to-image and text-to-speech generation.

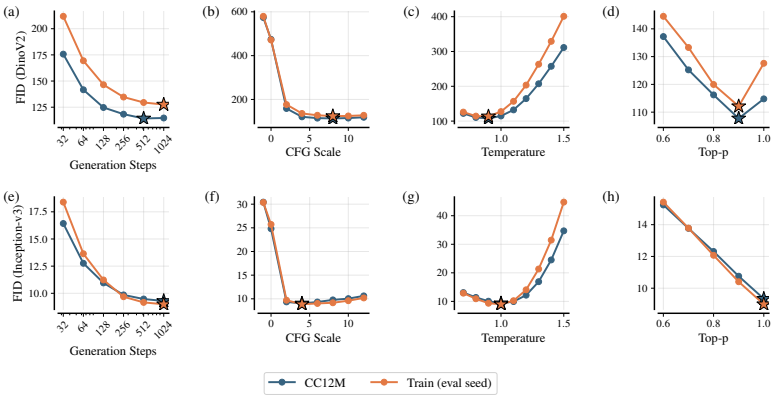


Figure 23: **Text-to-image hyperparameter ablations.** We generate images from text prompts and compute FID against reference images on two datasets: CC12M in blue and train (eval seed) in orange. Top row (a–d) shows FID computed using DINOv2-L features, bottom row (e–h) shows FID using Inception-v3 features. Panels (a,e) vary generation steps; (b,f) vary CFG scale; (c,g) vary temperature; (d,h) vary top-p. Stars indicate optimal values for each metric-dataset combination.

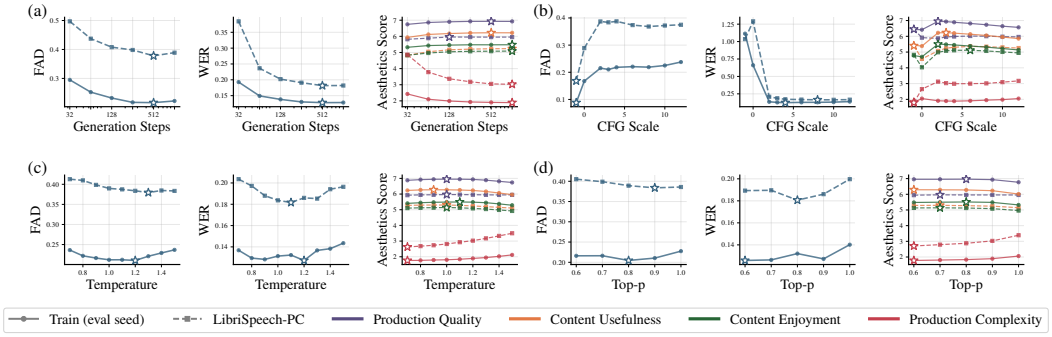


Figure 24: **Text-to-speech hyperparameter ablations.** We evaluate audio generation from text prompts on two datasets (train set, LibriSpeech-PC) using ground-truth durations, with three metrics: FAD, WER, and Audiobox Aesthetics. Top row shows (a) steps sweep and (b) CFG sweep. Bottom row shows (c) temperature sweep and (d) top-p sweep. $\star$ indicate the best hyperparameter for each metric. For Production Complexity, lower values indicate simpler audio and are considered as preferable.

L Masking Schedules for Image and Audio Generation

To determine the impact of the masking schedule on multimodal MDM training and generation quality, we evaluate four distinct schedules – linear, cosine, polynomial, and geometric – implemented using the continuous-time ELBO weighting in Shi et al. [2024]. We train a 1B model for 100k steps under each schedule, keeping all other hyperparameters the same.

First, we evaluate image generation quality at 256×256 resolution using 1024 diffusion steps with CFG scales ranging from 1.0 to 10.0, temperature $T = 0.9$, and nucleus sampling (top-$p = 0.9$). Image quality is measured using both FID-Inception and FID-DINOv2, computed over 8,192 generated samples on two datasets, CC12M and our train mixture (eval seed). Figure 26 shows that the polynomial schedule consistently achieves the best image quality across both metrics and datasets

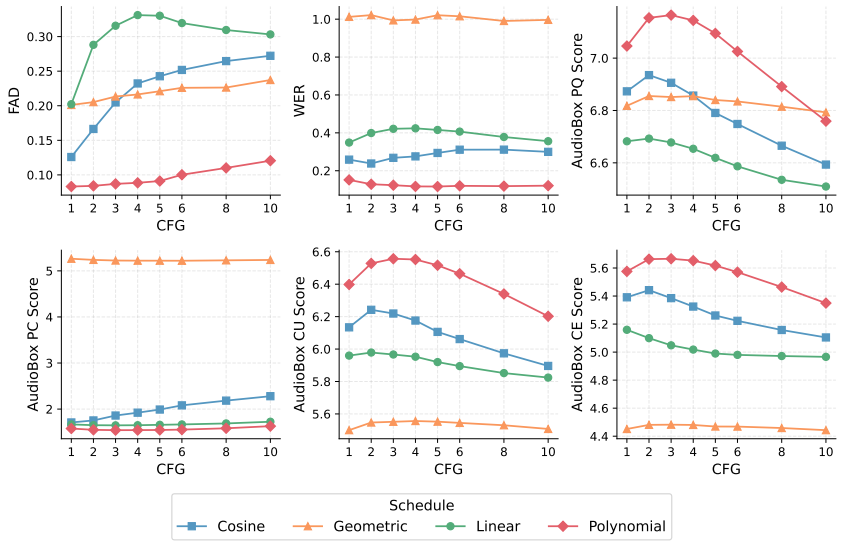


Figure 25: **Masking schedule ablation for audio generation across guidance scales.** We evaluate four masking schedules (linear, cosine, polynomial, geometric) on ground-truth length audio generation quality using FAD, WER, and AudioBox Aesthetics metrics on our train mixture. Models are evaluated across CFG scales 1.0-10.0.

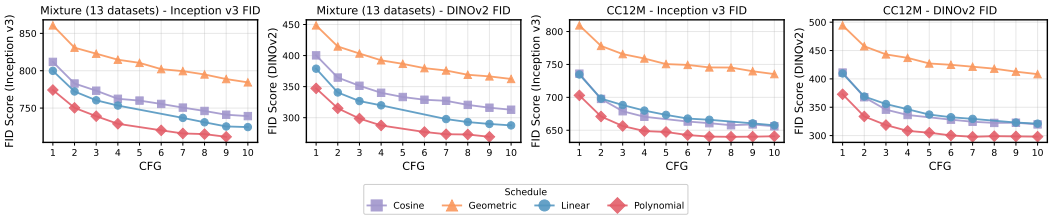


Figure 26: **Masking schedule ablation across guidance scales.** We evaluate the generation quality of four masking schedules (linear, cosine, polynomial, geometric) in CC12M and our train mixture (eval seed). Models are evaluated across CFG scales 1.0-10.0.

among the four schedules tested. Both metrics agree that polynomial yields superior generation quality, with optimal performance in the CFG range of 7 to 9.

Then, we evaluate audio generation quality with ground-truth durations using 512 diffusion steps with CFG scales ranging from 1.0 to 10.0, temperature $T = 1.0$, and nucleus sampling (top-$p = 0.9$). We measure audio quality using FAD, WER, and AudioBox Aesthetics computed over 10,000 generated samples from the dataset. Figure 25 shows that the polynomial schedule also consistently achieves the best audio generation quality across all six metrics evaluated. Unlike in image generation, the optimal CFG range is between 1 and 3.

M Inference Latency

At large canvas size, generation latency remains a major challenge for MDMs compared to AR models. We report the absolute wall-clock latency per sample of the 3B model for text, image and audio generation (Table 8). We benchmark each task on a single H100 in bfloat16. We disable EOS prediction to make the runtime predictable, so the reported latency is an upper bound. The number of steps is fixed, and we report the median over 5 independent seeds after `torch.compile` warmup. Bidirectional training forfeits KV-caching and context prefilling. The latency is linear in the step count, as expected (256, 512 and 1024 steps for 256px images). Images are generated at 256px resolution (1024 tokens) with 1024 steps. Tokenization and detokenization account for less than 1% of the total cost in every modality.

Table 8: **Wall-clock latency of the 3B tri-modal MDM.**

Task	Cond. tokens	Gen. tokens	Steps (NFE)	Median latency (s)
Text → Text	259	256	256	2.93
Text → Image (256px)	118	1024	256	5.88
Text → Image (256px)	118	1024	512	11.70
Text → Image (256px)	118	1024	1024	23.35
Text → Image (448px)	118	3136	1024	67.17
Image → Text	1029	100	100	1.78
Text → Audio ($\sim$10 s)	261	1000	250	4.60
Audio → Text	1437	100	100	2.42

N Compute costs

The 3B hero run took approximately 3 weeks on 1024 H100-80GB Nvidia GPUs, corresponding to 500k GPU-hours.

The grid search used for hyper-parameter search took $\approx$ 3h per run on 8xH100 GPUs, for a total cost below 100k GPU-hours.

The scaling law study required a total of 3072 GPUs, with variable GPU-count per model size, for a few days. Its total cost is estimated to sit below 200k GPU-hours on H100-80GB.

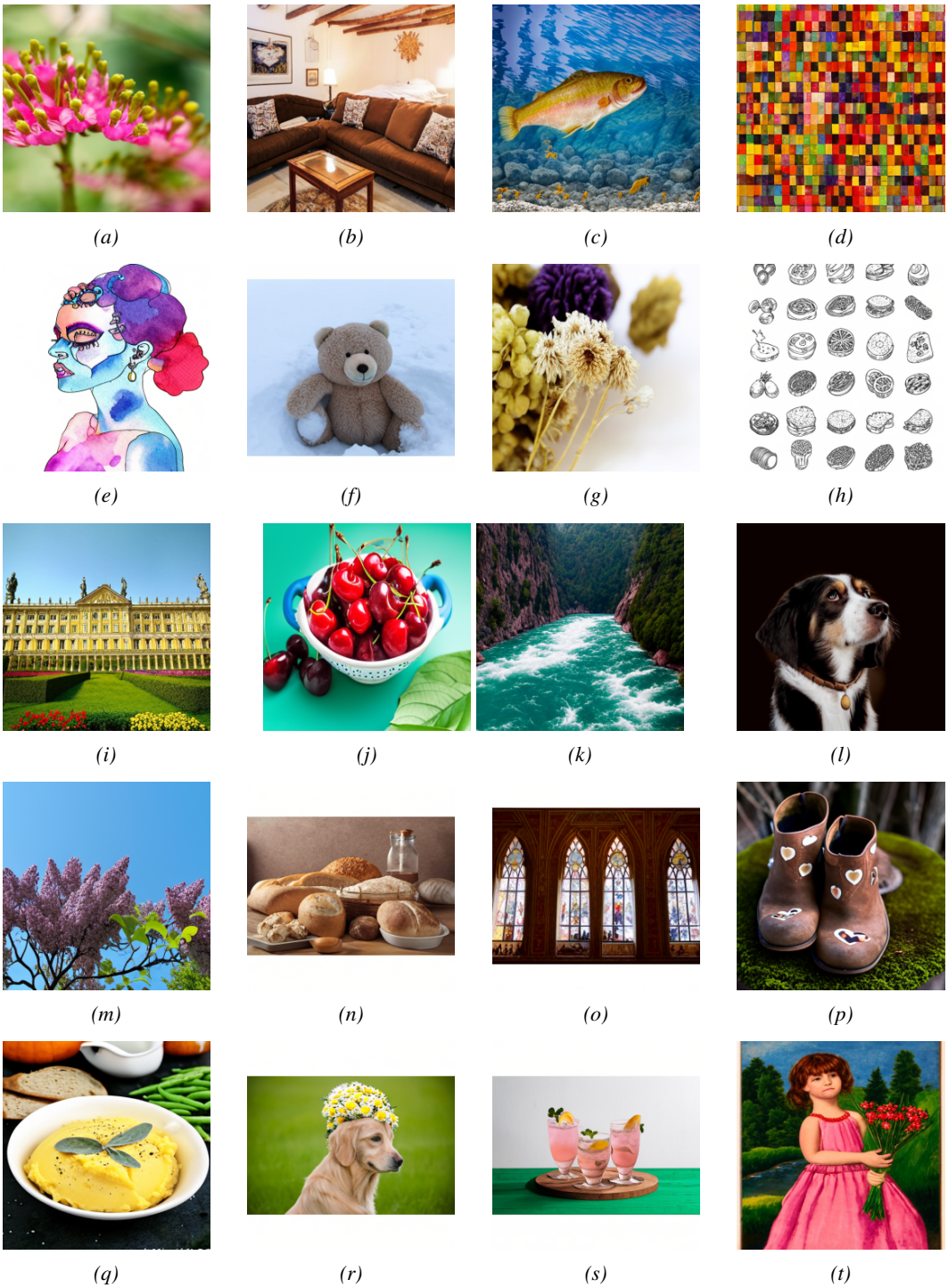
Pretokenization of audio and images was done on A100-40GB GPUs, with a much smaller total cost. The cost of other experiments mentioned in the paper is negligible compared to these main experiments.

O Broader Impacts

We introduced a method to train tri-modal models. These models exhibit generative capabilities in both text, image, and audio. As such, there are subjected to similar vulnerabilities as other generative models: *deepfakes*, voice cloning, and hallucinations. However, we do not release model weights, so no misuse is possible. Finally, all corpora are PII-filtered to prevent unintended memorization.

P Extended generations

We present extended generations in Figure 27, Figure 28, Figure 29, and Figure 30 and their respective complete list of prompts in Table 9, Table 10, Table 11, and Table 12. Note that the prompts come from synthetic captions and were selected among a larger set of generations based a mix of quality filtering and diversity.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Figure 27: **Samples generated by our model with different prompts. See Table 9 for the extensive list of prompts.**



(u) (v) (w) (x)



(y) (z) (aa) (ab)



(ac) (ad) (ae) (af)

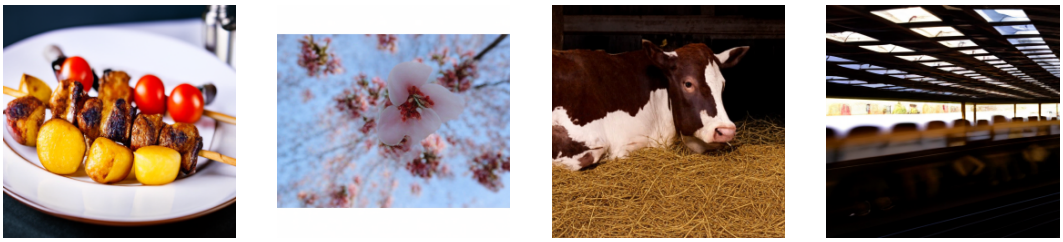


(ag) (ah) (ai) (aj)



(ak) (al) (am) (an)

Figure 28: **Samples generated by our model with different prompts.**



(ao) *(ap)* *(aq)* *(ar)*



(as) *(at)* *(au)* *(av)*



(aw) *(ax)* *(ay)* *(az)*

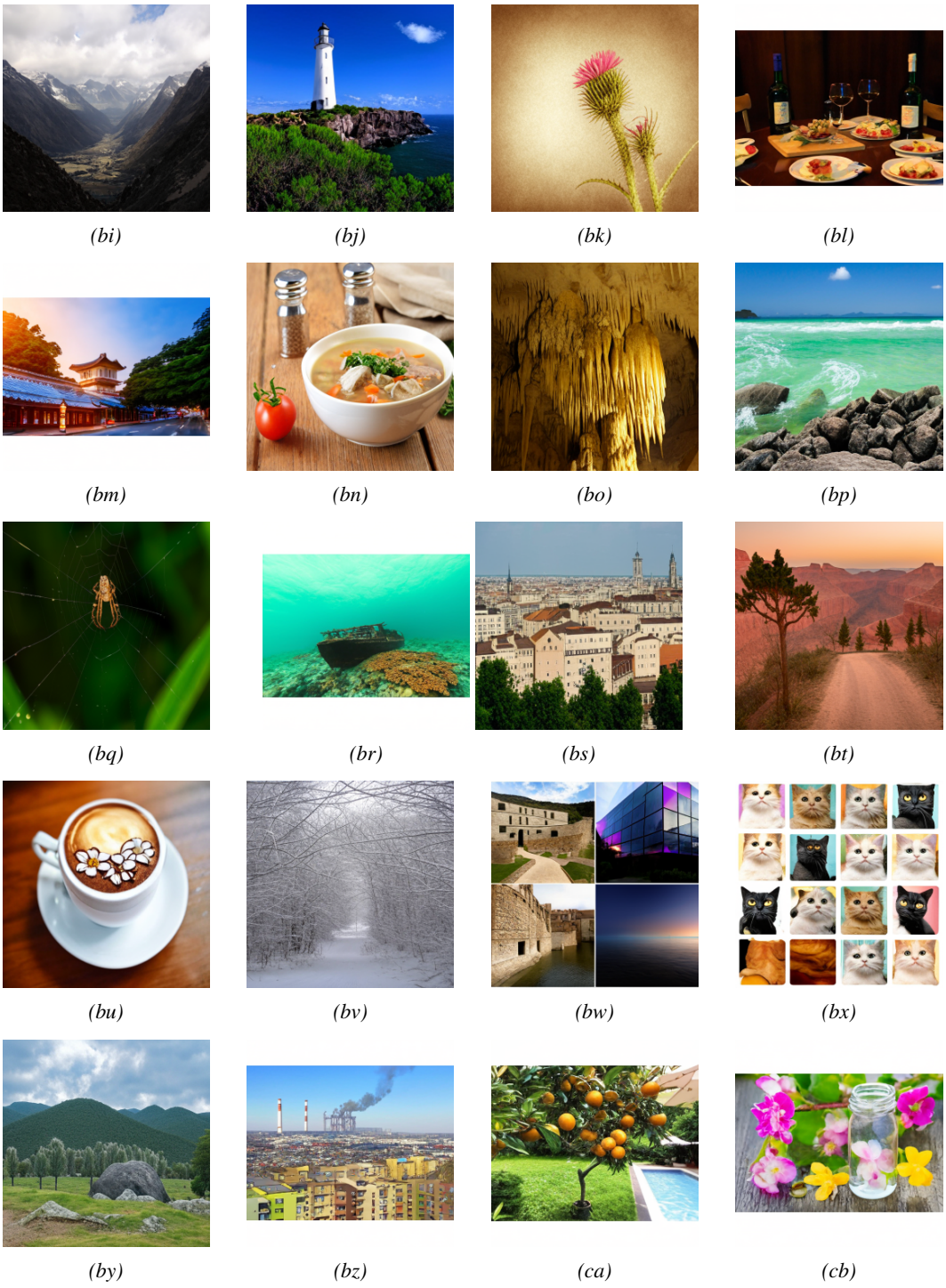


(ba) *(bb)* *(bc)* *(bd)*



(be) *(bf)* *(bg)* *(bh)*

Figure 29: **Samples generated by our model with different prompts.**



(bi) *(bj)* *(bk)* *(bl)*

(bm) *(bn)* *(bo)* *(bp)*

(bq) *(br)* *(bs)* *(bt)*

(bu) *(bv)* *(bw)* *(bx)*

(by) *(bz)* *(ca)* *(cb)*

Figure 30: **Samples generated by our model with different prompts.**

(a) Lagerstroemia macrocarpa Wall, Queen's flower, Queen's crape myrtle, Lythraceae. Close-up of a vibrant pink flower with a yellow center in full bloom, sharply focused against a blurred green outdoor background. Numerous small green buds at different stages surround the flower on thin green stems.
(b) A cozy living room with a plush brown sofa and patterned pillows. A wooden coffee table sits in front, framed wall art above, rustic wooden ceiling, and a bed with a white comforter in the background. Decorative lighting and wall hangings add warmth.
(c) A photograph of a rainbow trout in a body of water. The trout is facing the left side of the image. The water is clear and there are some small rocks at the bottom of the water. There are also some small fish in the water.
(d) A close-up photograph of a painting with a variety of colors. The colors are bright and vibrant. The painting is made up of small squares of color that are all the same size.
(e) Watercolor drawing of lady with punk makeup.
(f) A teddy bear lies alone among the snow.
(g) A close-up photo of dried white flowers with brown centers, surrounded by purple and green dried flowers. The arrangement is set against a plain white wall, emphasizing texture and muted tones.
(h) A digital illustration of various food items rendered in black and white, arranged in a grid pattern on a white background.
(i) A grand palace, likely the Palace of Versailles, with gold doors, statues, columns, and a clock tower. A formal garden with trimmed hedges and colorful flowers stands in front under a clear, sunny sky.
(j) Fresh red cherries with stems in a white bowl with a blue rim on a teal surface. A few cherries and a green leaf lie outside the bowl, with a softly blurred background emphasizing freshness.
(k) A wide shot of a fast-flowing blue river with white-water rapids, surrounded by steep rocky cliffs covered in dense green trees.
(l) A close-up photo of a black and white dog looking up to the right. The dog wears a brown collar with a small bronze pendant against a dark brown background.
(m) A tree in full bloom with deep lavender lilac flowers and vibrant green leaves, set against a clear blue sky with no clouds.
(n) Composition with bread and rolls on kitchen table.
(o) These windows and those in the next photo depict scenes from the Old Testament.
(p) Brown boots with white heart designs on the toes, resting on a moss-covered ring. A blurred tree trunk and branches appear in the background.
(q) A white bowl of mashed sweet potatoes garnished with sage, black pepper, and sea salt. In the background are a gravy boat, bread, green beans, and a pumpkin.
(r) A golden retriever wearing a crown of yellow and white flowers, looking to the right, set in a green grassy field.
(s) Three pink cocktails on a circular wooden tray, each garnished with lemon slices and mint, set against a white wall and green table.
(t) A painting of a young girl in a pink dress and red coral necklace holding flowers in a lush landscape with forest and stream, conveying innocence and childhood beauty.

Table 9: Prompts used to generate images in Figure 27.

(u) A serene landscape with a snow-capped mountain beneath a clear blue sky. Tall golden grass fills the foreground, distant trees add contrast, and warm lighting suggests early morning or late afternoon.
(v) A close-up photo of a circular apple pie on a wooden surface, with apples arranged in a circular pattern and dusted with powdered sugar.
(w) A cartoon parrot wearing a blue-and-white sailor hat with an anchor, smiling while holding a guitar. The parrot has green and yellow feathers, an orange beak, blue eyes, and stands on a white background.
(x) Five bowls of chia pudding arranged in a grid on a white surface, topped with peanut butter, berries, kiwi with chocolate chips, nuts with honey, and banana with granola under bright lighting.
(y) A close-up photo of a hand holding a rolled up 20 Euro note. The background is out of focus, but there are green leaves and what appears to be a tree.
(z) A close-up photo of a small pizza topped with red sauce, cheese, and herbs on a white plate placed on a wooden table.
(aa) Fresh silver fish with pinkish tones displayed on crushed ice in a supermarket. The fish overlap slightly, with finely shaved white ice keeping them fresh and the background cropped tightly.
(ab) An abandoned brick building with a sloped roof and chimney. Three doors are boarded up, one is open, and overgrown grass and trees suggest long-term neglect.
(ac) Metallic bowls arranged in a semi-circle, each filled with vibrant powdered pigments including pink, blue, yellow, and green, photographed in close-up on a dark surface.
(ad) A blue illustration of a skeleton running. The skeleton is highlighted in orange. The background is black.
(ae) Expanded horse nostrils with rime on facial hairs
(af) Rustic pumpkin soup in a white bowl on a woven mat, garnished with black pepper. Pumpkin seeds and a whole pumpkin appear nearby, creating a warm autumnal setting.
(ag) A black and white illustration of a shaggy dog with pointed ears, wide eyes looking forward, and a slightly open mouth against a white background.
(ah) A traditional Andalusian white village in Casares, Spain, with tightly packed white buildings and terracotta roofs built along a hillside, surrounded by greenery and lacking modern infrastructure.
(ai) Watercolor botanical illustrations of a pink lotus and a rose with soft brushstrokes and detailed petals, isolated on a white background in a delicate hand-painted style.
(aj) A colorful Day of the Dead illustration featuring a skull wearing a sombrero and holding maracas, surrounded by flowers on a clean white background.
(ak) A handshake between two men in suits, one seated at a desk with papers and a pen, the other standing and smiling, with a bookshelf visible behind them.
(al) Cactus plants with pink flowers in brown pots placed on a wooden table, set against tiled flooring and a brown-and-yellow tiled wall.
(am) Stones and shells arranged in the shape of a flower on a plain white background.
(an) Vintage hand drawn rustic wreath with cute spring flowers and hand written text Happy Mother's Day

Table 10: Prompts used to generate images in Figure 28.

(ao) Skewers with charred meat, roasted golden potatoes, and roasted red tomatoes served on a white plate with a brown rim. The background is softly blurred with table elements visible.
(ap) Blossom fruit. beautiful spring.
(aq) An adult cow and calf lying on straw inside a wooden barn. The cow is brown with white patches, the calf mostly white with brown markings, lit by soft natural light.
(ar) Slow Down Tours on the train.
(as) A calm beach scene with a large wet boulder in the foreground, sandy shore, gentle blue waves, and a clear sky creating a tranquil atmosphere.
(at) A forest landscape with green, red, and yellow foliage in the foreground and a mountain covered in autumn leaves in the background.
(au) Gorgeous portrait of a blue peacock with silky blue feathers.
(av) A large rock formation covered in green plants with light blue water in front and a cloudy sky above.
(aw) A tall glass of green beverage with a purple lid on a wooden surface, surrounded by colorful signage with Chinese characters, suggesting a café or food stall setting.
(ax) An ancient stone wall with Mayan carvings depicting figures, animals, and geometric shapes, weathered and textured in light gray stone.
(ay) Abandoned barn in Sauk County, Wisconsin - LOC's Public Domain Archive Public Domain Search.
(az) Wrapped Up In Wool Penguin.
(ba) A utility pole with multiple electrical wires and a street lamp in the foreground, with a dark wooden building and partly cloudy blue sky behind it.
(bb) A cozy bedroom with a large bed, white comforter, beige walls, framed pictures, seating furniture, a rug, and a vintage chandelier with candle-style lights.
(bc) A golden porcelain Turkish coffee cup and saucer with ornate detailing and a lion-shaped handle, isolated on a reflective surface against a white background.
(bd) A person in a white dress holding a bouquet of white and pink roses with green foliage, set against a natural green background.
(be) A detailed Carcassonne Mini board game tile depicting a medieval village with church, temple, river, and trees, rendered in a realistic illustrated style.
(bf) A carved Halloween pumpkin with triangular eyes and jagged mouth, wearing a black witch hat, resting on dark leaves in a spooky setting.
(bg) A white horse standing inside a fenced area with a red-and-white striped barrier, facing the camera with alert ears and trees in the background.
(bh) A collage of contemporary furniture pieces including beds, shelving, dressers, mirrors, and chandeliers displayed across modern bedroom interiors.

Table 11: Prompts used to generate images in Figure 29.

(bi) A wide shot of a valley surrounded by snow-covered mountains under a cloudy sky, evenly lit by natural daylight.
(bj) A white lighthouse on a rocky cliff surrounded by green vegetation, overlooking the ocean beneath a clear blue sky with scattered clouds.
(bk) A creeping thistle flower with a vibrant pink center on a sepia-toned old paper background, softly lit with blank space for text.
(bl) A dimly lit dining table set with plates of food and wine glasses. Three wine bottles are visible, chairs surround the table, and a wooden wall adds a cozy mood.
(bm) Streets of Kanazawa - japanese, japan, kanazawa, street, building, oriental, lantern.
(bn) A white bowl of soup with meat, carrots, and greens on a wooden table, with salt and pepper shakers, a tomato, and a beige napkin nearby.
(bo) A close-up view of stalactites hanging from a cave ceiling, with rugged textured walls and contrasting light and shadow highlighting natural formations.
(bp) A rocky shoreline at Nha Trang beach, Vietnam, with turquoise water creating white foam, calm sea beyond, and a clear sky with scattered clouds.
(bq) A spider positioned at the center of its web, facing the camera, surrounded by green leaves with a softly blurred background and light specks.
(br) Ship wreck "Superior Producer" in turquoise water of coral reef in Caribbean Sea.
(bs) A bright cityscape with numerous beige buildings of varying roof styles, trees on the left, and clear sunny weather.
(bt) A tree with a thick trunk near a dirt path leading into a canyon filled with orange rock formations and scattered pine trees.
(bu) A cappuccino in a white porcelain cup and saucer on a wooden table, topped with a floral chocolate design and photographed with a soft blur.
(bv) A quiet winter forest with a snow-covered path winding through densely packed trees under an overcast sky.
(bw) A four-image collage showing stone buildings, a glass structure with purple light, and buildings near large bodies of water.
(bx) A collage of 36 cat photos arranged in a uniform grid, each cat looking at the camera against a bright background.
(by) A lush hillside with green grass, trees, and a large boulder in the foreground, with cloudy skies over the Bulbul hills in the background.
(bz) A panoramic cityscape showing colorful residential buildings in the foreground and an industrial area with smokestacks under a clear blue sky.
(ca) A vibrant garden with a tree full of ripe oranges near a swimming pool, featuring potted plants and an umbrella in bright light.
(cb) Different spring blossoms in a little bottle with nature background.

Table 12: Prompts used to generate images in Figure 30.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: every claim is supported by empirical evidence (companion figures, tables, and protocol). If empirical results are too weak, we do not make a strong claim, and report it as anecdotic evidence instead.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: discussed in section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: the paper does not state new theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we disclosed hyper-parameters of every experiment, as well as the methodology used for to train and evaluate (see section 3).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: we are unable to release code and data at this time. We disclose some of the public datasets we rely on, which represent more than 50% of our data mixture.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: exact hyper-parameter found by grid search are specified in Table 5, while hero-runs hyper-parameters are specified in Table 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: for hyper-parameter search, we report uncertainty bars associated to a Gaussian Process fit of grid search results. For ablations, when applicable, we report standard deviation of measurements. Due to computational costs, not all experiments could be run twice (e.g. 3B hero run). In general, we don't report error bar when we show *feasability*, but we report them during *ablations*.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: in Appendix N we detail the compute cost of the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: no ethical violations can be reported from the methodology of this paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: we discuss potential societal impact of generative models in Appendix O.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: our datasets went through an extensive legal scrutiny (including licensing) to guarantee their compatibility with the use made in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.