Distilling Diversity and Control in Diffusion Models

Supplementary Material

A. Theoretical Support

We provide a theoretical justification for the empirical observation that distilled diffusion models lose most of their sample diversity at the *first denoising timestep*. Our argument relies on the standard DDPM forward process, the amplification structure of $\hat{x}_{0|t}$, and the statistical effect of mean-squared-error (MSE) based distillation.

A.1. Preliminaries

Let $x_0 \in \mathbb{R}^d$ be a clean data sample, and consider the forward diffusion process

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{A.1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_s \in (0,1)$.

The reverse model is often parameterized by predicting the noise $\varepsilon_\theta(x_t, t)$. From the forward relation, one can form an estimator of x_0:

$$\hat{x}_{0|t} \;=\; \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}. \tag{A.2}$$

Equation (A.2) will be central to our analysis, as it determines how prediction errors or randomness in ε_θ propagate into variability in $\hat{x}_{0|t}$.

A.2. Sensitivity

Lemma 1 (Sensitivity)**.** *Let $\Delta\varepsilon$ denote a perturbation in the noise prediction at timestep t. Then the induced change in $\hat{x}_{0|t}$ is*

$$\Delta\hat{x}_{0|t} \;=\; -\sqrt{\tfrac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\,\Delta\varepsilon. \tag{A.3}$$

Proof. Differentiate (A.2) with respect to ε:

$$\frac{\partial \hat{x}_{0|t}}{\partial \varepsilon} \;=\; -\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\, I.$$

Multiplying by $\Delta\varepsilon$ yields the result. □

The amplification factor

$$A_t \;=\; \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} \tag{A.4}$$

quantifies how strongly prediction variability at timestep t is magnified in the clean-sample estimate. Since $\bar{\alpha}_t \ll 1$ at early timesteps, A_t is very large.

A.3. Distillation and Conditional Variance

Distilled diffusion models are typically trained to minimize an MSE-style loss between student and teacher outputs. A standard fact from estimation theory is that the MSE minimizer of a random target is the conditional mean:

$$s^*(x) = \mathbb{E}[Y \mid X = x]. \tag{A.5}$$

Thus, when the teacher output Y has conditional variance $\mathrm{Var}(Y \mid X)$, the student collapses this variance and learns to reconstruct the mean.

By the law of total variance,

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y \mid X]) + \mathbb{E}[\mathrm{Var}(Y \mid X)]\,. \tag{A.6}$$

MSE distillation removes the second term, reducing sample diversity by precisely $\mathbb{E}[\mathrm{Var}(Y \mid X)]$.

$$\Delta\,\mathrm{Var} = \mathrm{Var}(Y) - \mathrm{Var}(s^*(x)) = \mathbb{E}[\mathrm{Var}(Y \mid X)]\,. \tag{A.7}$$

A.4. Amplification of Diversity Loss at Early Timesteps

Let the teacher produce a stochastic noise prediction $\varepsilon_{\mathrm{T}}(x_t, t; \xi)$, where ξ captures randomness in sampling. Then, from Lemma 1,

$$\mathrm{Var}(\hat{x}_{0|t} \mid x_t) \;=\; \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\;\mathrm{Var}(\varepsilon_{\mathrm{T}}(x_t,t) \mid x_t). \tag{A.8}$$

Taking expectation over x_t,

$$\mathbb{E}_{x_t}\big[\,\mathrm{Var}(\hat{x}_{0|t} \mid x_t)\big] \;=\; \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\;\mathbb{E}_{x_t}\big[\,\mathrm{Var}(\varepsilon_{\mathrm{T}}(x_t,t) \mid x_t)\big]. \tag{A.9}$$

Proposition 1 (Amplified Diversity Loss)**.** *For an MSE-trained student, the reduction in total variance of $\hat{x}_{0|t}$ due to distillation satisfies*

$$\Delta\,\mathrm{Var} \;\geq\; \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\;\mathbb{E}_{x_t}\big[\,\mathrm{Var}(\varepsilon_{\mathrm{T}}(x_t,t) \mid x_t)\big]. \tag{A.10}$$

Proof. Direct application of the law of total variance to the random variable $\hat{x}_{0|t}$, substituting from (A.9). □

A.5. Main Result

Theorem 1 (First-Timestep Dominance)**.** *Let $\bar{\alpha}_t$ denote the cumulative product of noise schedule parameters. Then for small $\bar{\alpha}_t$ (early timesteps), the amplification factor $(1-\bar{\alpha}_t)/\bar{\alpha}_t$ is maximal. Consequently, the diversity loss $\Delta\,\mathrm{Var}$ induced by MSE-based distillation is largest at the earliest timesteps. In particular, the first denoising step dominates the reduction of sample diversity in distilled diffusion models.*

Proof. Since $\bar{\alpha}_t$ is monotonically increasing in t with $\bar{\alpha}_0 \approx 0$, the fraction $(1-\bar{\alpha}_t)/\bar{\alpha}_t$ is strictly decreasing in t. Hence the bound on $\Delta\,\mathrm{Var}$ is largest at the earliest timestep, completing the proof. □

This proof applies in cases in which the teacher uses a stochastic $\epsilon_T(x_t, t; \xi)$ to create variance in its output. This is the case for the ADD distillation method [32] that is used to train SDXL-Turbo, which uses stochastically sampled teacher for reconstruction loss. The variance trade-off in such cases is also analyzed in previous works [11].

We have also tested our methods on models that do not use a stochastic teacher, such as DMD [39], which uses the Distribution Matching distillation method which uses a deterministic teacher, in which the teacher's $\mathrm{Var}(Y_t \mid x_t, t) = 0$. Despite this difference, we still measure a large drop in sample diversity concentrated at the first timestep. We hypothesize that in these cases, the loss in diversity is due to sparse sampling of the teacher relative to the large diversity of text prompts, which allows the student to collapse to sparsely-sampled modes in a similar way as seen in the stochastic teachers, despite the deterministic teachers' theoretical diversity: for example in DMD only 100k text-conditioned teacher samples are used. The loss in diversity due to sparse sampling of a teacher would be amplified in the early timesteps due to Eq (A.4).

B. Control Distillation: Reverse Transfer

In the main paper, we demonstrated that control mechanisms trained on base models can be seamlessly transferred to distilled models. Here, we present additional results for the reverse direction: transferring control mechanisms trained on distilled models to base models. This bidirectional transfer capability further validates our hypothesis that concept representations are preserved during the distillation process.

We note that while most control mechanisms transferred effectively, we encountered difficulties training LoRA adaptations on LCM due to its specialized architecture and training procedure. These challenges highlight potential avenues for future research in developing more universally transferable control mechanisms.

C. Skip Step Approach

In the main paper, we introduced a resource-efficient alternative to our hybrid approach: skipping the first timestep altogether in distilled model inference. We provide additional qualitative comparisons between this approach and our hybrid method in Figure C.1.

The skip-first-step approach provides a reasonable compromise when resource constraints prevent loading both models simultaneously. However, our quantitative analysis in the main paper and these qualitative examples demonstrate that the hybrid approach consistently achieves superior results in terms of both diversity and quality.

D. Generalization Across Model Backbones

To assess the generality of our findings, we extend our analysis to different diffusion model architectures beyond SDXL. We evaluate two additional model pairs: PixArt-Alpha (base) with PixArt-Delta (distilled), and SD 2.1 (base) with SD-Turbo (distilled).

Table D.1 shows that the diversity collapse phenomenon and the effectiveness of our solution generalize across different model architectures. PixArt-Delta exhibits similar sample diversity reduction compared to PixArt-Alpha, with our hybrid approach restoring diversity while maintaining efficiency. Similarly, SD-Turbo shows reduced sample diversity compared to SD 2.1, which our method successfully addresses. We show qualitative results in Figure D.1

Model	Architecture	Sample Diversity	Time (s)	Our Method
PixArt-Alpha	DiT	0.342	4.1	-
PixArt-Delta	DiT	0.198	0.9	0.339
SD 2.1	UNet	0.298	3.2	-
SD-Turbo	UNet	0.171	0.7	0.294

Table D.1. Sample diversity (DreamSim distance) across different model architectures for the prompt "image of a car" across 100 samples. Our hybrid approach consistently restores diversity regardless of the underlying architecture (DiT vs UNet) or distillation method. Please refer to Fig D.1 for qualitative samples

The consistent pattern across DiT-based (PixArt) and UNet-based (SD 2.1/Turbo) architectures demonstrates that the first-timestep diversity bottleneck is a fundamental characteristic of diffusion distillation, not specific to particular model designs.

E. Causal Validation: Testing Later Timesteps

To strengthen our causal claim that the first timestep is the critical bottleneck, we conduct a complementary experiment: replacing the *final* timesteps of distilled models with base model steps while keeping the first timestep from the distilled model.

Table E.1 shows that replacing the final timesteps does not yield diversity improvements compared to our first-timestep intervention. This result provides strong causal evidence that the diversity bottleneck is concentrated at the beginning of the generation process, not distributed throughout the timesteps. The minimal improvement from modifying later steps confirms our $\hat{\mathbf{x}}_0$ analysis: once the structural decisions are made in the first timestep, later steps primarily refine details rather than introduce fundamental variations. We show qualitative examples in Figure E.1

Figure B.1. Reverse Control Transfer: Control mechanisms (Custom Diffusion [17] and Concept Sliders [8]) trained on distilled models can be effectively transferred to base models without retraining. This bidirectional transferability confirms that concept representations are preserved during diffusion distillation. Note: LCM LoRA transfers were excluded due to training difficulties with the LCM architecture.

Method	Sample Diversity	Time (s)
SDXL-Base	0.357	9.22
SDXL-DMD2	0.264	0.64
Replace First Timestep (Ours)	0.350	0.64
Replace Final Timesteps	0.251	6.61

Table E.1. Causal validation experiment. Replacing final timesteps with base model steps provides minimal diversity improvement (DreamSim distance) compared to our first-timestep intervention, confirming that the first timestep is the critical bottleneck. The analysis is done for the prompt "image of a car" across 100 samples.

F. Extended $\hat{x}_0$ visualization Analysis

The main paper introduced $\hat{x}_0$ visualization technique for analyzing how diffusion models develop structural information during the denoising process. We present additional visualizations in Figures F.1, F.2 that further illuminate the differences between base and distilled models.

These visualizations reinforce our key finding: distilled models compress the diversity-generating behavior distributed across early timesteps in base models into a single initial step, explaining the observed mode collapse. This insight directly informed our hybrid inference approach, which strategically leverages the diversity-generating capabilities of base models in critical early steps.

G. Mode Collapse and Diversity

The main paper introduced our finding that distilled diffusion models suffer from reduced sample diversity (mode collapse) compared to their base counterparts. We provide additional qualitative examples in Figure G.1-G.4 that visually demonstrate this phenomenon across various prompts

Figure C.1. Qualitative comparison between (left) our hybrid approach, (right) skip-first-step approach. The skip-first-step approach improves diversity over the standard distilled model but exhibits reduced quality compared to our hybrid method, particularly in fine details and coherence.

and model variants.

These examples highlight the significant diversity loss in distilled models. While the distilled models produce high-quality images, they often converge to similar structural compositions regardless of random seed initialization. Our diversity distillation approach effectively addresses this limitation, restoring the variety of outputs comparable to the base model while maintaining computational efficiency.

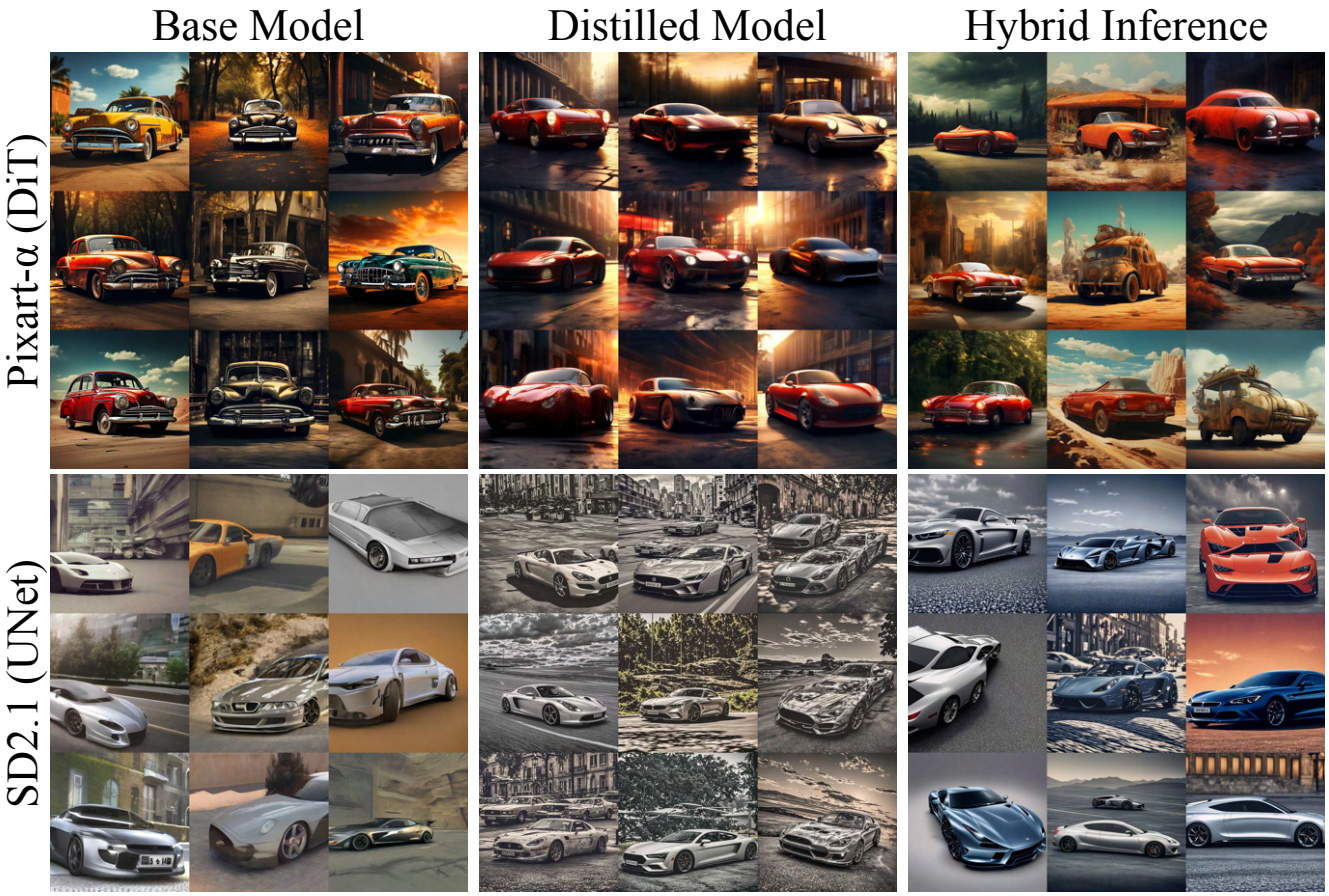

Figure D.1. **Generalization across model architectures.** Sample diversity comparison for the prompt "image of a car" across different diffusion model architectures. Top row: PixArt-α (DiT-based base model) shows diverse car types, colors, and contexts, while PixArt-δ (distilled) produces similar red sports cars with repetitive compositions. Bottom row: SD 2.1 (UNet-based base model) generates varied car styles and settings, while SD-Turbo (distilled) exhibits reduced diversity with similar silver/white cars in repetitive urban contexts. Our hybrid inference approach restores diversity in both architectures, demonstrating that the first-timestep bottleneck is architecture-agnostic.

Figure E.1. **Causal validation of first-timestep importance.** Visual comparison for the prompt "image of a car" showing: (left) SDXL-Base with diverse car types, colors, and contexts; (middle-left) SDXL-DMD with reduced diversity showing similar red sports cars; (middle-right) our hybrid approach using base model for first timestep (T=0) then DMD for remaining steps, successfully restoring diversity; (right) control experiment using DMD for first timestep (T=0) then base model for remaining steps, showing minimal diversity improvement. This demonstrates that the first timestep, not later steps, controls sample diversity.

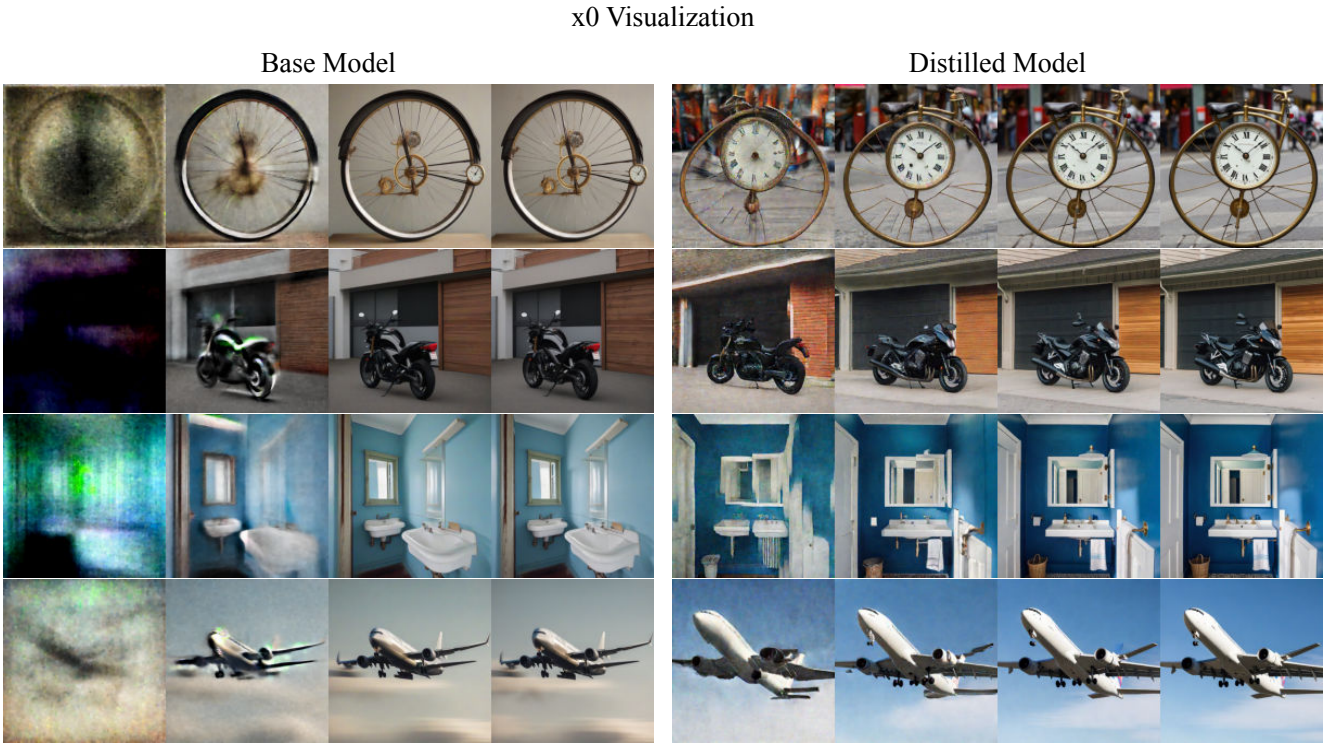

Figure F.1. Extended $\hat{\mathbf{x}}_0$ visualization comparison between SDXL-Base and SDXL-DMD for the prompt. The visualization reveals that DMD commits to final structural composition within the first timestep, while Base gradually develops structure across multiple steps. This pattern is consistent across different content types and prompts.

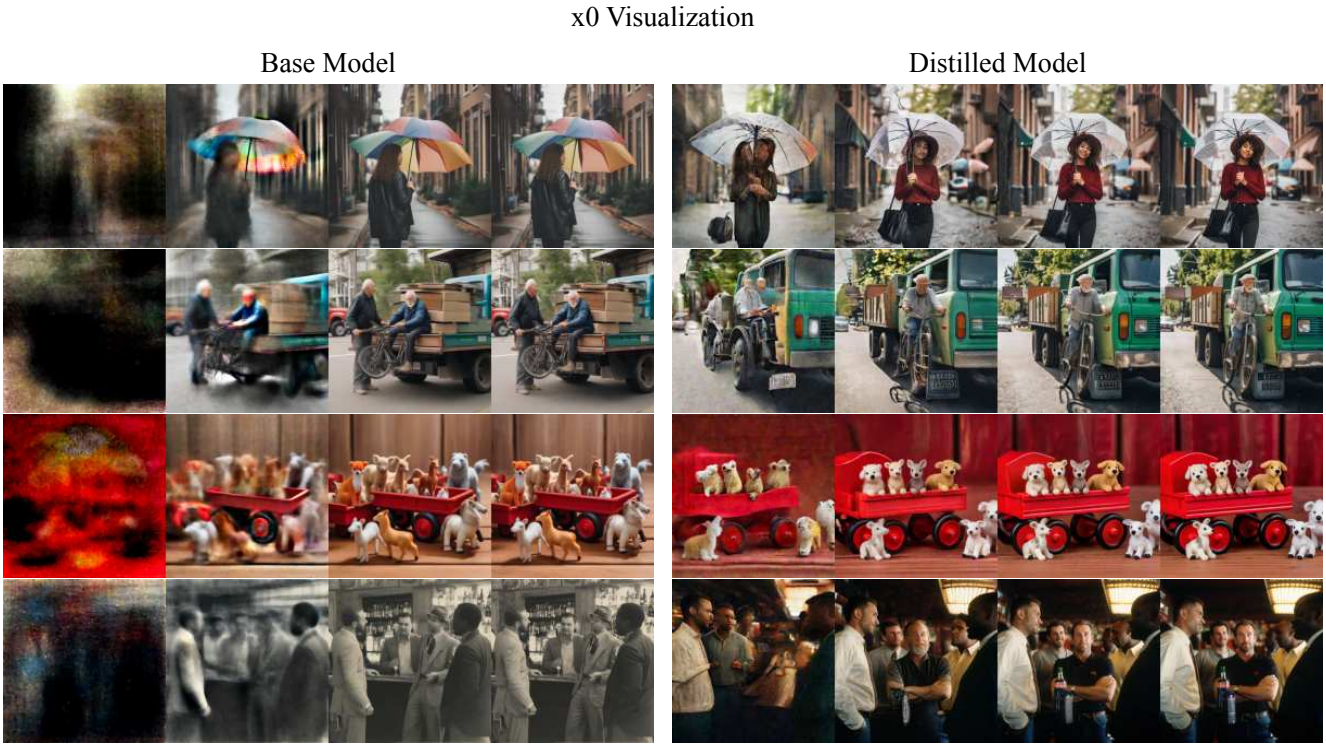

Figure F.2. Extended $\hat{\mathbf{x}}_0$ visualization comparison between SDXL-Base and SDXL-DMD for the prompt. The visualization reveals that DMD commits to final structural composition within the first timestep, while Base gradually develops structure across multiple steps. This pattern is consistent across different content types and prompts

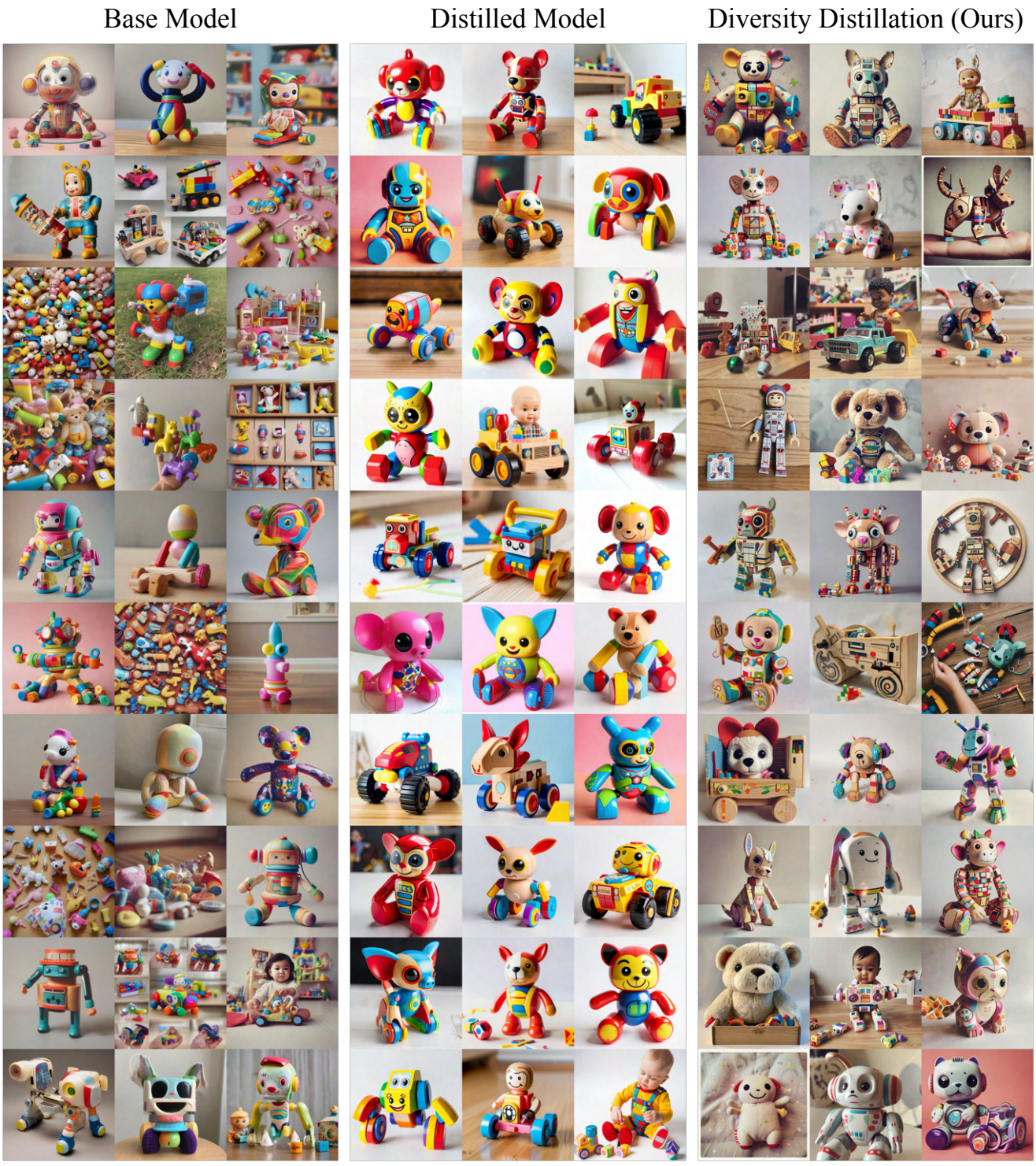

Figure G.1. Comparison of generation diversity across different models for the prompt ”image of a toy.” Each image shows different seeds for the same model. Note the structural similarity in distilled model outputs compared to the greater variation in base model and our hybrid approach.

Figure G.2. Comparison of generation diversity for ”image of a flower” Distilled models (middle column) produce structurally similar outputs across different seeds, while our approach (right column) restores diversity comparable to the base model (left column) while maintaining the speed advantage of distilled models.

Figure G.3. Additional diversity comparison for ”city street”Distilled models (middle column) produce structurally similar outputs across different seeds, while our approach (right column) restores diversity comparable to the base model (left column) while maintaining the speed advantage of distilled models.

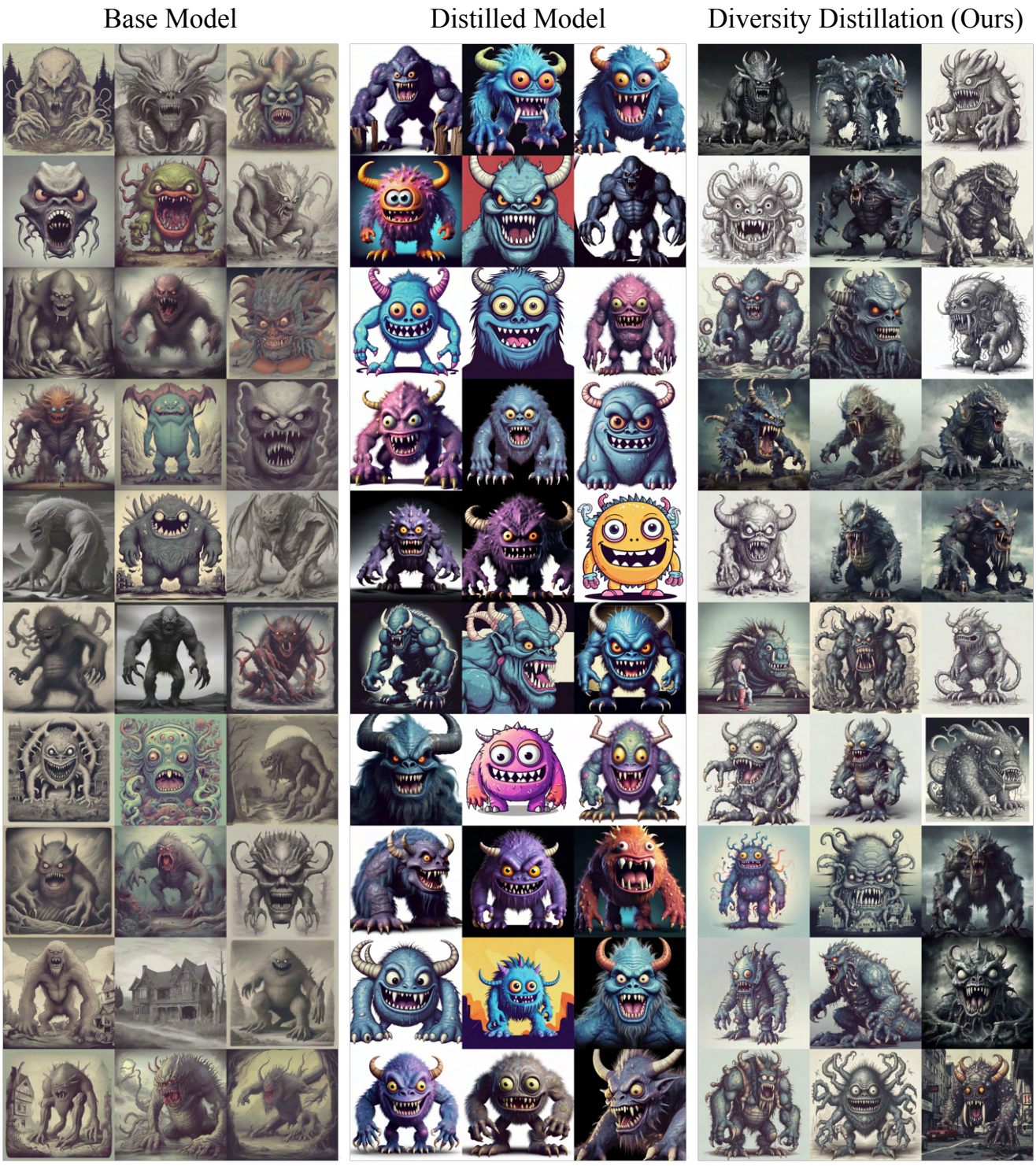

Figure G.4. Diversity comparison for abstract prompt: ”picture of a monster” Distilled models (middle column) produce structurally similar outputs across different seeds, while our approach (right column) restores diversity comparable to the base model (left column) while maintaining the speed advantage of distilled models.