DREAMAR: KNOWLEDGE DISTILLATION SAMPLING USING NEXT SCALE AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce DreamAR, an optimization framework that turns next-scale autoregressive (AR) models into priors for 2D and 3D representations, serving as the discrete AR counterpart to Score Distillation Sampling (SDS). While next-scale AR models rival diffusion in image fidelity and scale favorably, they cannot directly drive SDS, which relies on noise perturbation. Instead, these models operate across discrete multi-scale token maps, predicting categorical distributions over codebook indices in a single forward pass. To leverage this, we propose Knowledge Distillation Sampling (KDS), an objective that minimizes the cross-entropy between the model's predictive distribution and a differentiable tokenization of the rendered asset. Optimization progresses from coarse to fine, advancing to each scale after optimizing the preceding ones, with the scale index serving as the discrete analogue to diffusion timesteps. Because the prior's prediction is fully determined by the rendered view, the prompt and the current scale, KDS supervision involves no sampled noise and avoids the gradient variance that diffusion distillation needs dedicated machinery to reduce. Evaluated on text-to-2D generation (MJHQ-30K) using the Infinity-2B backbone, KDS achieves 24.24 FID, outperforming diffusion distillation baselines (36.55 for the strongest), and even surpasses distillation from SD3.5-Medium despite its lower native generation FID. For both text-to-3D generation and TRELLIS.2-based refinement, KDS performs on par with leading distillation sampling methods CFD and RFDS-Rev across appearance and geometry metrics.



Figure 1: DreamAR introduces a novel distillation sampling method leveraging a text-to-image next-scale autoregressive model. 2D Images (top) and 3D objects (bottom) generated with DreamAR.

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.

1 INTRODUCTION

Score Distillation Sampling (SDS) (Poole et al., 2023) marked a breakthrough in 3D content creation. A frozen text-to-image diffusion model supervises a 3D representation through differentiable rendering, enabling 3D asset generation using only a pretrained 2D prior. The same recipe carries to video, motion and editing (Ling et al., 2024; Bahmani et al., 2024b; Bekor et al., 2025; Li et al., 2024b; Haque et al., 2023; Gilo & Litany, 2026). Standard distillation sampling formulation is however tied to diffusion. The injected Gaussian noise at each step is what enables querying the diffusion model's estimated score, which then guides the optimization.

Next-scale autoregressive (AR) image models (Tian et al., 2024; Han et al., 2025; Voronov et al., 2024; Liu et al., 2025) have emerged as a promising generative class, with compelling scaling behavior and fidelity competitive with state-of-the-art diffusion models in text-to-image synthesis (Table 1). Despite their expressive power and computational advantages, no mechanism exists to harness these priors for 3D generation, since diffusion-centric distillation does not transfer to a discrete, multi-scale token space.

In this work we introduce *DreamAR*, the first framework to distill the generative prior of a next-scale AR model into an optimized continuous representation. Rather than estimating a score, these architectures tokenize images into a hierarchy of discrete maps at increasing spatial resolution, generating them coarse to fine, conditioned on preceding scales and a text prompt. A single forward pass yields a categorical distribution over the entire codebook vocabulary for every spatial token of the next scale. Adapting distillation sampling to this score-free setting replaces the continuous diffusion timestep with the discrete scale index and the estimated score with the model's predicted categorical distributions. However, this is not a direct substitution and requires carefully formulating the loss and gradient flow through the discrete token space.

We first examine a direct approach to adapting distillation sampling, building on recent image editing methods (El-Ghoussani et al., 2026). By truncating and regenerating the token hierarchy, we construct pseudo-targets and explore both stochastic and deterministic sampling. While deterministic sampling performs better, this approach remains limited, as it discards information contained in the full predicted distributions.

To address this, we introduce Knowledge Distillation Sampling (KDS), based on the original knowledge distillation objective (Hinton et al., 2015). Rather than sampling individual tokens, KDS minimizes the cross-entropy between the model's full predicted categorical distributions and differentiable probabilities induced by the host. Our formulation makes minimal assumptions about the codebook and applies across different next-scale architectures. We further introduce progressive scale optimization, lower-scale supervision, and frozen-host regularization to stabilize optimization.

Crucially, next-scale prediction models offers a natural solution to a persistent challenge in diffusion distillation sampling. Diffusion methods sample noise at each iteration, so views rendered from the same camera receive different gradients across steps, and variance-reduction machinery was built to contain it (Tang et al., 2023; Wang et al., 2023; Alldieck et al., 2024; McAllister et al., 2024; Yan et al., 2025). In contrast, next-scale predictions are fully determined by the render and its preceding scales, enabling deterministic supervision natively.

To summarize, our primary contributions are:

- We introduce DreamAR, the first framework to distill next-scale AR priors into optimized representations, opening the applications built on score distillation to a family of priors that had no route to them. The formulation makes minimal assumptions about the codebook and works with both bitwise (Infinity) and conventional (Switti) codebooks.
- We introduce KDS, a stable distillation objective that leverages the teacher's full token distribution through cross-entropy and differentiable soft quantization. Our analyses show that the chosen cross-entropy direction avoids mode seeking and outperforms both stochastic and deterministic sampling-based alternatives. A progressive scale schedule, lower-scale supervision, and frozen-host regularization further stabilize optimization.
- We demonstrate state-of-the-art text-to-2D performance among distillation sampling methods, with 24.24 FID and 5.20 CLIP-FID against 36.55 and 12.35 for CFD, the strongest

baseline. In 3D asset refinement, KDS performs on par with CFD and RFDS-Rev across appearance and geometry metrics.

2 Related Work

2.1 Visual Autoregressive Models

Autoregressive image generation has been extensively studied (Sun et al., 2024; Yu et al., 2025; Li et al., 2024c; Fan et al., 2025; Han et al., 2026). More recently, next-scale prediction approaches (Tian et al., 2024; Han et al., 2025; Yang et al., 2025b; Voronov et al., 2024; Liu et al., 2025) have shown promising results, with favorable scaling laws observed.

Notably, image editing methods based on next-scale prediction have shown strong performance (Wang et al., 2025b; El-Ghoussani et al., 2026; Zhang et al., 2026; Wang et al., 2025a). These methods avoid image inversion by directly tokenizing the input across scales, truncating the token hierarchy, and regenerating subsequent scales using different strategies. Our PT approach (Section 3.2) adapts this truncate-and-regenerate step to distillation sampling.

2.2 Text-to-3D Generation

Leveraging data-rich 2D priors for data-limited 3D generation is a long-standing problem. DreamFields (Jain et al., 2022) leveraged CLIP (Radford et al., 2021) as a 2D prior for 3D generation. DreamFusion (Poole et al., 2023) then extended this direction to 2D diffusion models, introducing SDS, inspired by knowledge distillation (Hinton et al., 2015). Numerous variants followed (Metzer et al., 2023; Lin et al., 2023; Huang et al., 2024; Tang et al., 2023; Chen et al., 2023; Wang et al., 2023; Liang et al., 2024; Alldieck et al., 2024; McAllister et al., 2024), including an adaptation to flow models (Yang et al., 2025a). Notably, CFD (Yan et al., 2025) reduces SDS stochasticity through complex noise-warping and sampling mechanisms. Beyond 3D generation, SDS-like objectives have been widely applied (Haque et al., 2023; Gilo & Litany, 2026; Ling et al., 2024; Bahmani et al., 2024b;a; Bekor et al., 2025; Li et al., 2024b; Mann et al., 2026).

More recent text-to-3D approaches train feed-forward 3D generative models on large-scale 3D datasets (Deitke et al., 2023b;a; Zhang et al., 2025; Shi et al., 2024; Team Hunyuan3D et al., 2025; Xiang et al., 2025b;a; Li et al., 2025a;b). Many employ a two-stage pipeline, first generating a 2D image using a text-to-image model, then passing it to an image-to-3D generator. These methods mitigate limitations of 2D-prior-based approaches, such as the Janus problem (Shi et al., 2024), but remain constrained by the distribution of available 3D training data.

3 Method

Our goal is to leverage the generative prior of a frozen, text-conditioned next-scale AR model with parameters θ to optimize a differentiable representation, which we call the *host* and parameterize by ϕ. In our experiments, the host is a 3D NeRF or a 2D latent, though our formulation is not limited to these representations. Given a text prompt c, we render a 2D image $x = g(\phi, \pi)$ from the host, where π denotes the camera view in 3D generation. We seek a host whose renders agree with that prior,

$$\phi^\star \;=\; \arg\min_{\phi}\; \mathbb{E}_\pi\big[\mathcal{L}\big(g(\phi,\pi), c; \theta\big)\big], \tag{1}$$

where $\mathcal{L}$ is the distillation sampling objective developed in this section, given in Equation (10).

As in SDS, each step renders a view, applies the loss, and backpropagates through the renderer to the host (Figure 2). We first adapt the pseudo-target view of SDS to next-scale models (Section 3.2), then introduce KDS, which matches the full predicted distributions to differentiable host-induced token probabilities (Section 3.3), and stabilize optimization with lower-scale supervision and frozen-host regularization.

3.1 Preliminaries

Next-scale autoregressive models. Introduced by VAR (Tian et al., 2024), a next-scale AR model represents an image as K token maps $r_{1:K} = \{r_1, \ldots, r_K\}$ of increasing resolution $h_i \times w_i$, with

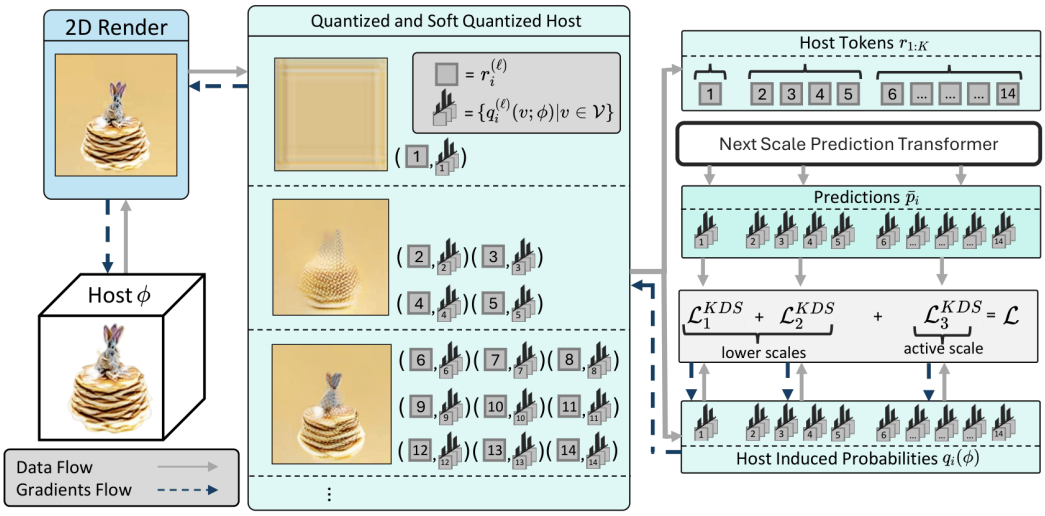

Figure 2: Overview of DreamAR. A host render x is tokenized into per-scale tokens $r_{1:K}$ and host-induced probabilities $q_i(\phi)$. Conditioned on preceding tokens $r_{<i}$, the frozen next-scale prediction transformer produces teacher probabilities $\bar{p}_i$. The KDS loss supervises the active and lower scales, backpropagating through $q_i(\phi)$ to the host. Frozen-host term $\mathcal{L}_{\text{fh}}$ is omitted for clarity.

$i \in \{1, \dots, K\}$ the scale index and $r_i \in \mathcal{V}^{h_i \times w_i}$. The vocabulary $\mathcal{V} = \{1, \dots, N\}$ holds N tokens, each token $v \in \mathcal{V}$ looked up in the codebook as $e[v] \in \mathbb{R}^d$. The model generates the maps coarse to fine, one forward pass per scale predicting all $h_i w_i$ tokens of that scale in parallel, conditioned on a text prompt c, so

$$p_\theta(r_{1:K} \mid c) \;=\; \prod_{i=1}^{K} \prod_{\ell=1}^{h_i w_i} p_\theta\Big(r_i^{(\ell)} \,\Big|\, r_{<i}, c\Big)\,, \tag{2}$$

with $r_i^{(\ell)}$ the token at position ℓ of r_i. Writing $p_{\theta,i}^{(\ell)}(v) = p_\theta(v \mid r_{<i}, c)$ for its prediction at that position, training minimizes the cross-entropy against the one-hot ground-truth tokens q_i,

$$\mathcal{L}_{\text{CE}}(\theta) \;=\; \sum_{i=1}^{K} \lambda_i \, H_i\big(q_i,\, p_{\theta,i}\big), \qquad H_i(a,b) \;:=\; -\sum_{\ell=1}^{h_i w_i} \sum_{v \in \mathcal{V}} a^{(\ell)}(v)\, \log b^{(\ell)}(v), \tag{3}$$

with per-scale weights λ_i (Appendix A.1). H_i is the per-scale cross-entropy between any two token distributions over scale i, and we reuse it below with other arguments. A tokenizer maps between images and token maps. An image x is encoded to a latent $z = \mathcal{E}(x) \in \mathbb{R}^{h_K \times w_K \times d}$ and quantized into the maps $r_{1:K} = \mathcal{Q}(z)$. Conversely, the maps are dequantized back to a latent,

$$\mathcal{A}(r_{1:m}) \;=\; \sum_{i=1}^{m} \text{up}_K\big(e[r_i]\big), \tag{4}$$

and decoded to $\mathcal{D}(\mathcal{A}(r_{1:K})) \approx x$, where $e[r_i] \in \mathbb{R}^{h_i \times w_i \times d}$ looks up each token's code and up_K interpolates to the finest grid $h_K \times w_K$. A prefix $m < K$ gives a coarser latent.

SDS and its pseudo-target view. SDS optimizes the host so that its renders $x = g(\phi, \pi)$ from arbitrary views appear likely to a frozen text-to-image model (Poole et al., 2023). We omit π below. Following HiFA (Zhu et al., 2024), let $z = \mathcal{E}(x)$ be the render's latent, $z_t = \alpha_t z + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ its noised version, $\hat{\epsilon} = \epsilon_\psi(z_t; c, t)$ the frozen denoiser's prediction, and $\hat{z} := \frac{1}{\alpha_t}(z_t - \sigma_t \hat{\epsilon})$ the resulting one-step estimate of z. Written as a latent residual, the SDS gradient is $\nabla_\phi \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\big[\omega(t) \frac{\alpha_t}{\sigma_t}(z - \hat{z}) \frac{\partial z}{\partial \phi}\big]$. Absorbing the t-dependent coefficients into $\omega(t)$, this is, up to that weighting, the gradient of a plain MSE against a pseudo-target held under the stop-gradient

operator $\mathrm{sg}[\cdot]$,

$$\mathcal{L}_{\mathrm{SDS}}(\phi) \;=\; \mathbb{E}_{t,\epsilon}\, \omega(t)\, \big\| z - \mathrm{sg}[\hat{z}] \big\|^2. \tag{5}$$

We now explore how this pseudo-target formulation can be adapted to next-scale AR models.

3.2 Pseudo-Target Approach

The optimization loop underlying Equation (5) can be adapted to next-scale models by replacing noise addition and denoising with re-generation (El-Ghoussani et al., 2026). Tokenizing the rendered view into its scale maps $r_{1:K} = \mathcal{Q}(\mathcal{E}(g(\phi)))$, at the active scale k being optimized, we fix the coarse prefix $r_{<k}$ and re-sample only the current scale from the frozen model, $\tilde{r}_k \sim p_\theta(\cdot \mid r_{<k}, c)$. The latent $\hat{z}_k$ reassembled from $(r_{<k}, \tilde{r}_k)$ then plays the role of the pseudo-target $\hat{z}$, and the truncation scale k that of the timestep t. With the dequantizer $\mathcal{A}$ of Equation (4), Equation (5) yields us the Pseudo-Target (PT) loss

$$\mathcal{L}_i^{\mathrm{PT}}(\phi) \;=\; \big\| \mathcal{A}\big(r_{1:i}(\phi)\big) \;-\; \mathrm{sg}\big[\mathcal{A}\big(r_{<i}(\phi), \tilde{r}_i\big)\big] \big\|^2, \tag{6}$$

taken at an active scale $i = k$ drawn from any schedule, uniform or annealed, in place of t. Both terms share the host prefix $r_{<k}(\phi)$. The target is conditioned on and reassembled from this prefix, but is fully detached, so gradients reach ϕ only through the encoder and quantizer of the left term. While resampling can extend to all remaining scales $k : K$, in practice we resample and optimize only the current scale, as it both reduces runtime and yields slightly better results. Since the model predicts from the uncorrupted render, the resampling need not be stochastic, unlike the noise in SDS. It can use the model's default inference sampling or select the most likely tokens, and we experimentally find the deterministic variant works better (Section 4.3).

3.3 Knowledge Distillation Sampling

Consider two views that favor different tokens for the same surface point. For example, one assigns probabilities 0.6 and 0.4 to tokens A and B, respectively, while the other assigns 0 and 1.0. A sampling based loss may sample A from the first view and B from the second, producing equally strong but conflicting targets. Yet the information needed to resolve this conflict already exists in the complete distributions, which jointly favor B. Sampling a single token discards this information. We therefore supervise the host with the full distributions. Inspired by the original knowledge distillation objective (Hinton et al., 2015), for a host render decomposed into per scale token maps $r_{1:K}$, we optimize scale k by minimizing the cross entropy between the host induced token probabilities and the predicted distribution.

Host Induced Probabilities. During tokenization of each scale, a given scale-wise continuous vector $z_i^{(\ell)}(\phi) \in \mathbb{R}^d$ is quantized to $r_i^{(\ell)}$ by $\mathcal{Q}$. In order to use the Knowledge Distillation loss, we need to phrase these host tokens as continuous probabilities. We use soft-quantization with a fixed quantization temperature τ,

$$q_i^{(\ell)}(v; \phi) \;=\; \frac{\exp\big(-\|z_i^{(\ell)}(\phi) - e[v]\|^2/\tau\big)}{\sum_{u \in \mathcal{V}} \exp\big(-\|z_i^{(\ell)}(\phi) - e[u]\|^2/\tau\big)}. \tag{7}$$

The teacher $\bar{p}_i^{(\ell)} = \mathrm{sg}\big[p_\theta(\cdot \mid r_{<i}(\phi), c)\big]$ is the frozen model's predicted probabilities, conditioned on the host's hard tokens for all $i \le k$. Our KDS term at scale i is

$$\mathcal{L}_i^{\mathrm{KDS}}(\phi) \;=\; \lambda_i\, H_i\big(\bar{p}_i,\, q_i(\phi)\big), \tag{8}$$

using the scale-wise weighting λ_i as in the original model's training objective. The active scale k gradually increases throughout optimization, with no loss applied to higher scales. Note that unlike the training loss in Equation (3) in which the VAR-predicted probabilities are on the right side, our loss uses a reversed order avoiding mode-seeking order, see Appendix A.6.

3.4 Conflicting Targets

A hidden assumption underlying this loss is worth making explicit. While optimizing scale k, its lower scales are assumed to be already optimized and fixed. Although conditioning predictions on

the host's lower scales enables deterministic supervision (Section 3.2), conflicting targets can still arise. Since the host is a shared representation, optimizing higher scales can also change lower ones, violating this assumption and shifting the target itself.

Lower-scale stabilization. To better satisfy this assumption, we gradually anneal the active scale k upward, allowing lower scales to converge before advancing. However, they may still be imperfect, especially in 3D, where different views update the same underlying host (Section 3.3), and should remain adjustable. We therefore continue supervising lower scales with the same $\mathcal{L}_i$, encouraging them to remain aligned with the teacher's predicted distributions without freezing them.

Specifically, we apply $\mathcal{L}_i^{\mathrm{PT}}$ or $\mathcal{L}_i^{\mathrm{KDS}}$ jointly to all scales $i < k$ (Equation (10)), using the same per scale weights λ_i. This adds no extra model cost, since the lower scales already serve as context for predicting scale k, so their probability maps are available at every step.

Frozen host. In 3D, the host may drift during optimization, most visibly through stray densities outside the object and color shifts. Unlike in SDS, where added noise can suppress such artifacts, here they enter the lower-scale prefix, are treated as valid structure or appearance, and are reinforced at finer scales. We observe this for a small subset of prompts, mainly at later scales, and add a regularizer that keeps the lower scales stable.

As optimization gradually advances the active scale k, we snapshot the current host as $\bar{\phi}$ whenever the schedule moves to the next scale, starting from k_0. At each subsequent active scale k, we render the same camera π from both hosts, tokenize the frozen render as $\bar{r}_{1:K} = \mathcal{Q}(\mathcal{E}(g(\bar{\phi}, \pi)))$, and anchor the lower scale latents up to depth k',

$$\mathcal{L}_{\mathrm{fh}}(\phi) \;=\; \big\|\mathcal{A}\big(r_{1:k'}(\phi,\pi)\big) \;-\; \mathrm{sg}\big[\mathcal{A}\big(\bar{r}_{1:k'}\big)\big]\big\|^2. \tag{9}$$

We choose k' a fixed number of scales below the active scale k, capped at k_{fh}, allowing recent scales to refine. The limit k_{fh} is chosen coarse enough to preserve structural flexibility, yet fine enough to capture density drift and pronounced color shifts.

3.5 Final Loss

Writing $\mathcal{L}_i$ for the base per-scale term, either $\mathcal{L}_i^{\mathrm{PT}}$ of Equation (6) or $\mathcal{L}_i^{\mathrm{KDS}}$ of Equation (8), the loss at active scale k is

$$\mathcal{L}(\phi) \;=\; \underbrace{\mathcal{L}_k(\phi)}_{\text{active scale}} \;+\; \lambda_{\mathrm{low}} \underbrace{\sum_{i=1}^{k-1} \mathcal{L}_i(\phi)}_{\text{lower scales}} \;+\; \lambda_{\mathrm{fh}} \underbrace{\mathcal{L}_{\mathrm{fh}}(\phi)}_{\text{frozen host}}. \tag{10}$$

Our final method loss is $\mathcal{L} = \mathcal{L}^{KDS}$, being stable enough to use $\lambda_{\mathrm{low}} = \lambda_{\mathrm{fh}} = 1$.

4 Experiments

Methods in comparison. We compare against SDS (Poole et al., 2023), SDS-Bridge (McAllister et al., 2024), VSD (Wang et al., 2023), and the leading methods CFD (Yan et al., 2025) and RFDS-Rev (Yang et al., 2025a). The former methods distill SD2.1-base (Rombach et al., 2022), while RFDS-Rev distills the newer rectified-flow SD3.5-Medium (Esser et al., 2024), which still uses classifier-free guidance at inference. KDS distills Infinity-2B (Han et al., 2025), and additionally Switti (Voronov et al., 2024) in 2D. Every baseline runs from its authors' released code at its recommended operating point. Since base models differ in quality, we also report each one's native sampling, so a method can be read against the prior it distills.

4.1 Text-to-2D Generation

We introduce a new distillation method and loss, and therefore seek to evaluate their core capabilities without the additional complexities of 3D optimization. As commonly done for qualitative or quantitative comparisons (Wang et al., 2023; McAllister et al., 2024; Yang et al., 2025a), the 2D setting eliminates potential sources of error, such as camera sampling and method-specific NeRF regularization, providing a clean, albeit simpler, evaluation of the distillation loss.

Evaluation Protocol and Metrics. We follow the text-to-2D protocol of SDS-Bridge (McAllister et al., 2024), optimizing one latent per prompt against the frozen base model and reporting zero-shot FID and CLIP-FID (clean-fid) against the corresponding real images. In place of its COCO captions we use the MJHQ-30K (Li et al., 2024a), newer 30K category-balanced Midjourney images its authors propose as an FID reference for human-aligned aesthetic rather than photographic quality. We draw 5,000 prompts, 500 per category. We also evaluate each base model's native generation quality at its original resolution, before distillation sampling. For the comparison methods, we use their default 2D configurations from the released code whenever available. For CFD (Yan et al., 2025), we use 800 instead of the default 1,000 optimization steps, as we observed performance degradation during the final 200 steps. Quantitative results are reported in Table 1, with qualitative results shown in Figures 6 and 8.

Table 1: **Text-to-2D on MJHQ-30K (Li et al., 2024a).** KDS achieves the best FID and CLIP-FID among all distillation sampling methods.

Method	Base model	FID ↓	CLIP-FID ↓
Standard sampling			
SD2.1 (Rombach et al., 2022)	—	23.59	7.31
Infinity-2B (Han et al., 2025)	—	20.38	4.47
Switti (Voronov et al., 2024)	—	16.45	3.07
SD3.5-M (Esser et al., 2024)	—	18.07	4.67
Distillation sampling			
KDS (ours)	Infinity-2B	**24.24**	**5.20**
KDS (ours)	Switti	34.31	7.53
CFD (Yan et al., 2025)	SD2.1	36.55	12.35
RFDS-Rev (Yang et al., 2025a)	SD3.5-M	38.66	13.47
VSD (Wang et al., 2023)	SD2.1	52.41	16.51
SDS-Bridge (McAllister et al., 2024)	SD2.1	59.50	27.02
SDS (Poole et al., 2023)	SD2.1	64.16	20.95

Table 2: PT sampling variants versus KDS in text-to-2D, using Switti (Voronov et al., 2024) and Infinity (Han et al., 2025). KDS achieves the lowest FID and CLIP-FID for both base models.

Method	FID ↓	CLIP-FID ↓
Infinity-2B		
PT (stochastic)	28.57	8.07
PT (deterministic)	25.61	5.49
KDS (ours)	**24.24**	**5.20**
Switti		
PT (stochastic)	39.38	10.40
PT (deterministic)	34.79	7.85
KDS (ours)	**34.31**	**7.53**

Results. KDS outperforms all distillation sampling methods on both metrics, achieving 24.24 FID and 5.20 CLIP FID, compared with 36.55 and 12.35 for CFD, the strongest baseline. With Switti as its base model KDS reaches 34.31 and 7.53. It also outperforms the recent flow based RFDS Rev baseline, despite its underlying model outperforming Infinity, the base model used by KDS.

Examining the relationship between base model quality and distillation performance reveals several interesting trends. Although Infinity has worse native FID than SD3.5-M, KDS closely preserves its base model's performance, outperforming RFDS-Rev. Conversely, despite Switti's better native FID than Infinity, Switti-based KDS exhibits a larger performance gap. This may stem from Switti's conventional codebook, compared to Infinity's bitwise quantization. Nevertheless, Switti-based KDS remains the second-best distillation sampling method, behind Infinity-based KDS.

4.2 Text-to-3D Generation

Recent text-to-3D approaches are predominantly feed-forward based (Section 2.2). To ensure a relevant comparison, we follow CFD (Yan et al., 2025) in using distillation sampling to refine objects generated from a 3D prior. This mitigates common failures such as the Janus problem while leveraging the 2D prior to refine both geometry and appearance beyond the limited 3D training distribution. CFD's original refinement recipe uses MVDream-based initializations (Shi et al., 2024). Our ablation (Appendix A.3) shows that CFD benefits from feed-forward initializations using TRELLIS.2 (Xiang et al., 2025a) or Hunyuan3D-2.1 (Team Hunyuan3D et al., 2025), while KDS is less sensitive to initialization, slightly favoring TRELLIS.2, which we select for refinement.

Evaluation Protocol and Metrics. We use 100 prompts from the DreamFusion prompt set, rendering each asset from 60 uniformly spaced azimuths at 30° elevation with a fixed camera (Appendix A.2). We additionally evaluate TRELLIS.2 and Hunyuan3D-2.1 as feed-forward baselines, each with its own recommended text-to-image stage, FLUX.1-schnell (Black Forest Labs, 2024) for TRELLIS.2 and HunyuanDiT-v1.1-distilled (Li et al., 2024d) for Hunyuan3D-2.1, rendering both at the same evaluation cameras.

The appearance metrics, CLIP score and 3D-FID, can demonstrate the benefits of 2D prior refinement over the 3D prior based feed-forward methods. To assess structure, we additionally report Uni3D-T and ULIP-T, the shape metrics used by the baselines (Xiang et al., 2025a; Team Hunyuan3D et al., 2025), which measure cosine similarity between surface point cloud and text embeddings. Since point-cloud embeddings are learned on limited 3D data distributions and may miss fine geometric detail, we complement these metrics with depth 3D-FID (d-3D-FID). Computed analogously to 3D-FID, it compares depth maps estimated from asset renders with those from images generated by the same FLUX.1-schnell teacher, enabling structural evaluation against the more diverse 2D distribution. However, as it relies on depth estimated from RGB renders, d-3D-FID alone cannot rule out view overfitting or poor underlying geometry and should be interpreted only when Uni3D-T and ULIP-T scores are reasonable.

We evaluate distillation sampling both from scratch and as refinement, initializing all refinement methods from the same TRELLIS.2 asset. Quantitative results are reported in Tables 3 and 4, respectively, with qualitative results shown in Figures 9 and 10 for from-scratch generation and Figures 11 and 12 for refinement.

4.2.1 Distillation Sampling from Scratch

Table 3: Plain text-to-3D evaluation over 100 prompts. All methods optimize their representations from scratch, without a 3D prior.

Method	Base model	CLIP-L/14 ↑	CLIP-B/16 ↑	3D-FID ↓	ULIP-T ↑	Uni3D-T ↑
SDS	SD2.1	23.08	28.24	131.8	0.196	0.175
VSD	SD2.1	25.10	30.05	108.9	0.121	0.131
SDS-Bridge	SD2.1	23.14	27.95	141.4	0.183	0.169
RFDS-Rev	SD3.5-M	**28.32**	**33.21**	91.4	**0.198**	**0.178**
CFD	SD2.1	27.95	32.64	88.7	0.182	0.160
KDS (ours)	Infinity-2B	27.61	32.44	**84.5**	0.141	0.132

The appearance metrics reveal two trends. RFDS-Rev, CFD, and KDS substantially outperform SDS, VSD, and SDS-Bridge on both CLIP scores and 3D-FID. Among these methods, RFDS-Rev achieves the highest CLIP scores, while KDS achieves the best 3D-FID.

Without a 3D prior, all methods achieve substantially lower ULIP-T and Uni3D-T scores than in the refinement setting (Table 4), highlighting the structural limitations of distillation sampling from scratch. Despite its strong appearance scores, KDS also performs poorly on these shape metrics. We attribute this to the limited sample diversity of bitwise AR models such as Infinity (Yang et al., 2025b), which we hypothesize also limits how reliably it follows view-dependent prompts, encouraging Janus artifacts.

4.2.2 Distillation Sampling as 3D Prior Refinement

Table 4: **Distillation Sampling as 3D Prior Refinement.** Comparison over 100 prompts. Top: feed-forward text-to-3D methods based on 3D priors. Bottom: 2D prior based distillation sampling methods refining 3D initializations. † All refinement methods start from the same TRELLIS.2 asset (appendix A.3).

Method	Base model	CLIP-L/14 ↑	CLIP-B/16 ↑	3D-FID ↓	d-3D-FID ↓	ULIP-T ↑	Uni3D-T ↑
Hunyuan3D-2.1	feed-fwd	25.13	29.71	115.8	59.6	0.344	0.295
TRELLIS.2	feed-fwd	25.96	30.44	97.7	53.5	0.358	0.300
CFD†	SD2.1	28.20	33.04	87.3	52.4	0.252	0.218
RFDS-Rev†	SD3.5-M	**28.62**	**33.81**	**67.4**	**43.4**	0.263	0.232
KDS† (ours)	Infinity-2B	28.37	33.39	76.0	51.2	**0.286**	**0.239**

In the refinement setting (Table 4), TRELLIS.2 and Hunyuan3D-2.1 achieve the highest ULIP-T and Uni3D-T scores, reflecting their strong alignment with the 3D training distribution. Distillation sampling achieves substantially higher shape scores than in the from-scratch setting (Table 3),

although still below the feed-forward baselines, lending greater significance to d-3D-FID. Refinement methods improve d-3D-FID over these baselines, suggesting that 2D-prior refinement enhances structural diversity beyond appearance alone.

For appearance, refinement improves both CLIP scores and 3D-FID over the initializations. These metrics also improve relative to the corresponding plain distillation methods, demonstrating that 3D initialization benefits appearance as well as structure. KDS performs on par with the leading methods, achieving the highest ULIP-T and Uni3D-T scores, while RFDS-Rev leads by a small margin on CLIP-L/B, 3D-FID, and d-3D-FID. These results demonstrate the effectiveness of KDS and the potential of next-scale AR models, which we enable for distillation sampling for the first time.

4.3 Ablations

PT Sampling and KDS. As discussed in Section 3.2, PT supports both stochastic and deterministic sampling. We compare these variants against KDS in the isolated 2D setting, using both Switti and Infinity (Table 2).

KDS outperforms both PT variants across both metrics and base models, demonstrating the benefit of utilizing the full predicted distributions for more informative gradients. Deterministic PT also improves over stochastic PT, highlighting the benefit of more coherent gradients. Qualitative results are shown in Appendix A.4.

Loss Component Ablation. We ablate the terms of Equation (10) in the 3D refinement setting on the same 30-prompt subset used in Appendix A.3, removing one term at a time. Quantiative results are reported in Table 5, with qualitative results shown in Appendix A.5.

Table 5: Loss term ablation, KDS refinement on a 30-prompts subset, otherwise the protocol of Table 4.

Recipe	CLIP L/14 ↑
$\mathcal{L}_i = \mathcal{L}_i^{\mathrm{PT}}$ in place of $\mathcal{L}_i^{\mathrm{KDS}}$	26.67
No frozen host ($\lambda_{\mathrm{fh}} = 0$)	27.62
No lower scales ($\lambda_{\mathrm{low}} = 0$)	27.04
Ours (eq. (10))	**28.18**

Results demonstrate the significant quality increase when using KDS informative gradients compared to the deterministic PT loss, aligning with Section 4.3. The important effect of the lower-scale loss λ_{low}, which we find important for most prompts, is also evident. The frozen-host loss λ_{fh} provides a smaller yet noticeable improvement, primarily benefiting a small subset of prompts by preventing NeRF drift and the emergence of floaters.

5 Conclusion, Limitations and Future Work

We introduced DreamAR, enabling distillation sampling with next-scale AR models for the first time. Its KDS objective matches the teacher's full predicted token distributions to differentiable host-induced probabilities. The teacher's prediction is fully determined by the render and the active scale, whereas diffusion distillation samples fresh noise at every step and needs dedicated machinery to contain the resulting variance. In text-to-2D, KDS outperforms all diffusion and flow distillation baselines and stays within 4 FID of its base model's native sampling, while RFDS-Rev falls over 20 FID behind SD3.5-M. Diffusion distillation has matured over four years since DreamFusion, and as the first distillation sampling method for next-scale AR models, KDS already performs on par with its strongest variants, CFD and RFDS-Rev, in 3D refinement, with the highest shape scores among them.

The 2D margin does not yet carry over to 3D. From scratch, KDS matches the strongest baselines in appearance but trails them on shape metrics, and in refinement RFDS-Rev leads on appearance. We attribute this to the limited sample diversity of its bitwise base model, which we hypothesize leaves it unable to reliably produce view-dependent outputs, weakening supervision of the full 3D shape and encouraging Janus artifacts. Optimization takes 1–3 hours on a single GPU, dominated by the per-step cost of the AR teacher. Looking forward, KDS makes minimal assumptions about the codebook and runs on both Infinity and Switti, so it stands to benefit directly as next-scale AR models improve, including in view-dependent generation. Extending it to applications built on score distillation, such as editing, video and motion, is a natural next step.

AI USE STATEMENT

We used generative AI tools to formulate mathematical claims, implement methods, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis. We did not use them to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or give feedback on our research methodology or experiments, or interpret results. The remaining disclosure tasks do not apply to this work, since it contains no synthetic data and no formal proofs. Among tasks with recommended disclosure, we also used AI assistants to edit and shorten prose and to help locate related work.

We have reviewed all AI-assisted work. Every AI-drafted derivation was rederived by hand by at least one author, AI-assisted code was read and tested by the authors and all reported numbers come from runs of that code, and every AI-suggested reference was checked against its primary source. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method, and Appendix A.2 the 3D evaluation protocol. We will release code, configs, and prompt lists with the camera ready version.

REFERENCES

Thiemo Alldieck, Nikos Kolotouros, and Cristian Sminchisescu. Score distillation sampling with learned manifold corrective. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 1–18. Springer, 2024. arXiv:2401.05293.

Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, Andrea Tagliasacchi, and David B. Lindell. Tc4d: Trajectory-conditioned text-to-4d generation. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 53–72. Springer, 2024a. arXiv:2403.17920.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7996–8006, 2024b. arXiv:2311.17984.

Yarin Bekor, Gal Michael Harari, Or Perel, and Or Litany. Gaussian see, gaussian do: Semantic 3d motion transfer from multiview video. In *SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10. ACM, 2025. arXiv:2511.14848.

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

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22246–22256, 2023. arXiv:2303.13873.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pp. 35799–35813, 2023a. arXiv:2307.05663.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, 2023b. arXiv:2212.08051.

Amir El-Ghoussani, Marc Hölle, Gustavo Carneiro, and Vasileios Belagiannis. Prompt-guided image editing with masked logit nudging in visual autoregressive models. *arXiv preprint arXiv:2604.14591*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, pp. 12606–12633, 2024. arXiv:2403.03206.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. In *International Conference on Learning Representations (ICLR)*, pp. 100218–100231, 2025. arXiv:2410.13863.

Daniel Gilo and Or Litany. Instructmix2mix: Consistent sparse-view editing through multi-view model personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29000–29011, 2026. arXiv:2511.14899.

Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Shu-Tao Xia, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. In *International Conference on Learning Representations (ICLR)*, pp. 93867–93889, 2026. arXiv:2508.10711.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Oral; arXiv:2412.04431.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19740–19750, 2023. arXiv:2303.12789.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NeurIPS 2014 Deep Learning Workshop.

Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.12422.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 867–876, 2022. arXiv:2112.01455.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. Coin: Control-inpainting diffusion prior for human and camera motion estimation. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 426–446. Springer, 2024b. arXiv:2408.16426.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 56424–56445, 2024c. arXiv:2406.11838.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, Xiao Chen, Feipeng Tian, Jianxiong Pan, Zeming Li, Gang Yu, Xiangyu Zhang, Daxin Jiang, and Ping Tan. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025a.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025b.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024d. Open weights: `https://huggingface.co/Tencent-Hunyuan/HunyuanDiT-v1.1-Diffusers-Distilled`.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6526, 2024. arXiv:2311.11284.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, 2023. arXiv:2211.10440.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8576–8588, 2024. arXiv:2312.13763.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 170054–170072, 2025. Oral; arXiv:2511.04675.

Amir Mann, Gal Michael Harari, Merav Keidar, and Or Litany. Videomdm: Towards 3d human motion generation from 2d supervision. *arXiv preprint arXiv:2606.13364*, 2026.

David McAllister, Songwei Ge, Jia-Bin Huang, David W. Jacobs, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.09417.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12663–12673, 2023. arXiv:2211.07600.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.14988.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. arXiv:2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. arXiv:2112.10752.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*, pp. 39838–39859, 2024. arXiv:2308.16512.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Boshi Tang, Jianan Wang, Zhiyong Wu, and Lei Zhang. Stable score distillation for high-quality 3d generation. *arXiv preprint arXiv:2312.09305*, 2023.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. Open weights: `https://huggingface.co/tencent/Hunyuan3D-2.1`.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Best Paper Award; arXiv:2404.02905.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Jiahuan Wang, Yuxin Chen, Jun Yu, Guangming Lu, and Wenjie Pei. Editinfinity: Image editing with binary-quantized generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 50763–50790, 2025a. arXiv:2510.20217.

Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. arXiv:2503.23897.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Spotlight; arXiv:2305.16213.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025a. TRELLIS.2; open weights: `https://github.com/microsoft/TRELLIS.2`.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, 2025b. arXiv:2412.01506.

Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.05445.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.09414.

Xiaofeng Yang, Cheng Chen, Xulei Yang, Fayao Liu, and Guosheng Lin. Text-to-image rectified flow as plug-and-play priors. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2406.03293.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, and Chenyang Si. Diversear: Boosting diversity in bitwise autoregressive image generation. *arXiv preprint arXiv:2512.02931*, 2025b.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18431–18441, 2025. arXiv:2411.00776.

Shengqiang Zhang, Ruotong Liao, Volker Tresp, Barbara Plank, and Hinrich Schütze. Edit the bits, diff the codes: Bitwise residual editing for visual autoregressive models. *arXiv preprint arXiv:2606.13558*, 2026.

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3d objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025.

Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. HiFA: High-fidelity text-to-3d generation with advanced diffusion guidance. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.18766.

A Appendix

A.1 Implementation details

Scale weights. Equations (3) and (8) weight each token of scale i by $\lambda_i \propto \sqrt{h_K w_K/(h_i w_i)}$, Infinity's training weighting (Han et al., 2025). Coarse tokens thus weigh more, yet a scale's total weight $h_i w_i \lambda_i \propto \sqrt{h_i w_i}$ grows with resolution. We normalize over the computed scales, $\sum_{i=1}^{k} h_i w_i \lambda_i = 1$.

Guidance scale. Teacher logits use classifier-free guidance and temperature 0.5, Infinity's sampling default. A guidance scale of 3 works in 2D, while 3D appears to need 100.

A.2 Text-to-3D evaluation protocol

Setup. From the 415 DreamFusion prompts (Poole et al., 2023) we take all 81 used by later papers plus 19 random unused ones. 5 scene-level prompts were excluded (landmarks, aerial views, extended architecture, scene-embedded subjects) and replaced by further random unused ones. Every method renders at camera distance 1.8 and field of view $65°$, overriding its configuration's defaults.

Metrics. CLIP (L/14, B/16) is $100\cos$ of normalized image and prompt embeddings (not CLIP-Score, nor comparable to CFD's maximum over 120 views). FID uses `torchmetrics` (2048-d Inception pool3) against cached statistics of the reference set, 500 images per prompt from FLUX.1-schnell (Black Forest Labs, 2024). Depth maps (Depth Anything V2 Large (Yang et al., 2024)) are min-max normalized per image and replicated to three channels. d-3D-FID applies FID to them. ULIP-T and Uni3D-T follow Team Hunyuan3D et al. (2025) and Xiang et al. (2025a): 8,192 area-uniform surface points with albedo, sampled from the generated mesh for the feed-forward rows and from a marching-cubes surface of the NeRF (iso-level 25, 256^3) for the refined ones, mean-centred and scaled to the unit sphere in the glTF frame, embedded by ULIP-2 (colored PointBERT, ViT-bigG-14 text tower) and Uni3D-g (EVA02-E-14-plus text tower), and scored as the cosine to the text embedding of the prompt. Every row keeps its NeRF or mesh. A NeRF with no surface at iso-level 25 is extracted at the highest lower level that yields one (four SDS runs, levels 0.02 to 1), and a NeRF with no density at all is left out (three SDS-Bridge runs). Absolute values are not comparable across papers (the ULIP column of Team Hunyuan3D et al. (2025) uses a different checkpoint); we compare within the table.

A.3 Initialization Ablation

We evaluate four initializations with both refinement methods on a representative 30-prompt subset of our 100-prompt benchmark (240 runs). The subset is selected by stratifying prompts according to their average CLIP-L/14 scores across optimization-based methods. MVDream+CFD corresponds to CFD's first stage, while MVDream+SDS uses the same setup with consistent noise disabled. Since FID estimates from this smaller subset are less reliable, we report only per-sample metrics. Our ablation (Appendix A.3) shows that CFD benefits from feed-forward initializations using TRELLIS.2 (Xiang et al., 2025a) or Hunyuan3D-2.1 (Team Hunyuan3D et al., 2025), while KDS is less sensitive to initialization, slightly favoring TRELLIS.2, which we select for refinement.

A.4 PT resampling, qualitative

Figure 3 shows the three rows of Table 2 on four prompts, for both base models.

A.5 Loss term ablation, qualitative

Figure 4 shows the rows of Table 5 on three prompts, one chosen per ablated term (frozen host, lower scales, PT, left to right), the first a classic floater failure.

Table 6: Initialization ablation on 30 prompts, before and after each refiner. Cells are the mean over prompts, best refined cell per column in bold.

Stage	Initialization	CLIP L/14	CLIP B/16
Initialization only	MVDream+CFD	24.75	29.94
	MVDream+SDS	24.52	29.97
	TRELLIS.2	24.43	29.44
	Hunyuan3D-2.1	25.06	29.94
+ CFD refine	MVDream+CFD	26.78	31.95
	MVDream+SDS	26.46	32.10
	TRELLIS.2	27.77	32.80
	Hunyuan3D-2.1	27.83	32.95
+ KDS refine	MVDream+CFD	28.27	33.19
	MVDream+SDS	28.13	33.17
	TRELLIS.2	**28.42**	**33.82**
	Hunyuan3D-2.1	27.73	33.34

A.6 Which distribution sits inside the cross-entropy log

Fix a scale i and position ℓ, and write $q = q_i^{(\ell)}(\cdot\,;\phi)$ for the host distribution of Equation (7), $\bar{p} = \bar{p}_i^{(\ell)}$ for the teacher's, and $H(a,b) = -\sum_{v\in\mathcal{V}} a(v)\log b(v)$, so that H_i of Equation (3) sums H over positions. Each term of Equation (8) is then $H(\bar{p}, q)$, the reverse of the training order of Equation (3). The reverse order $H(q,\bar{p})$ is linear in q, so over the simplex it is minimized by the one-hot at $\arg\max_v \bar{p}(v)$ (mode-seeking), whereas $H(\bar{p},q)$ is minimized at $q=\bar{p}$, keeping host mass on every token the teacher finds plausible.

Both orders depend on the host code $z_i^{(\ell)}$ only through q. Let $\nabla := \partial/\partial z_i^{(\ell)}$, with all sums over $\mathcal{V}$. By the chain rule, each gradient pairs a factor specific to the order with a factor the two share,

$$\nabla H = \sum_v \frac{\partial H}{\partial q(v)}\,\nabla q(v). \tag{11}$$

Shared factor. Write Equation (7) as $\log q(v) = s_v - \log\sum_u \exp s_u$ with $s_v = -\|z_i^{(\ell)} - e[v]\|^2/\tau$, so $\nabla s_v = \frac{2}{\tau}(e[v] - z_i^{(\ell)})$. Then

$$\nabla \log q(v) = \nabla s_v - \sum_u q(u)\,\nabla s_u = \frac{2}{\tau}\big(e[v] - \mathbb{E}_q[e]\big), \qquad \nabla q(v) = q(v)\,\nabla\log q(v) = \frac{2}{\tau}\,q(v)\big(e[v] - \mathbb{E}_q[e]\big), \tag{12}$$

where $\mathbb{E}_q[e] = \sum_v q(v)\,e[v]$ is the mean codebook entry under q (likewise $\mathbb{E}_{\bar{p}}[e]$), and the $z_i^{(\ell)}$ terms cancel. So $\nabla q(v)$ is proportional to $q(v)$ itself.

Order-specific factor. The reverse order is linear in q, so its factor is a constant, whereas ours carries a $1/q(v)$ weight,

$$\frac{\partial H(q,\bar{p})}{\partial q(v)} = -\log\bar{p}(v), \qquad \frac{\partial H(\bar{p},q)}{\partial q(v)} = -\frac{\bar{p}(v)}{q(v)}. \tag{13}$$

The two gradients. Substituting Equations (12) and (13) into Equation (11), and using $\sum_v \bar{p}(v) = 1$ for ours,

$$\begin{aligned}\nabla H(q,\bar{p}) &= -\frac{2}{\tau}\sum_v q(v)\log\bar{p}(v)\,\big(e[v] - \mathbb{E}_q[e]\big),\\ \nabla H(\bar{p},q) &= -\frac{2}{\tau}\sum_v \bar{p}(v)\big(e[v] - \mathbb{E}_q[e]\big) = \frac{2}{\tau}\Big(\mathbb{E}_q[e] - \mathbb{E}_{\bar{p}}[e]\Big).\end{aligned} \tag{14}$$

The reverse order weights each token by the host's $q(v)$, ours by the teacher's $\bar{p}(v)$, since its $1/q(v)$ cancels the $q(v)$ of Equation (12) exactly.

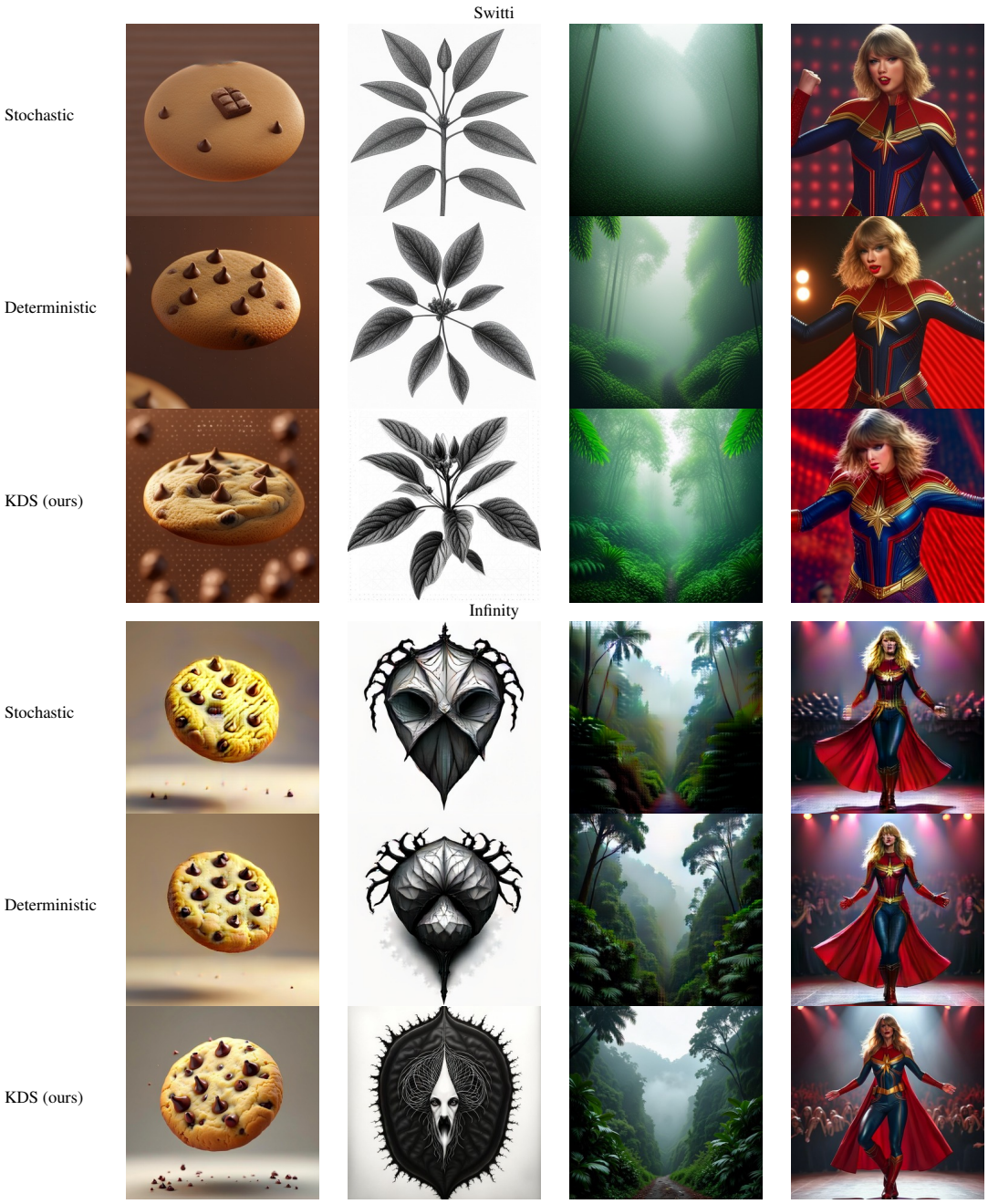

Figure 3: The rows of Table 2 on four MJHQ-30K prompts (a floating chocolate chip cookie, a black and white geometric drawing of Atropa belladonna, a misty rainforest, Taylor Swift dressed as Captain Marvel), same seed per column.

A far token. Let the teacher favor a token $v^\star$ whose entry $e[v^\star]$ is far from $z_i^{(\ell)}$, so $q(v^\star) \propto \exp(-\|z_i^{(\ell)} - e[v^\star]\|^2/\tau) \approx 0$. In the reverse order the pull toward $v^\star$ scales with $q(v^\star)$ and vanishes. The host sits on a plateau, where any small move leaves $q(v^\star) \approx 0$. In our order the pull toward $e[v^\star]$ is weighted by $\bar{p}(v^\star)$ and does not fade with distance.

Bitwise and lookup codebooks. Infinity's bitwise codebook is $e[v] \in \{\pm 1/\sqrt{d}\}^d$ with $N = 2^d$, so Equation (7) factorizes over the d coordinates of $z_i^{(\ell)}$, each a bit with two entries. A far token

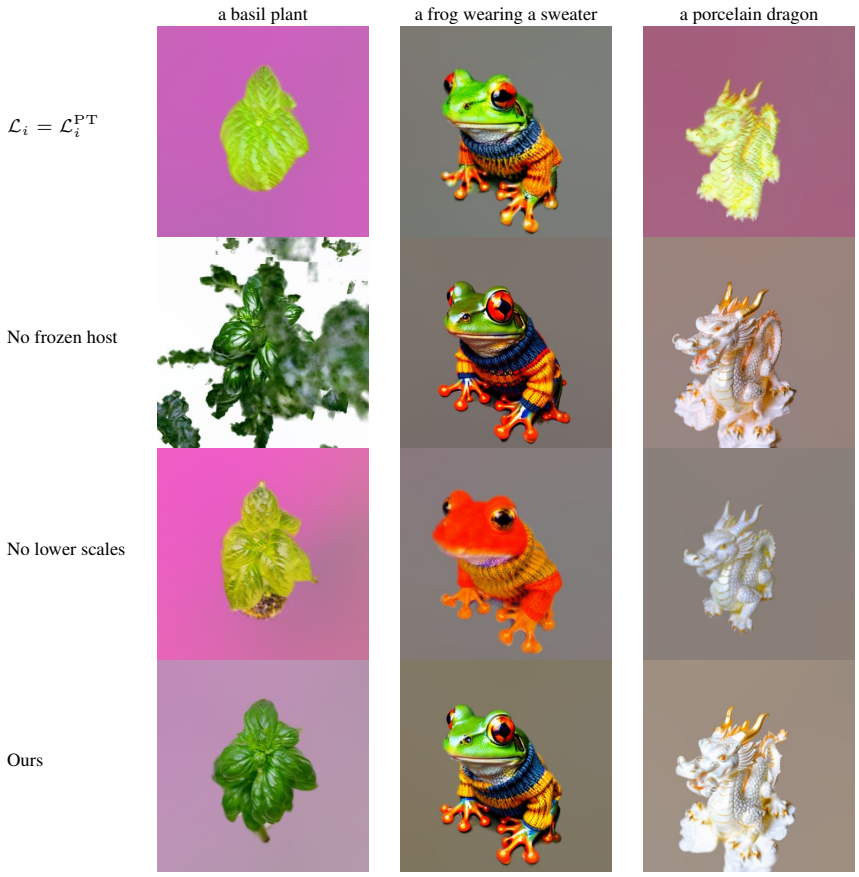

Figure 4: Loss term ablation of Table 5, one render per run. Prompts are "a DSLR photo of" followed by the column title.

is unlikely there, since each coordinate moves along a single axis between its two entries and its range is bounded. In practice both orders work there. Under a large lookup codebook, as in Switti (Voronov et al., 2024), most entries are far from any given code, and empirically only our order works (Figure 5).

A.7 Qualitative results

Figures 6 to 8 show 2D KDS samples on MJHQ-30K prompts and on prompts from the Infinity paper. Figures 9 and 11 show text-to-3D results from scratch and from TRELLIS.2 initializations, and Figures 10 and 12 their failure cases.

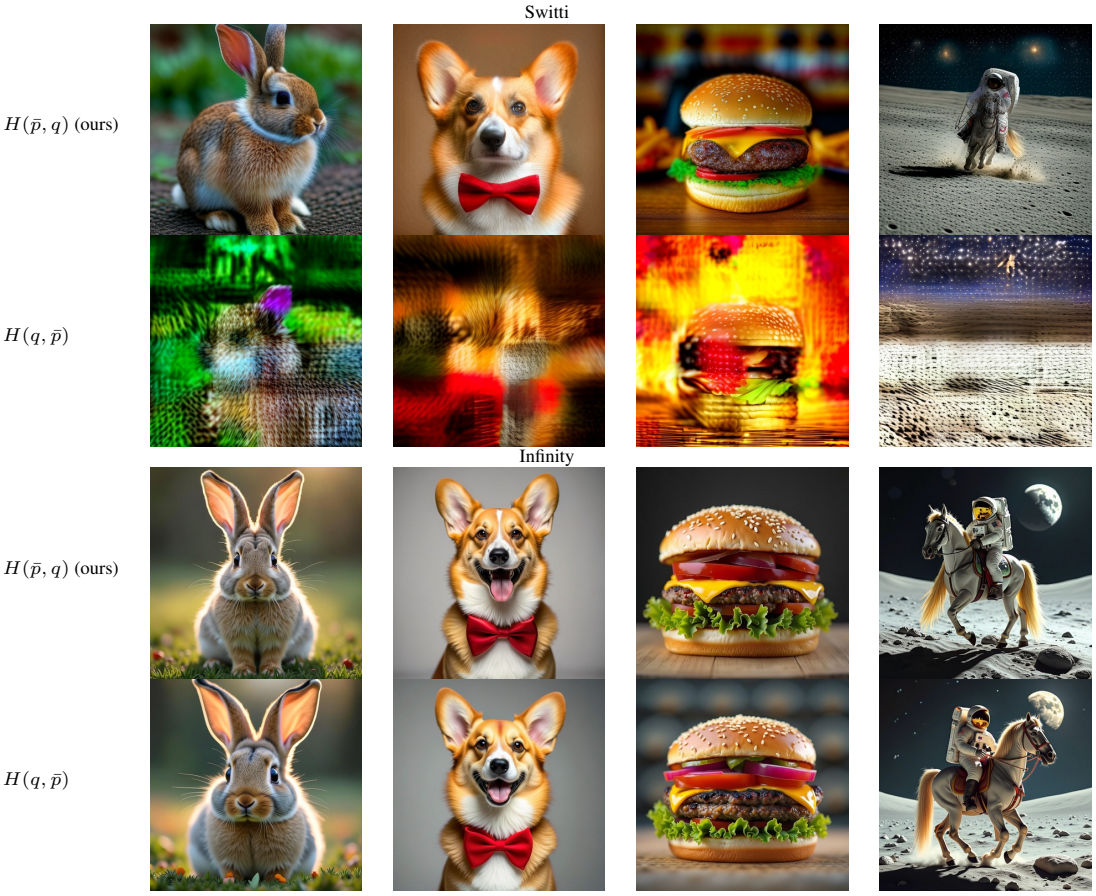

Figure 5: 2D KDS under both cross-entropy orders with Switti (top) and Infinity-2B (bottom), one prompt per column (a bunny, a corgi wearing a red bowtie, a hamburger, an astronaut riding a horse on the moon). Same recipe and seed per backbone, only the order differs. Only Switti's lookup codebook needs our order.

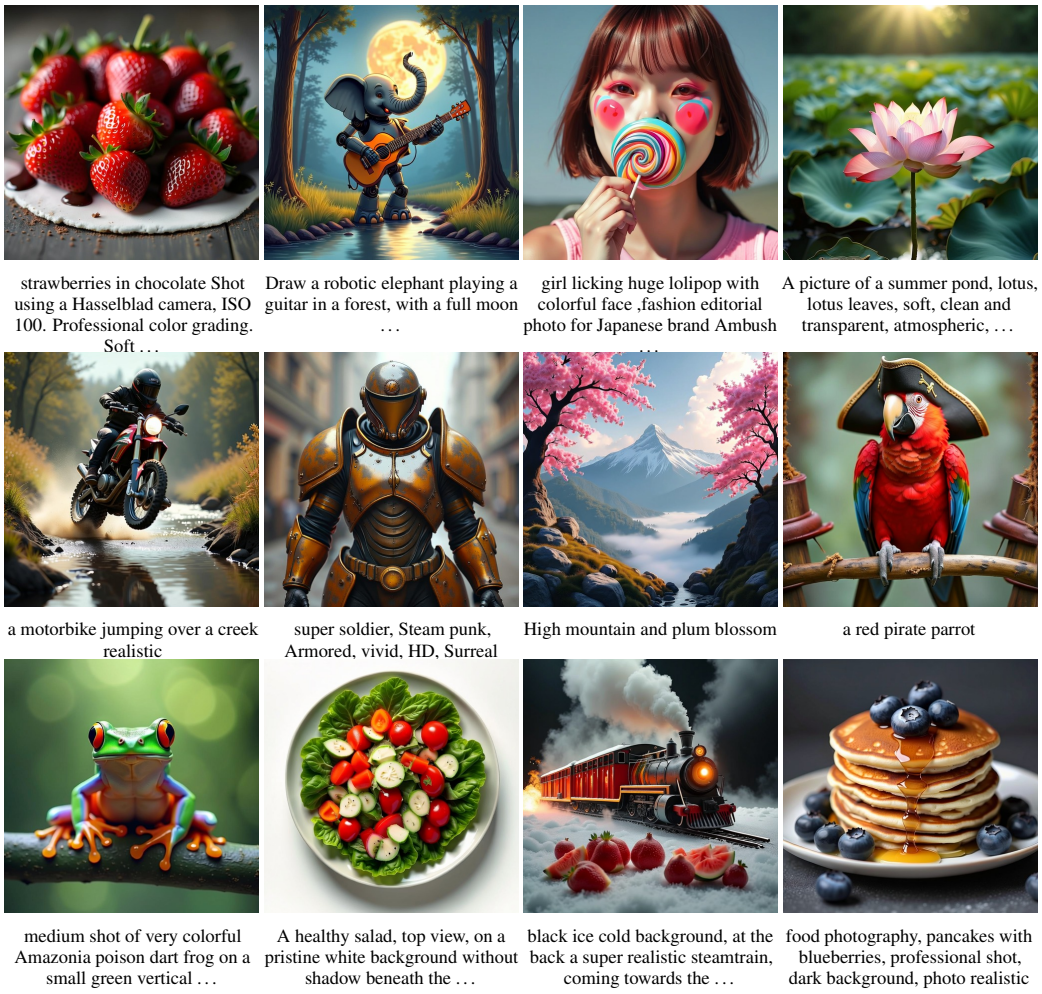

Figure 6: Text-to-image with KDS and Infinity-2B on MJHQ-30K prompts (truncated).



a giant, towering cloud in the shape of a man looms over the earth, . . .

An image of a futuristic robot at work in a high-tech laboratory, with advanced . . .

The text ‘Peace and Love’ placed elegantly on a serene ocean horizon during sunset, . . .

a tiny astronaut hatching from an egg on the moon

Insect made from vintage 1960s electronic components, capacitors, resistors, transistors, wires, diodes, solder, circuitboard

House: white; pink tinted windows; surrounded by flowers; cute; scenic; garden; fairy-like; epic; photography; . . .

Two superheroes called Spacefrog (a dashing green cartoon-like frog with a red cape) and . . .

A close-up photograph of a Corgi dog. The dog is wearing a black hat . . .

Three fuzzy muppet-like monsters, orange, green and orange, standing arm in arm at a . . .

The words ‘DREAM BIG’ in bold colorful graffiti lettering with dripping paint on a . . .

A glowing orange jellyfish floating above a green mountain valley at sunset

Macro photo of glowing translucent crystals, amber and pale blue, scattered on a wet . . .

Figure 7: Text-to-image with KDS and Infinity-2B on prompts from the Infinity paper (Han et al., 2025) (truncated).

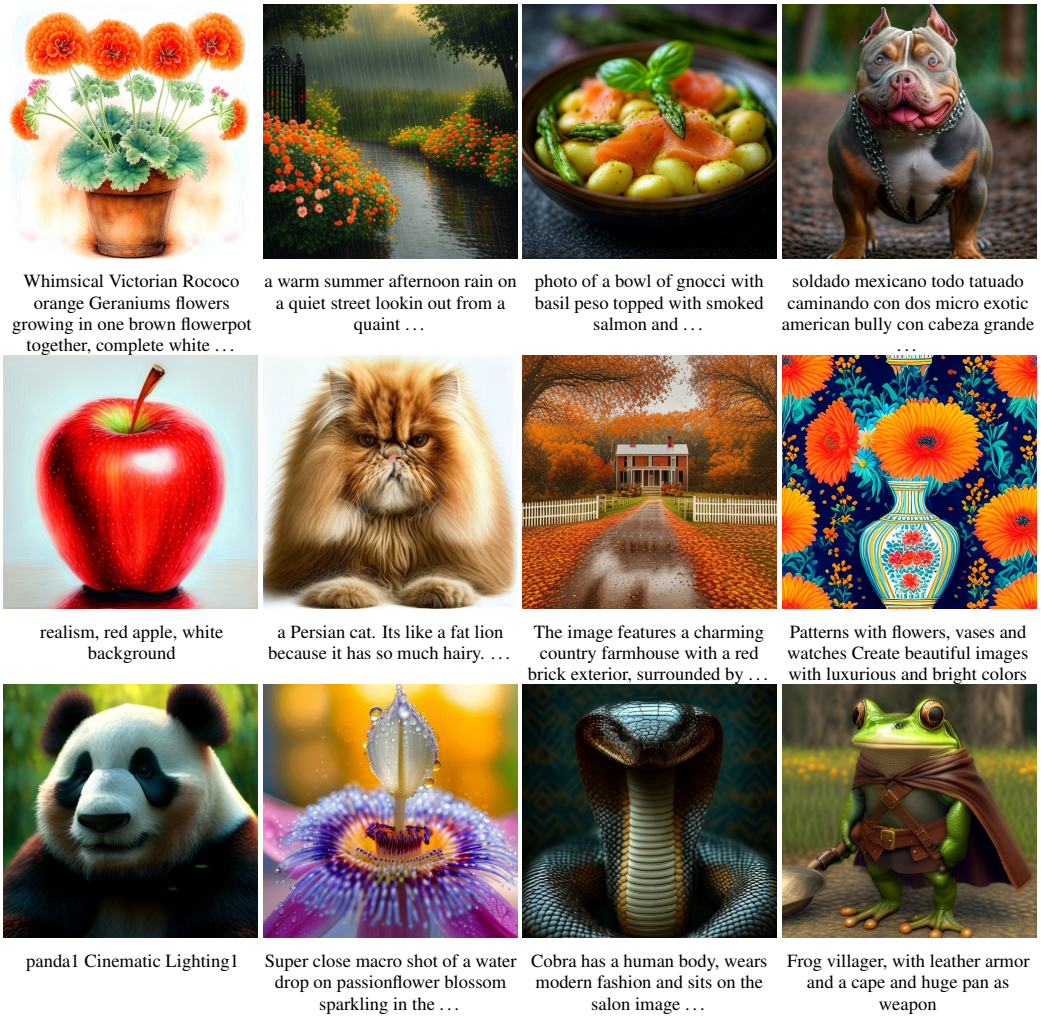

Figure 8: Text-to-image with KDS and Switti on MJHQ-30K prompts (truncated).

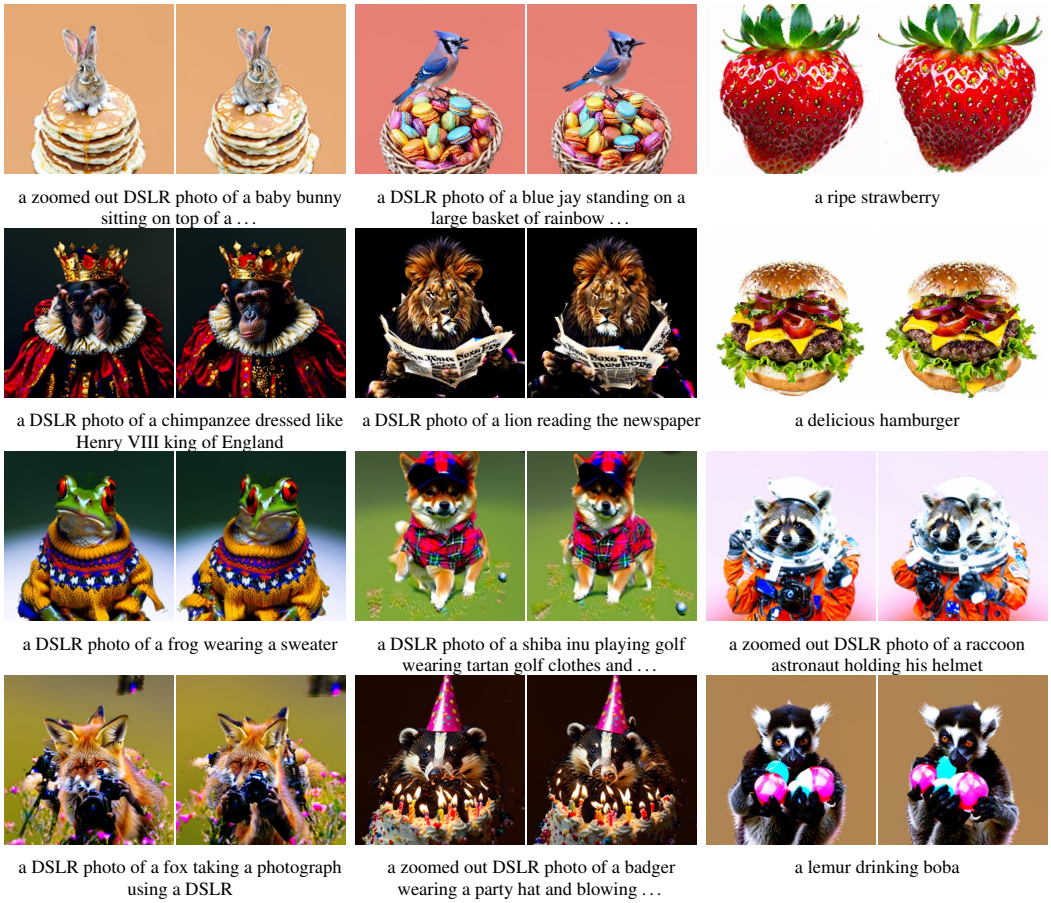

Figure 9: Text-to-3D with KDS from scratch (Infinity-2B), two nearby views per prompt.

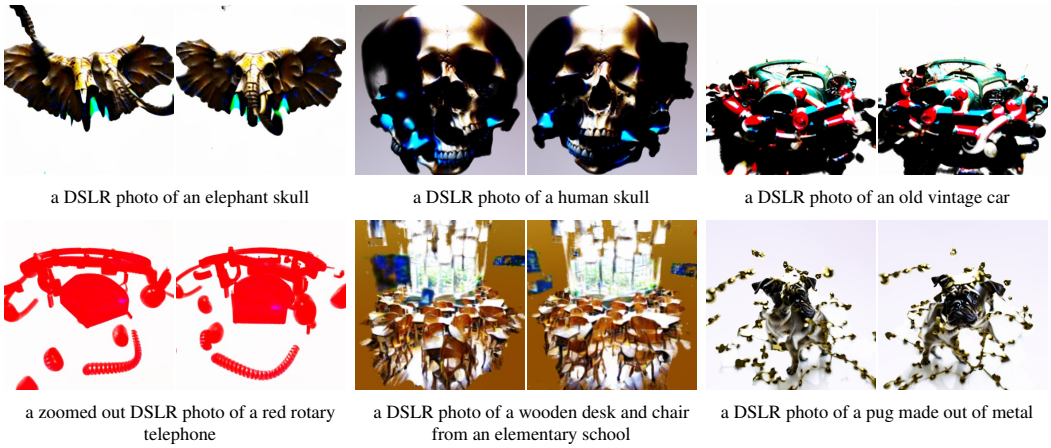

Figure 10: Failure cases of text-to-3D with KDS from scratch, two nearby views per prompt.

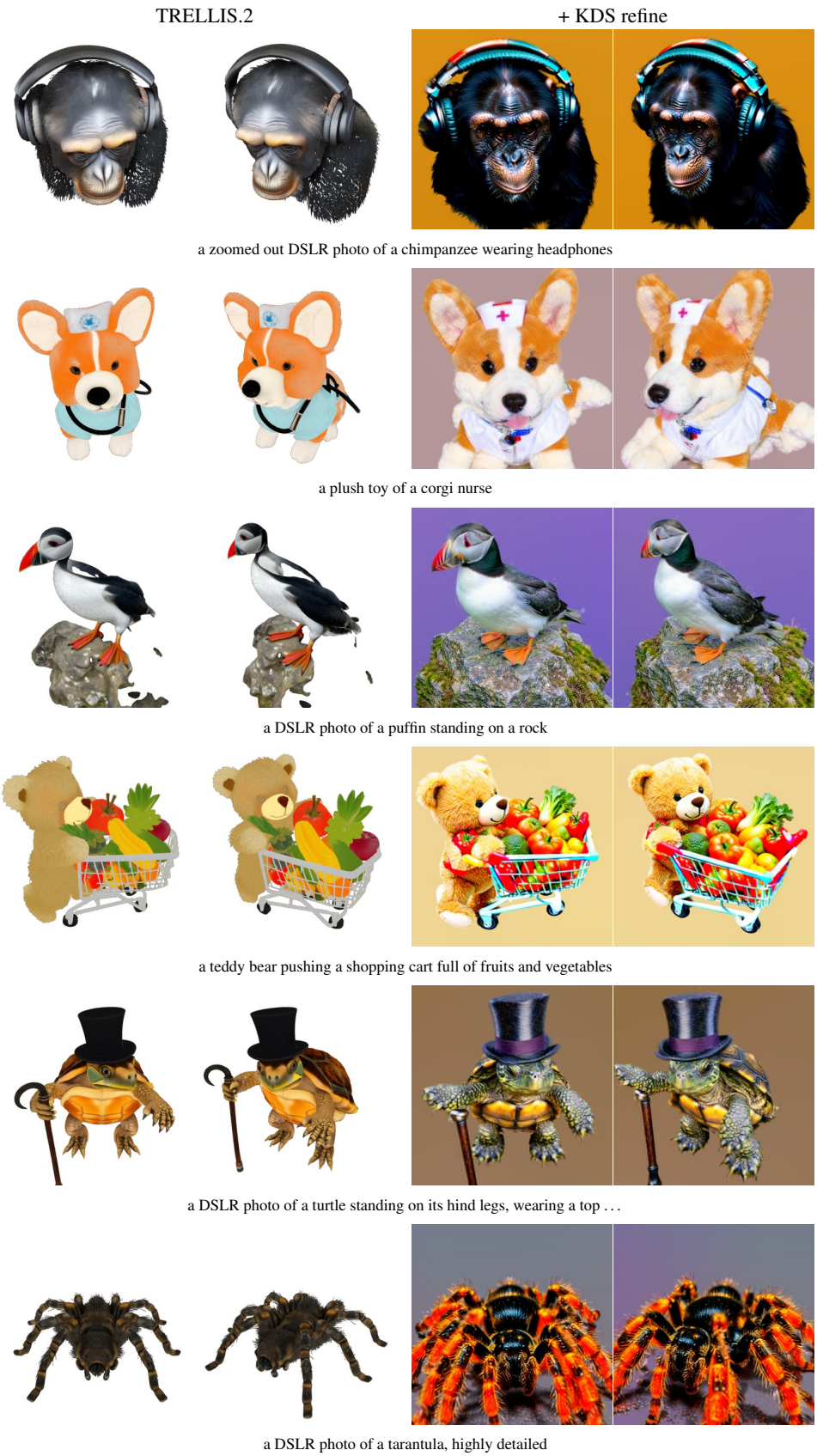

Figure 11: KDS refinement of TRELLIS.2 initializations, two nearby views of the initialization (left) and of the refined object (right).

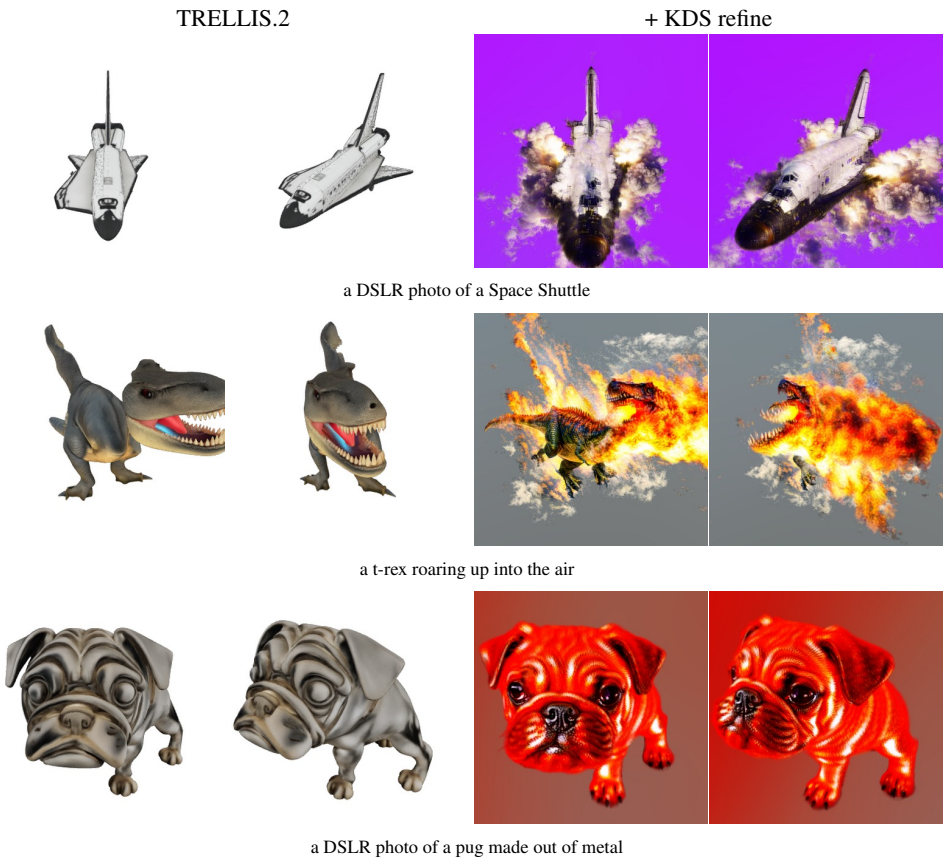

Figure 12: Failure cases of KDS refinement, laid out as in Figure 11. They are much rarer than in generation from scratch (Figure 10).