HOLISTIC LATENT DIFFUSION ACCELERATION: UNIFYING SPATIAL, TEMPORAL, AND ARCHITECTURAL EFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent Diffusion Models (LDM) face three compounding efficiency challenges in practical deployment: *i)* the temporal latency of iterative sampling; *ii)* the architectural overhead of heavy backbone parameter counts; and *iii)* the spatial cost of high-dimensional latent grids. While recent acceleration methods have made substantial progress on temporal distillation and architectural compression, the spatial axis is often inherited from the teacher tokenizer and treated as fixed. In this work, we recast latent resolution as an optimizable efficiency axis and introduce a unified framework that optimizes all three dimensions simultaneously. We introduce a novel strategy of Score-Compatible Tokenizer Distillation (SCTD), which leverages score-matching principles to align a spatially compact latent space with the induced distribution of a frozen, powerful teacher model, distilling the teacher's generative prior into a compressed, lower-dimensional compatible manifold. With flexibility provided by SCTD, we can surrogate a computationally heavy teacher backbone with a lightweight student architecture operating strictly within this new compressed space. Finally, we apply temporal distillation to collapse the sampling trajectory, producing a one-step generator that operates at peak efficiency. Rather than claiming quality parity, we frame our approach as an explicit efficiency–quality tradeoff: at matched one-step efficiency it occupies a Pareto-optimal operating point unreachable by single- or dual-axis methods, cutting total inference FLOPs by over $3000\times$ relative to the multi-step teacher and by $3.8\times$ relative to a one-step temporal-only baseline (at 32% lower peak memory), for a small and explicitly reported increase in FID. With reduced peak memory usage and latency, our method enables resource-constrained deployment and high-volume serving of high-fidelity LDM.

1 INTRODUCTION

The generative AI landscape is currently defined by a fundamental tension between generation fidelity and computational tractability. Latent Diffusion Models (LDMs) (Rombach et al., 2022) initially bridged this gap by moving the generative process from pixel space to the compressed coordinate system of a pre-trained visual tokenizer, enabling high-fidelity photorealistic capabilities of the state-of-the-art systems (Esser et al., 2024; Black Forest Labs, 2024; Baldridge et al., 2024). However, this progress comes with an escalating computational cost. This escalating cost is not monolithic; rather, it compounds across three interacting structural bottlenecks: *i)* Model Size: diffusion backbones have grown from ~800M to over 12B parameters to capture complex distributions, drastically increasing the cost per forward pass; *ii)* Temporal Depth: inference remains inherently iterative, requiring tens of heavy forward passes; and *iii)* Spatial Resolution: models operate on increasingly dense latent grids, incurring quadratic computational costs in attention.

While prevailing acceleration paradigms have successfully compressed the temporal axis via step distillation (Yin et al., 2024a; Zhou et al., 2024a; Kim et al., 2024a) and the architectural axis via lightweight backbones (Wang et al., 2024; Fang et al., 2023), they predominantly inherit the teacher's original visual tokenizer. Consequently, the student model remains tethered to a high-dimensional latent grid. This fixed spatial resolution acts as an unavoidable efficiency floor that dominates memory bandwidth, inflates operational costs, and precludes lightweight local execution for end users. A

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

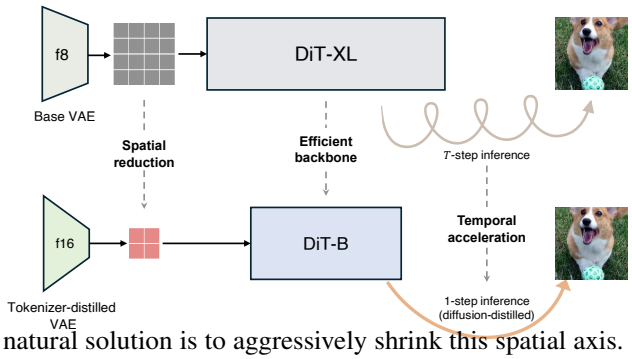

Figure 1: We tackle the overlooked third axis spatial resolution in diffusion model efficiency by distilling a more compressed, score-compatible tokenizer through SCTD, while simultaneously optimizing the temporal axis via step-distillation and the architectural axis via a more efficient diffusion backbone for holistic LDM acceleration.

natural solution is to aggressively shrink this spatial axis. However, isolating spatial compression as a pure reconstruction problem overlooks a fundamental mechanism that an image tokenizer does not merely reconstruct pixels; it defines the exact probability measure upon which the generative prior is trained. When a new compressed tokenizer, it induces a fundamentally different measure in a lower-dimensional space. Existing spatial compression methods (Chen et al., 2025c;d; He et al., 2025) optimize for pointwise reconstruction without distributional constraints; the resulting latent geometry is rarely amenable to the pre-trained generative prior for further efficiency improvement on the diffusion model, as the learned score field becomes geometrically misaligned. This creates a rigid trade-off: either retain the computationally heavy teacher tokenizer to enable score-based distillation, or achieve spatial compression but lose diffusion efficiency improvement space.

In this work, we break this lock by elevating the latent space itself to a distillable efficiency axis. Our key insight is that while score-matching distillation is architecture-agnostic, it is strictly space-dependent: the teacher's score field can only guide a student whose samples reside in a compatible probability domain. Because naive spatial compression fundamentally shifts this domain, we introduce a bootstrapping mechanism that mathematically reconciles the compressed tokenizer with the teacher's generative prior before any downstream distillation occurs. Specifically, we propose Score-Compatible Tokenizer Distillation (SCTD) that employs a lightweight, differentiable bridge to project the output of a heavily compressed spatial tokenizer into the teacher's latent space. Through this bridge, we align the noised marginals of the compressed tokenizer with those of the teacher in distribution, that are compatible with the future distillation objective. The pre-trained diffusion teacher act as a density oracle that pull the score field back through the projection bridge to directly shape the topological geometry of the new compressed tokenizer. Once this foundational geometric compatibility is established, the compressed latent space supports further efficiency gains by flexibly integrating a lightweight student backbone and apply standard temporal distillation without modification. This yields a "three-in-one" student model that simultaneously minimizes memory footprint and inference compute across the spatial, architectural, and temporal axes.

Our contribution can be summarized as follows:

(i) We identify latent resolution as the missing link in LDM acceleration and appeal to optimize it in tandem with temporal and architectural axes. To our knowledge, this is the first work to address all three dimensions, maximizing throughput by aligning an architecturally smaller LDM via diffusion-to-tokenizer distillation followed by diffusion-to-diffusion temporal distillation.

(ii) Technically, we introduce a novel distillation mechanism, SCTD, to align the implicit marginal distribution of a compressed tokenizer using the teacher's score as a geometric guide. Using the frozen teacher as a density oracle, we pull its score field back through the compressed tokenizer, shaping the student's latent geometry to be natively compatible with the generative prior.

(iii) Empirical experiments demonstrate that our method yields "all-around" efficient models. By combining a compressed latent space, a lightweight backbone, and one-step generation, our unified framework successfully aggregates the distinct efficiency advantages of individual axes, achieving high-throughput generation with minimal memory footprint and computational cost, making efficient generation deployment-ready by facilitating dynamic server load-balancing and execution on consumer-grade hardware.

2 Background and Related Work

2.1 Diffusion Models and Latent Diffusion Models

Diffusion Models: Diffusion models generate samples by learning to reverse a fixed Gaussian noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Given a clean sample

$\boldsymbol{z}_0$, the forward process progressively corrupts the data following a noise schedule $\{\alpha_t, \sigma_t\}_{t=1}^T$: $q(\boldsymbol{z}_t \mid \boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \alpha_t \boldsymbol{z}_0, \sigma_t^2 \mathbf{I})$ where $\boldsymbol{z}_T$ approaches a standard Gaussian. A noise-predicting network $\epsilon_\phi(\boldsymbol{z}_t, t)$ is trained to predict the injected noise by:

$$\mathcal{L}_{\text{Diffusion}}(\phi) = \mathbb{E}_{\boldsymbol{z}_0, t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\phi(\boldsymbol{z}_t, t)\|_2^2\right]. \tag{1}$$

Sampling starts from $\boldsymbol{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively applies the learned reverse process to recover $\boldsymbol{z}_0$, typically requiring tens to hundreds of network evaluations.

Latent Diffusion Models: Latent diffusion models (LDMs) reduce the computational cost of diffusion by moving generation from pixel space to the latent space of a pretrained visual tokenizer (Rombach et al., 2022). A tokenizer consists of an encoder-decoder pair $E : \mathcal{X} \to \mathcal{Z}$ and $D : \mathcal{Z} \to \mathcal{X}$ which maps an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ to a latent tensor $\boldsymbol{z}_0 = E(\boldsymbol{x}) \in \mathbb{R}^{h \times w \times c}$ and reconstructs $\hat{\boldsymbol{x}} = D(\boldsymbol{z}_0)$. The tokenizer is trained once and then kept fixed during diffusion training. Beyond compressing images, the tokenizer defines the probability space of latent diffusion by inducing the latent measure $p_{\mathcal{Z}} = E_{\#} p_{\text{data}}$ where $(\cdot)_{\#}$ denotes pushforward. The diffusion model is trained on noised marginals of this tokenizer-induced measure. Therefore, changing the tokenizer changes the distribution on which the diffusion model operates, leaving tokenizer compression a generative modeling problem rather than only an autoencoding problem.

2.2 Diffusion Model Efficiency

Recent efforts to accelerate diffusion models primarily act along two axes. The first axis reduces the cost of each denoising step by lowering the FLOPs or memory footprint via pruning (Yuan et al., 2022; Fang et al., 2023; Yuan et al., 2024; Kim et al., 2024b; Bolya & Hoffman, 2023), quantization (Li et al., 2023a; He et al., 2023; Wang et al., 2024; Liu et al., 2024; Zhao et al., 2025; Huang et al., 2024; Roy et al., 2025; Sui et al., 2024), activation and attention caching (Ma et al., 2024; Liu et al., 2025; Wimbauer et al., 2024), and architectural redesigns (Li et al., 2024; Fang et al., 2025; Li et al., 2023b; You et al., 2025). The second axis reduces the number of denoising steps through progressive distillation (Luhman & Luhman, 2021; Salimans & Ho, 2022), and more recently through distribution-matching distillation (Luo et al., 2023; Wang et al., 2023b; Yin et al., 2024b; Zhou et al., 2024b; Xie et al., 2024; Luo et al., 2024; Wang et al., 2025), which achieves one-step high-fidelity generation. We discuss more techinical details of such methods in Section 2.3. Due to the diffusion model's strict manifold dependence on the tokenizer, latent spatial resolution is typically treated as fixed, and efficiency improvements focus mainly on the diffusion network itself.

2.3 Preliminaries on score-based distribution matching

Score-based distribution matching has become an effective mechanism for reducing diffusion sampling to even one step. Its starting point is that a trained diffusion model provides both a score estimate and an x_0-prediction for a noised sample. For a diffused sample $\boldsymbol{z}_t$, the score of the diffused sample can be estimated by $\nabla_{\boldsymbol{z}_t} \log p_t(\boldsymbol{z}_t) \approx -\sigma_t^{-1} \epsilon_\phi(\boldsymbol{z}_t, t)$ (Song et al., 2021) while the posterior mean of the clean latent can be estimated as $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, t) = \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_t] \approx \alpha_t^{-1}(\boldsymbol{z}_t - \sigma_t \epsilon_\phi(\boldsymbol{z}_t, t))$. This motivates one-step or few-step students G_θ: by restricting the reverse process to one step, the student can be initialized or interpreted as a fast x_0-predictor.

However, this direct x_0-prediction is mean-seeking: it approximates a conditional expectation rather than the full distribution represented by the teacher reverse process. Distribution-based distillation methods therefore refine the student by matching the distribution induced by G_θ to the teacher distribution. Let $q_\theta = (G_\theta)_{\#} p(\epsilon)$ denote the student sample distribution, and let p_ϕ denote the teacher distribution. Instead of comparing clean distributions directly, these methods compare their noised marginals with a general form of objective:

$$\mathcal{L}_{\text{score-distill}}(\theta) = \mathbb{E}_{t \sim \rho(t)}\left[\mathcal{D}_t(q_{\theta,t} \| p_{\phi,t})\right], \tag{2}$$

where $\mathcal{D}_t$ is a divergence. For well-behaved divergence choices (Luo et al., 2023; Zhou et al., 2024b), the update can be expressed through score fields rather than explicit likelihoods. The teacher score is provided by the frozen teacher denoiser, *i.e.*, $\nabla \log p_{\phi,t} \approx -\sigma_t^{-1} \epsilon_\phi$. The student score $\nabla \log q_{\theta,t}$, however, is unavailable because G_θ defines an implicit distribution. These methods therefore train an auxiliary denoiser ϵ_ψ on diffused output of G_θ using the same diffusion objective in Eq. 1 on diffused output of G_θ, *i.e.*, $\mathcal{L}_{\text{aux}}(\psi; \theta) = \mathbb{E}_{\boldsymbol{z}_0 \sim q_\theta, t, \boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \epsilon_\psi(\alpha_t \boldsymbol{z}_0 + \sigma_t \boldsymbol{\epsilon}, t)\|_2^2]$ which yields $\nabla \log q_{\theta,t} \approx -\sigma_t^{-1} \epsilon_\psi$. at optimum. Training alternates between updating ψ to track the current student distribution and updating θ using the resulting teacher–student score gap.

3 METHOD

As established, a structural limitation persists in current diffusion acceleration paradigms: even when temporal distillation reduces sampling to single step, and architectural compression yields lightweight backbones, student models remain tethered to the spatial grid produced by the teacher tokenizer. This inherited latent resolution acts as a rigid computational floor. For $N = hw$ spatial tokens, halving both dimensions reduces the sequence length by $4\times$. This directly cuts convolution FLOPs by $4\times$ and shrinks Transformer self-attention compute by up to $16\times$ due to quadratic $O(N^2)$ scaling. Spatial compression reduces the baseline problem size upon which all subsequent accelerations operate.

Nevertheless, the visual tokenizer in an LDM is not merely an interchangeable preprocessing module; it defines the underlying coordinate system and the exact measure upon which the generative diffusion prior is trained. Let E_T denote the teacher tokenizer. The teacher diffusion model learns the score fields of the noised marginals of the induced latent measure, $p_T = (E_T)_{\#}p_{\text{data}}$. Naively replacing E_T with a more spatially compressed tokenizer fundamentally alters this pushforward measure. Consequently, the pre-trained teacher score field becomes geometrically misaligned and obsolete. This severs the link required for score-based distillation, rendering distribution-matching infeasible since the student and teacher no longer share a compatible distributional geometry. Therefore, a successful compressed tokenizer faces a dual mandate: it must preserve sufficient high-frequency image information for decoding, while simultaneously inducing a latent distribution natively compatible with downstream diffusion modeling.

We address this challenge via Score-Compatible Tokenizer Distillation (SCTD). By formulating tokenizer compression as a distribution-aware optimization, SCTD leverages the teacher diffusion model to provide direct score-based distributional supervision to the compressed encoder. This ensures the newly compressed latent space aligns smoothly with the pre-trained generative prior, unlocking computational savings while allowing inference within the reduced spatial dimensions.

3.1 SCORE-COMPATIBLE TOKENIZER DISTILLATION

Let p_{data} denote the image distribution. The frozen teacher tokenizer E_T induces the teacher latent measure $p_T = (E_T)_{\#}p_{\text{data}}$. Here, a compressed tokenizer $E_\eta : \mathcal{X} \to \mathcal{Z}_S$, where $\mathcal{Z}_S$ has lower spatial resolution than the teacher latent space $\mathcal{Z}_T$. Since the teacher score is defined on $\mathcal{Z}_T$, it cannot be evaluated directly on compressed latents in $\mathcal{Z}_S$. To use the teacher score as a training signal, we introduce a lightweight bridge map $h_\xi : \mathcal{Z}_S \to \mathcal{Z}_T$ and define the projected tokenizer map $T_\varphi(x) = h_\xi(E_\eta(x))$ where $\varphi = (\eta, \xi)$ to induce the projected compressed-tokenizer measure on the teacher's coordinate system:

$$q_\varphi = T_{\varphi\#}p_{\text{data}} = (h_\xi \circ E_\eta)_{\#}p_{\text{data}}. \tag{3}$$

Thus, a geometric view of SCTD is that the teacher tokenizer induces the teacher latent manifold $\mathcal{M}_T = \text{supp}(p_T) \subset \mathcal{Z}_T$, while the compressed tokenizer, after projection through h_ξ, induces the projected student manifold $\widehat{\mathcal{M}}_{S,\xi} = \text{supp}(q_\varphi) = T_\varphi(\mathcal{X}) = h_\xi(E_\eta(\mathcal{X})) \subset \mathcal{Z}_T$.. SCTD does not require pointwise equality between these coordinate charts, such as $h_\xi(E_\eta(x)) = E_T(x)$. Such a constraint would force the lower-resolution tokenizer to imitate the teacher representation sample by sample. Instead, SCTD aligns the projected distribution q_φ with p_T in the teacher score domain, while the decoder D_η ensures that the compressed code remains image-decodable.

Direct comparison of the clean measures q_φ and p_T can be ill-conditioned, especially early in training when the projected compressed latents may poorly overlap the teacher latent distribution. We therefore compare diffusion-regularized marginals:

$$q_{\theta,t} = (\alpha_t \text{Id})_{\#}q_\varphi * \mathcal{N}(0, \sigma_t^2 I), \quad p_{T,t} = (\alpha_t \text{Id})_{\#}p_T * \mathcal{N}(0, \sigma_t^2 I), \tag{4}$$

for $t \sim \rho(t)$, Equivalently, for $x \sim p_{\text{data}}$, the perturbed latent is $\hat{z} = T_\theta(x), z_t = \alpha_t \hat{z} + \sigma_t \epsilon$ for $\epsilon \sim \mathcal{N}(0, I)$ and we optimize a scored–induced discrepancy $\mathcal{D}$ over diffusion times:

$$\mathcal{L}_{\text{score}}(\varphi) = \mathbb{E}_{t\sim\rho}\left[\mathcal{D}_t(q_{\varphi,t} \| p_{T,t})\right]. \tag{5}$$

We detail the choice and implementation of $\mathcal{D}$ in Section 3.2.

For $\sigma_t > 0$, both noised marginals have smooth positive densities in the same teacher latent space. Given that Gaussian convolution is injective, *i.e.*, exact matching $q_{\varphi,t} = p_{T,t}$ at any fixed positive noise level implies matching of the clean measures $q_\varphi = p_T$. Thus, perturbation preserves the exact matching optimum while making the comparison better conditioned.

Importantly, the perturbation serves as a diffusion-aware comparison space. We align $q_{\varphi,t}$ and $p_{T,t}$ over the same timestep distribution $t \sim \rho(t)$ and noise schedule used by the teacher diffusion model to perform the distillation in the future as in Eq. 2. This produces distillation-friendly alignment across the diffusion trajectory, where score supervision is evaluated in a multi-scale setting.

Remark 3.1 (Scope of the matching guarantee). The injectivity of Gaussian convolution establishes *population-level identifiability*: the unique minimizer of the score-matching objective at any single positive noise level $\sigma_t > 0$ recovers the clean-measure equality $q_\varphi = p_T$, so averaging over $t \sim \rho(t)$ preserves this optimum while improving conditioning. It is *not* a convergence guarantee under empirical sampling, timestep mixing, or approximate scores; in the approximate regime, the update inherits the same auxiliary-critic error as all distribution-matching distillation methods (*e.g.*, DMD (Yin et al., 2024a), SiD (Zhou et al., 2024b)) rather than a weaker one.

3.2 Pulling Back Teacher-Score Guidance

We now formalize how a teacher-space score signal updates the compressed tokenizer. Unlike likelihood-based models, SCTD does not parameterize the density $q_{\varphi,t}$ explicitly. The learnable object is the projected tokenizer map $T_\varphi = h_\xi \circ E_\eta$ and changing φ moves samples through $z_t = \alpha_t T_\varphi(x) + \sigma_t \epsilon$. Therefore, we cannot optimize $\mathcal{D}_t(q_{\varphi,t} \| p_{T,t})$ by differentiating an explicit likelihood or density ratio with respect to φ. Instead, the objective must provide a sample-level direction in the teacher latent space: given a point $z_t \sim q_{\varphi,t}$, it should indicate how this point should move to reduce the discrepancy to $p_{T,t}$. Only then can the teacher-space signal be pulled back through T_φ to update the encoder and bridge.

This motivates the requirement that for a discrepancy $\mathcal{D}_t$ to be directly usable in SCTD, $\mathcal{D}_t(q_{\varphi,t} \| p_{T,t})$ to be transport-differentiable with respect to its first argument (Ambrosio et al., 2005; Santambrogio, 2015), that is, for any smooth infinitesimal perturbation $z \mapsto z + \varepsilon u(z)$, there exists a vector field g_t:

$$\frac{d}{d\varepsilon} \mathcal{D}_t \left((\mathrm{Id} + \varepsilon u)_{\#} q_{\varphi,t} \| p_{T,t} \right) \Big|_{\varepsilon=0} = \mathbb{E}_{z \sim q_{\varphi,t}} \left[g_t(z)^\top u(z) \right]. \tag{6}$$

In our setting, however, $q_{\varphi,t}$ is not available as an explicit density. What is available is the way an update of φ moves each noised sample $\Delta z_t = \alpha_t \nabla_\varphi T_\varphi(x) \Delta\varphi$.

Once the discrepancy induces a teacher-space first-variation field $g_t(z_t)$, the remaining step is to convert this desired movement in $\mathcal{Z}_T$ into an update of the tokenizer parameters, pulling back the teacher-space field through the projected tokenizer map, *i.e.*, the field $g_t(z_t)$ is composed with the Jacobian $\nabla_\varphi T_\varphi(x)$ (Lang, 2012). This yields

$$\nabla_\varphi \mathcal{D}_t = \alpha_t \mathbb{E}_{x,\epsilon} \left[g_t(z_t)^\top \nabla_\varphi T_\varphi(x) \right], \tag{7}$$

where the factor α_t can be absorbed into the timestep weight.

In implementation, g_t is evaluated at the current marginal and treated as a first-variation field. The teacher score is provided by the frozen teacher diffusion model $s_\phi(z_t, t) \approx \nabla_{z_t} \log p_{T,t}(z_t)$ while the source score $\nabla_{z_t} \log q_{\varphi,t}(z_t)$ is unavailable because q_φ is implicit. Leveraging the auxillary-score trick in Section 2.3, an auxillary score model (s_ψ is alternately trained on diffused samples from the current projected tokenizer marginal by $\mathcal{L}_{\mathrm{aux}}(\psi; \varphi) = \mathbb{E}_{x,t,\epsilon} \left[\omega_{\mathrm{aux}}(t) \| \epsilon - \epsilon_\psi(\alpha_t T_\varphi(x) + \sigma_t \epsilon, t) \|_2^2 \right]$ so that we perform score approximation by $s_\psi(z_t, t) \approx \nabla_{z_t} \log q_{\varphi,t}(z_t)$.

To implement the score update, we need to have the objective to induce a teacher-space field g_t that is both a valid first-variation direction and estimable from availbale quantities. We use two such choices: VSD-style KL divergence and SiD-style score projection for Fisher divergence. Their derivations and alternative discrepancy choices are discussed in Appendix B. In both cases, SCTD leverages the score-matching principle by treating the conditional perturbation of projected tokenizer samples $z_t \sim q(z_t \mid T_\varphi(x))$ as samples from the current source marginal $q_{\varphi,t}$.

VSD-style KL divergence: For the KL divergence, we have the approximated induced transport field of $g_t^{\mathrm{KL}}(z_t) \approx s_\psi(z_t, t) - s_\phi(z_t, t)$ where its variational field can be written purely using scores. Using stop-gradient surrogate whose gradient matches the desired pullback field (Luo et al., 2023; Yin et al., 2024b), we have

$$\mathcal{L}_{\text{SCTD-KL}}(\varphi) = \mathbb{E}_{x,t,\epsilon} \left[\omega(t) \, \mathrm{sg} \left[s_\psi(z_t, t) - s_\phi(z_t, t) \right]^\top z_t \right]. \tag{8}$$

SiD-style Fisher divergence: For the Fisher realization, we use score-projection identities (Zhou et al., 2024b; Luo et al., 2024) to obtain a score-implicit update and obtain a pullback-realizable surrogate direction based on the Fisher score gap. We define the stopped score-gap direction $g_t^{\mathrm{SiD}}(z_t) \approx$

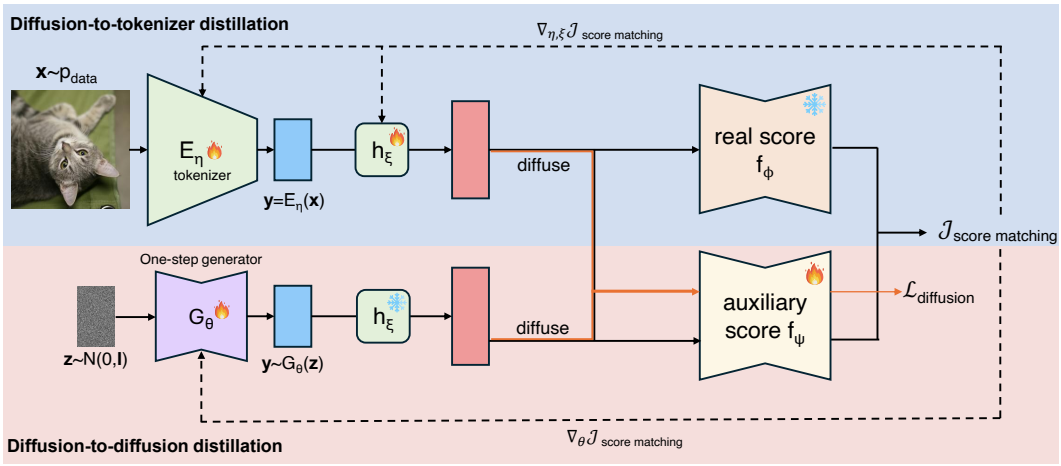

Figure 2: **Top: Diffusion-to-Tokenizer Distillation**: We train a compact encoder E_η and a projector h_ξ to define a latent space compatible with the teacher's frozen, diffused probability flow. The score matching gradient $\nabla_{\eta,\xi}\mathcal{J}$ minimizes the divergence between the projected encoder aggregated posterior and the teacher's prior f_ϕ. **Bottom: Diffusion-to-Diffusion Distillation**: with the projector h_ξ frozen, we distill a one-step generator G_θ on a lightweight network that generate by mapping noise $\mathbf{z}$ directly to data. The generator is optimized via score distillation $\nabla_\theta\mathcal{J}$ to match the teacher's distribution in the projected space. In both stages, the auxiliary score network f_ψ acts as a critic to estimate the score of the distribution defined by the current parameters, enabling tractable gradient computation.

$s_\phi(z_t,t) - s_\psi(z_t,t)$ and we have the known conditional perturbation score $s_{\text{cond}}(z_t, T_\varphi(x), t) = \nabla_{z_t} \log q(z_t \mid T_\varphi(x)) = -\sigma_t^{-2}(z_t - \alpha_t T_\varphi(x))$. The SCTD-Fisher update is implemented as:

$$\mathcal{L}_{\text{SCTD-Fisher}}(\varphi) = \mathbb{E}_{x,t,\epsilon}\left[\omega(t)\,\text{sg}\left[s_\phi(z_t,t) - s_\psi(z_t,t)\right]^\top \left(s_\phi(z_t,t) - s_{\text{cond}}(z_t, T_\varphi(x), t)\right)\right]. \quad (9)$$

With the SCTD update defined, we train the compressed tokenizer under the usual tokenizer constraint: the learned code must remain decodable. Equivalently, SCTD can be viewed as a reconstruction-constrained distributional alignment problem. With , we optimize the coupling objective:

$$\mathcal{L}_{\text{SCTD-tokenizer}}(\eta,\xi) = \mathcal{L}_{\text{VAE}}(\eta) + \lambda_{\text{SCTD}}\mathcal{L}_{\text{score}}(\eta,\xi), \quad (10)$$

where $\mathcal{L}_{\text{VAE}}$ is the tokenizer loss used for LDM tokenizer training (Rombach et al., 2022).

3.3 Training the Student Generator and Unifying Acceleration Axes

With the training done, E_η, D_η define a spatially compressed latent coordinate system. As generation is performed in $\mathcal{Z}_S$. We first warm-up train a student diffusion model with parameters θ directly on compressed latents $z_S = E_\eta(x)$ using the standard denoising objective in Eq. 1. Unlike same-space distillation, where a student can often inherit teacher weights, in our cases, due to the measure difference, a warm up is needed. Given that we have the chance not to inhereit, here, the backbone of the student can be heterogenous choice that need not to be identical to the teacher.

With the connected field that set up favorable conditional over timestep, We then apply temporal distillation in the compressed space. To compare the with the teacher, we reuse the frozen bridge h_ξ. Let $y_0 = h_\xi(z_{S,0})$, $y_t = \alpha_t y_0 + \sigma_t\epsilon$ and $p^h_{\theta,t}$ be the distribution of y_t. The projected student distribution is distilled toward the teacher by matching noised marginals:

$$\mathcal{L}_{\text{distill}}(\theta) = \mathbb{E}_{t\sim\rho(t)}\left[\mathcal{D}_t(p^h_{\theta,t} \| p_{T,t})\right], \quad (11)$$

where $\mathcal{D}$ follows standard score distillation gradients to the specific choice of $\mathcal{D}$ (Zhou et al., 2024b; Wang et al., 2023b). This objective mirrors the tokenizer-stage score alignment, but now the distribution-defining map is the student generator rather than the encoder. The auxiliary score model s_ψ follows the standard distribution-matching critic loop (Yin et al., 2024a; Zhou et al., 2024b) (alternating updates with EMA and a high-capacity critic), so its stability and convergence behavior match those established methods rather than introducing a new failure mode. Because the tokenizer-stage critic is fit on real-data posteriors while the generator subsequently drifts, we additionally include a brief calibration step that freezes θ and pre-fits s_ψ before enabling joint updates, intended to prevent mode collapse from generator drift.

The final model combines three acceleration axes. SCTD reduces the spatial latent grid; the student backbone provides architectural compression; and Eq. (11) enables few-step or one-step temporal distillation. At inference, the teacher model, auxiliary score model, and bridge h_ξ are discarded. The

generator samples directly in $\mathcal{Z}_S$ and decodes with D_η, so the deployed model does not inherit the teacher's dense latent representation.

4 EXPERIMENT

Our experiments engage in comparisons to assess stable and effective diffusion training and distillation at reduced latent resolution, and the holistic efficiency gains by jointly improving latency, FLOPs, and memory usage while preserving generation quality. We conduct primary experiments on the ImageNet dataset (Deng et al., 2009). We follow standard data preparation procedures, normalizing images and applying random horizontal flips during training.

We evaluate based on static complexity and sampling efficiency, alongside standard quality metrics. For deployment efficiency, we report the parameter count of the backbone and FLOPs per denoising step. Crucially, we calculate total inference FLOPs to highlight the compound reductions in both resolution and sampling steps. We further validate practical accessibility via wall-clock latency and peak memory usage averaged over three trials of inference. We quantitatively assess the quality of our generated samples using Fréchet Inception Distance (FID) (Heusel et al., 2017), which measures the perceptual similarity between the distribution of 50k generated samples and the distribution of real images, the Inception Score (IS) (Salimans et al., 2016) to evaluate image fidelity and diversity.

We adopt DMD (Yin et al., 2024a) and SiD (Zhou et al., 2024b) which represent the current standard for single-step generation that optimizes the temporal axis. In addition, we provide evaluations using progressive distillation (Salimans & Ho, 2022) and consistency model (Song et al., 2023). For architectural efficiency, rather than relying on rigid pruning heuristics that require structural isomorphism, we employ a heterogeneous surrogate strategy: we replace the heavy teacher backbone with a standard, parameter-efficient architecture trained from scratch. For the spatial axis, we use DC-Gen (He et al., 2025) to compress the grid size. We re-implement DC-Gen *only* as a spatial-compression tokenizer baseline under our matched protocol (identical DiT-XL teacher, ImageNet, and compute budget), so that the tokenizer is the sole variable; consequently its numbers here are not directly comparable to those reported in the original DC-Gen setup, which uses a different, larger, multi-step model at higher resolution.

We evaluate on ImageNet Deng et al. (2009) at 256 and 512 resolution with standard preprocessing (normalization and random horizontal flips). We report generation quality with FID Heusel et al. (2017) and IS Salimans et al. (2016) over 50k samples, and deployment efficiency with backbone parameter count, FLOPs per denoising step, *total* inference FLOPs (which expose the compound resolution×step savings), wall-clock latency, and peak memory averaged over three inference trials.

We instantiate the three axes with standard components so that only the axis under study varies. **Temporal:** DMD Yin et al. (2024a) and SiD Zhou et al. (2024b) (with progressive distillation Salimans & Ho (2022) and consistency models Song et al. (2023) for reference). **Architectural:** a heterogeneous surrogate that replaces the heavy teacher with a smaller backbone trained from scratch, rather than pruning to a layer-matched student. **Spatial:** DC-Gen He et al. (2025). We re-implement DC-Gen *only* as a spatial-compression tokenizer baseline under our matched protocol (identical DiT-XL teacher, ImageNet, and compute budget), so the tokenizer is the sole variable; its numbers here are therefore not comparable to the original DC-Gen setup, which uses a larger, multi-step model.

The teacher is a standard DiT-XL/2 and the bridge h_ξ is a lightweight upsample-ResBlock-conv1×1 projector; all pipelines are implemented in JAX Bradbury et al. (2018) on TPU-v5p (training) and TPU-v5e (inference). Settings of the tokenizer, backbone, guidance, and distillation hyperparameters are given in Appendix C.

4.1 RESULTS

Table 1 reports the efficiency–quality tradeoff on ImageNet 256×256 across spatial, architectural, and temporal axes. The single-axis baselines reveal that each acceleration strategy alone reaches a distinct bottleneck that temporal distillation reduces the number of function evaluations from 1000 to 1, bringing latency down, but still inherits the teacher's DiT-XL backbone and 32×32 latent grid, so their per-step FLOPs and peak memory remain tied to the original configuration. Architectural compression reduces parameters and memory by replacing DiT-XL with DiT-L or DiT-B, but with 1000 denoising steps the total inference cost remains high. Spatial compression

Table 1: We compare methods that optimize diffusion efficiency along individual axes (temporal, Arch., or latent resolution) against our unified approach on ImageNet 256×256. Columns are grouped by static model complexity and sampling-time efficiency and quality.

		Static Model Complexity				Sampling Efficiency & Quality					
Optimization	**Model**	Latent Res.	Backbone	Params	FLOPs / step	NFE	Total FLOPs	Latency (ms)	Mem (GB)	FID_{50k}	IS
Reference: No Acceleration											
Baseline	Teacher	$\mathcal{Z}_T$ (32 × 32)	DiT-XL	461M	5.5e11	1000	5.5e14	2600	2.2	4.41	176
Single-Axis Acceleration											
Temporal	Teacher + step distill (DMD)	$\mathcal{Z}_T$ (32 × 32)	DiT-XL	461M	5.5e11	1	6.4e11	4.9	2.2	5.08	141
Temporal	Teacher + step distill (SiD)	$\mathcal{Z}_T$ (32 × 32)	DiT-XL	461M	5.5e11	1	6.4e11	4.9	2.2	5.09	149
Arch.	Student	$\mathcal{Z}_T$ (32 × 32)	DiT-L	314M	3.6e11	1000	3.6e14	2350	1.5	5.73	169
Arch.	Student	$\mathcal{Z}_T$ (32 × 32)	DiT-B	130M	1.7e11	1000	1.7e14	2250	1.1	7.41	132
Spatial	Teacher-size + DC-Gen tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.3e11	1000	1.3e14	2450	1.7	5.60	168
Spatial	Teacher-size + SCTD-KL tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.3e11	1000	1.3e14	2450	1.7	5.42	132
Spatial	Teacher-size + SCTD-Fisher tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.3e11	1000	1.3e14	2450	1.7	5.55	144
Spatial	Teacher-size + SCTD-KL tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (8 × 8)	DiT-XL	461M	3.8e10	1000	3.8e13	2400	1.4	6.70	118
Spatial	Teacher-size + SCTD-Fisher tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (8 × 8)	DiT-XL	461M	3.8e10	1000	3.8e13	2400	1.4	7.03	114
Two-Axis Acceleration											
Spatial + Temporal	DC-Gen tok. + step distill	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.3e11	1	–	–	–	*not directly applicable*	
Spatial + Temporal	Teacher-size + SCTD-KL tok. + DMD	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.4e11	1	2.5e11	5.3	1.7	6.02	124
Spatial + Temporal	Teacher-size + SCTD-Fisher tok. + SiD	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-XL	461M	1.3e11	1	2.5e11	5.7	1.7	6.22	128
Spatial + Arch.	Student (de novo) + DC-Gen tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-L	314M	8.1e10	1000	8.1e13	2350	1.5	25.31	45
Spatial + Arch.	Student + SCTD-KL tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-L	314M	8.1e10	1000	8.1e13	2350	1.5	5.89	133
Spatial + Arch.	Student + SCTD-Fisher tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-L	314M	8.1e10	1000	8.1e13	2350	1.5	5.78	136
Spatial + Arch.	Student + SCTD-Fisher tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (16 × 16)	DiT-B	130M	3.9e10	1000	3.9e13	2250	1.0	6.35	119
Spatial + Arch.	Student + SCTD-Fisher tok.	$\mathcal{Z}_T \to \mathcal{Z}_S$ (8 × 8)	DiT-B	130M	1.2e10	1000	1.2e13	2200	0.9	7.20	101
Three-Axis Acceleration (Ours)											
Spatial + Arch. + Temporal	Student + SCTD-KL tok. + DMD	$\mathcal{Z}_\mathbf{T} \to \mathcal{Z}_\mathbf{S}$ **(16 × 16)**	DiT-L	314M	8.1e10	1	1.7e11	3.5	1.5	6.56	101
Spatial + Arch. + Temporal	Student + SCTD-Fisher tok. + SiD	$\mathcal{Z}_\mathbf{T} \to \mathcal{Z}_\mathbf{S}$ **(16 × 16)**	DiT-L	314M	8.8e10	1	1.7e11	3.5	1.5	6.41	98
Spatial + Arch. + Temporal	Student + SCTD-Fisher tok. + SiD	$\mathcal{Z}_\mathbf{T} \to \mathcal{Z}_\mathbf{S}$ **(16 × 16)**	DiT-B	130M	3.9e10	1	8.2e10	3.2	1.1	7.35	91
Spatial + Arch. + Temporal	Student + SCTD-Fisher tok. + SiD	$\mathcal{Z}_\mathbf{T} \to \mathcal{Z}_\mathbf{S}$ **(8 × 8)**	DiT-B	130M	1.2e10	1	1.6e10	3.0	0.9	8.40	78

areduces the denoising problem size by shrinking the latent grid; however, when used alone, latency is still dominated by the large number of sampling steps.

The table also separates tokenizer compression from score-compatible tokenizer distillation. Existing compressed-tokenizer baselines such as DC-Gen, are effective for single-axis spatial reduction, but they are not directly designed to support the subsequent score-based temporal distillation stage nor provide a meaningful space for *de novo* training of another diffusion model from scratch. In contrast, SCTD trains the compressed tokenizer to remain aligned with the teacher score domain. At the same 16×16 latent resolution and DiT-L backbone, and under an *identical* de novo training protocol and budget, replacing the DC-Gen tokenizer with the SCTD-guided tokenizer improves downstream generation from FID 25.31 to 5.78. Because the training protocol is held fixed, this $4.4\times$ gap isolates score-compatibility as the causal factor. It shows that the tokenizer must be optimized, not only for reconstruction but also for diffusion compatibility.

Table 2: Comparisons of quality and efficiency on Imagenet 512×512 resolution.

Model	**FLOPs**	**Mem(GB)**	**FID**	**IS**
32×32 *teacher latent*				
Base DiT-XL, 1000 steps	5.5×10^{14}	2.2	3.57	239
Teacher + SiD, 1 step	6.4×10^{11}	2.2	4.18	218
SCTD-KL + DMD, DiT-L	4.4×10^{11}	1.9	5.05	185
SCTD-Fisher + SiD, DiT-L	4.4×10^{11}	1.9	4.90	190
SCTD-Fisher + SiD, DiT-B	3.6×10^{11}	1.3	5.80	177
64×64 *teacher latent*				
Base DiT-XL, 1000 steps	5.5×10^{14}	2.2	3.19	235
Teacher + SiD, 1 step	6.4×10^{11}	2.2	4.48	207
SCTD-KL + DMD, DiT-L	4.5×10^{11}	1.5	6.35	105
SCTD-Fisher + SiD, DiT-L	4.5×10^{11}	1.5	6.02	110
SCTD-Fisher + SiD, DiT-B	2.4×10^{11}	1.3	7.11	99

Combining all three axes yields the best deployment-oriented tradeoff. By simultaneously reducing latent resolution ($32 \times 32 \to 16 \times 16$), sampling steps ($1000 \to 1$), and architectural overhead (DiT-XL → DiT-L), our unified approach breaks the efficiency ceilings of single-axis methods. The method achieves the lowest inference latency, total FLOPs, outperforming the temporal-only baselines. For the memory footprint, by decoupling the grid resolution and using a more efficient backbone, we reduce peak memory usage significantly, making the model more deployable on lower-end consumer devices, at a bounded, explicitly reported quality cost (FID 6.41 for the one-step DiT-L 16×16 model). More aggressive variants, such as DiT-B or 8×8 latents, further reduce FLOPs and memory but incur some quality degradation. We will further examine qualitatively the degree of degradation in the later part of this section. This is an explicit efficiency–quality tradeoff rather than quality parity: at matched one-step efficiency our model reaches a Pareto-optimal operating point (1.7×10^{11} total FLOPs, 1.5 GB, FID 6.41) that no single- or dual-axis method attains—it is $3.8\times$ cheaper in FLOPs and 32% lower in memory than a one-step temporal-only baseline (FID 5.08), and over $3000\times$ cheaper in FLOPs than the multi-step teacher (FID 4.41).

Table 2 evaluates whether the same efficiency pattern holds at ImageNet 512×512. Across both teacher latent resolutions, SCTD substantially reduces total sampling FLOPs by combining spatial compression with one-step generation. Two common tokenizer of $f8$ and $f16$ is used. For the 32×32 teacher-latent setting ($f16$), the SCTD-Fisher + SiD DiT-L student reduces total FLOPs and peak memory at a quantified FID cost (e.g., $3.57 \to 4.90$ relative to the multi-step teacher), consistent with the efficiency–quality tradeoff above. A similar trend is observed on 64×64 setting ($f8$) where we notice that using a one-step student yields large compute savings under less aggressive compression.

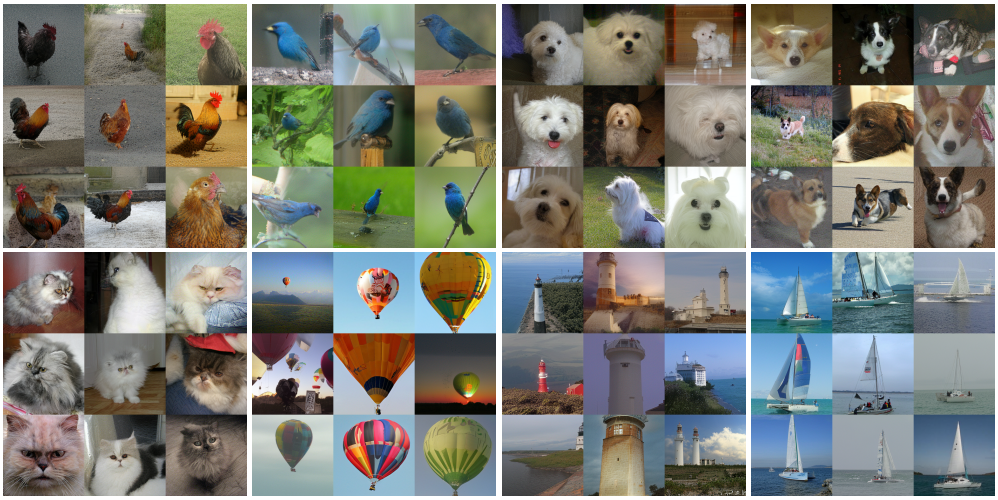

Figure 3: ImageNet samples conditionally generated in a single step by DiT-B on the SCTD-guided 8×8 latent space (the most aggressive, lowest-fidelity operating point in Table 1). Despite the simultaneous three-way holistic acceleration, samples remain semantically coherent with plausible detail.

We qualitatively assess the generation quality of our method in Fig. 3. We display random, uncurated samples from our one-step SiD-distilled student on the spatially compressed latent space. The samples are semantically coherent with plausible high-frequency detail across a diverse range of ImageNet classes; consistent with the quantitative tradeoff, they do not reach the fidelity of the multi-step teacher, but they demonstrate that generative capability transfers across mismatched latent manifolds even under aggressive spatial and temporal compression.

4.2 Further Empirical Study

We ablate the architecture of the training-time bridge h_ξ in Table 3. We report both reconstruction FID (rFID) and generation FID (gFID) (Yao et al., 2025), which evaluates the quality of a diffusion model warmed up on the resulting compressed latent space before temporal distillation. Like with prior observations that reconstruction quality and generative quality are poorly correlated (Yao et al., 2025; Xu et al., 2026), the bridge with the best rFID does not yield the best gFID. We use the lightweight upsample–ResBlock–1×1 conv bridge as our default, as it achieves the best gFID. Additional empirical studies are presented in Appendix D.

Table 3: Architecture choice of h. All models use 32×32 → 16×16 SCTD on ImageNet 256×256. We use SCTD-KL in the experiments.

Bridge h_ξ	**rFID** ↓	**gFID** ↓
Upsample + 1×1 Conv	0.83	5.74
MLP projector	0.72	6.00
One-layer Transformer	**0.60**	6.67
Upsample + ResBlock + 1×1 Conv	0.73	**5.60**

5 Limitation and Conclusion

Limitations and Future Work. SCTD uses image samples $x \sim p_{\text{data}}$ to instantiate the projected tokenizer distribution $(h_\xi \circ E_\eta)_{\#}p_{\text{data}}$. This poses a practical limitation in deployment scenarios where a teacher model's weights are open-source, but its massive original training corpus remains proprietary or inaccessible, though synthetically generated samples can provide close approximation. Furthermore, our empirical study focuses on ImageNet generation, which allows us to isolate the spatial, architectural, and temporal acceleration axes under a standard benchmark. Because open-vocabulary text-to-image (T2I) pipelines incur immense inference costs, they stand to reap the greatest benefits from SCTD's spatial compression. While scaling this framework to such massive systems entails substantially larger training corpora and infrastructural compute budgets, it represents a highly promising, natural avenue for future application of the holistic acceleration.

Conclusion. In this work, we introduced SCTD, a framework for jointly addressing the temporal, architectural, and spatial efficiency axes of latent diffusion models. By pulling teacher score guidance back through a compressed tokenizer, SCTD elevates spatial compression from an isolated autoencoding task into a distribution-aware alignment with the generative prior. This geometrically compatible coordinate system seamlessly integrates with lightweight student backbones and few-step distillation, yielding a "three-in-one" acceleration pipeline. By shrinking the latent computational floor, SCTD reduces FLOPs and memory footprint, enabling more cost-effective generation for both high-volume serving and resource-constrained devices. While democratizing the usage, easier local generation also lowers the barrier to misuse, increasing the need for safeguards such as watermarking and on-device content filtering to support responsible deployment.

AI USE STATEMENT

We used generative AI tools only for language editing and polishing of author-written text. We have reviewed all AI-assisted work. All technical content, derivations, experiments, and analyses were produced by the authors.

ETHICS STATEMENT

SCTD is, at its core, an accessibility contribution. State-of-the-art generation increasingly underpins research, education, and creative work, yet its cost concentrates access in a handful of well-resourced labs and cloud providers. By collapsing the spatial, architectural, and temporal cost of latent diffusion at once, SCTD moves high-fidelity generation off data-center clusters and onto commodity and consumer hardware, with three concrete community benefits: (i) it lowers the barrier for students, independent researchers, and institutions in resource-constrained settings to study and build on modern generative models; (ii) it enables privacy-preserving, on-device generation that keeps user data local rather than routed to third-party APIs; and (iii) it cuts the energy and carbon footprint of large-scale serving, since most demand can be met by lightweight distilled models with heavy compute reserved only when necessary. We view efficiency not as a mere engineering convenience but as a prerequisite for the equitable and sustainable deployment of generative AI. We also recognize that cheaper local generation can lower the barrier to misuse by bypassing the safeguards of centralized services, and we encourage releasing efficient on-device generators together with lightweight safeguards (e.g., watermarking and on-device content filtering) so that broadened access does not come at the expense of responsible use.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models. *arXiv preprint arXiv:2510.18457*, 2025.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPR*, 2023.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. JAX: composable transformations of python+ numpy programs, 2018. URL http://github.com/google/jax.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In *ICLR*, 2026.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *ICML*, 2025a.

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. SoftVQ-VAE: Efficient 1-dimensional continuous tokenizer. In *CVPR*, 2025b.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *ICLR*, 2025c.

Junyu Chen, Dongyun Zou, Wenkun He, Junsong Chen, Enze Xie, Song Han, and Han Cai. DC-AE 1.5: Accelerating diffusion model convergence with structured latent space. In *ICCV*, 2025d.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *NeurIPS*, 2023.

Jiarui Fang, Jinzhe Pan, Aoyu Li, Xibo Sun, and WANG Jiannan. PipeFusion: Patch-level pipeline parallelism for diffusion transformers inference. In *NeurIPS*, 2025.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *arXiv preprint arXiv:2512.07829*, 2025.

Ming Gui, Johannes Schusterbauer, Timy Phan, Felix Krause, Josh Susskind, Miguel Angel Bautista, and Björn Ommer. Adapting self-supervised representations as a latent space for efficient generation. In *ICLR*, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Wenkun He, Yuchao Gu, Junyu Chen, Dongyun Zou, Yujun Lin, Zhekai Zhang, Haocheng Xi, Muyang Li, Ligeng Zhu, Jincheng Yu, et al. Dc-Gen: Post-training diffusion acceleration with deeply compressed latent space. *arXiv preprint arXiv:2509.25180*, 2025.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. PTQD: accurate post-training quantization for diffusion models. In *NeurIPS*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. TFMQ-DM: temporal feature maintenance quantization for diffusion models. In *CVPR*, 2024.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *ICLR*, 2024a.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In *ICCV*, 2025.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *WACV*, 2024b.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In *NeurIPS*, 2021.

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *NeurIPS*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

Serge Lang. *Fundamentals of differential geometry*. Springer Science & Business Media, 2012.

Junho Lee, Jeongwoo Shin, Hyungwook Choi, and Joonseok Lee. Latent diffusion models with masked autoencoders. In *ICCV*, 2025.

Jianze Li, Yong Guo, Yulun Zhang, and Xiaokang Yang. Asymmetric VAE for one-step video super-resolution acceleration. *arXiv preprint arXiv:2509.24142*, 2025a.

Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Kai Li, and Song Han. DistriFusion: Distributed parallel inference for high-resolution diffusion models. In *CVPR*, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-Diffusion: Quantizing diffusion models. In *ICCV*, 2023a.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In *NeurIPS*, 2023b.

Yizhuo Li, Yuying Ge, Yixiao Ge, Ying Shan, and Ping Luo. Aligning latent spaces with flow priors. *arXiv preprint arXiv:2506.05240*, 2025b.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion through temporal attention decomposition. *Trans. Mach. Learn. Res.*, 2025.

Xuewen Liu, Zhikai Li, Junrui Xiao, Mengjuan Chen, Jianquan Li, and Qingyi Gu. EDA-DM: Enhanced distribution alignment for post-training quantization of diffusion models. *arXiv preprint arXiv:2401.04585*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *NeurIPS*, 2023.

Weijian Luo, Zemin Huang, Zhengyang Geng, J. Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. In *NeurIPS*, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free. In *CVPR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Kai Qiu, Xiang Li, Jason Kuen, Hao Chen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Robust latent matters: Boosting image generation with sampling error synthesis. *arXiv preprint arXiv:2503.08354*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Arani Roy, Shristi Das Biswas, and Kaushik Roy. Slimdiff: Training-free, activation-guided hands-free slimming of diffusion models. *arXiv preprint arXiv:2509.21498*, 2025.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. LiteVAE: Lightweight and efficient variational autoencoders for latent diffusion models. In *NeurIPS*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer, 2015.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Yang Sui, Yanyu Li, Anil Kag, Yerlan Idelbayev, Junli Cao, Ju Hu, Dhritiman Sagar, Bo Yuan, Sergey Tulyakov, and Jian Ren. BitsFusion: 1.99 bits weight quantization of diffusion model. In *NeurIPS*, 2024.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. In *CVPR*, 2024.

Yifei Wang, Weimin Bai, colin zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-Instruct: One-step diffusion model through unified diffusion divergence instruction. In *NeurIPS*, 2025.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training gans with diffusion. In *ICLR*, 2023a.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *NeurIPS*, 2023b.

Felix Wimbauer, Bichen Wu, Edgar Schönfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam S. Tsai, Jonas Kohler, Christian Rupprecht, Daniel Cremers, Peter Vajda, and Jialiang Wang. Cache me if you can: Accelerating diffusion models through block caching. In *CVPR*, 2024.

Sirui Xie, Zhisheng Xiao, Diederik P. Kingma, Tingbo Hou, Ying Nian Wu, Kevin P. Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. EM distillation for one-step diffusion models. In *NeurIPS*, 2024.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Haotian Zhang, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *arXiv preprint arXiv:2603.05630*, 2026.

Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. In *ICLR*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Haoran You, Connelly Barnes, Yuqian Zhou, Yan Kang, Zhenbang Du, Wei Zhou, Lingzhi Zhang, Yotam Nitzan, Xiaoyang Liu, Zhe Lin, Eli Shechtman, Sohrab Amirghodsi, and Yingyan Celine Lin. Layer- and timestep-adaptive differentiable token compression ratios for efficient diffusion transformers. In *CVPR*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025.

Geng Yuan, Yanyu Li, Sheng Li, Zhenglun Kong, Sergey Tulyakov, Xulong Tang, Yanzhi Wang, and Jian Ren. Layer freezing & data sieving: Missing pieces of a generic framework for sparse training. In *NeurIPS*, 2022.

Zhihang Yuan, Hanling Zhang, Lu Pu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. DiTFastAttn: Attention compression for diffusion transformer models. In *NeurIPS*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Rui Wan, Widyadewi Soedarmadji, Enshu Liu, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. ViDiT-Q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *ICLR*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *ICLR*, 2024a.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *ICML*, 2024b.