

Distilling Diversity and Control in Diffusion Models

Rohit Gandikota*

David Bau

Northeastern University

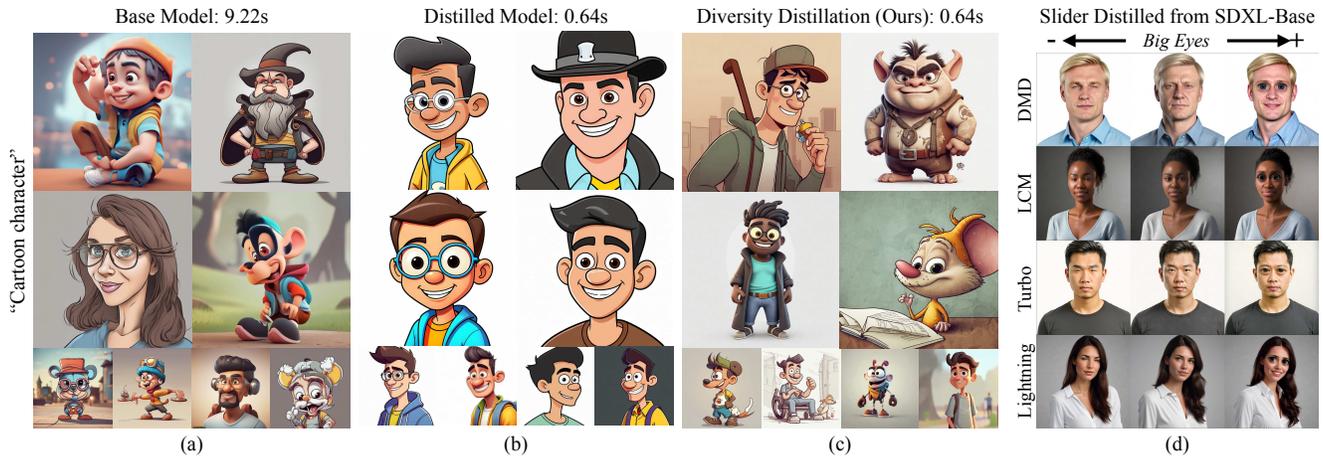


Figure 1. **Diversity Distillation:** (a) SDXL-Base is very slow (9.22s) but has good *sample diversity* for the prompt “Cartoon character”, sampling a wide range of styles, creatures, backgrounds, and poses. (b) SDXL-DMD2 is fast (0.64s) but sacrifices diversity. With the same prompt, samples all have the same style, pose, species, and context. (c) We show how the diversity of the base model can be distilled into the fast model by substituting the first timestep, achieving both speed and diversity (0.64s). **Control Distillation:** (d) Despite the lack of sample diversity in distilled models, control mechanisms like Concept Sliders trained on base models transfer perfectly to distilled variants, demonstrating that the representational structure for diversity exists but is not spontaneously activated during generation.

Abstract

Distilled diffusion models generate images in far fewer timesteps but suffer from reduced sample diversity when generating multiple outputs from the same prompt. To understand this phenomenon, we first investigate whether distillation damages concept representations by examining if the required diversity is properly learned. Surprisingly, distilled models retain the base model’s representational structure: control mechanisms like Concept Sliders and LoRAs transfer seamlessly without retraining, and Slider-Space analysis reveals distilled models possess variational directions needed for diversity yet fail to activate them. This redirects our investigation to understanding how the generation dynamics differ between base and distilled models. Using \hat{x}_0 trajectory visualization, we discover distilled models commit to their final image structure almost immediately at the first timestep, while base models distribute structural decisions across many steps. To test whether this first-step commitment causes the diversity loss, we introduce diversity distillation, a hybrid approach using the base

model for only the first critical timestep before switching to the distilled model. This single intervention restores sample diversity while maintaining computational efficiency. We provide both causal validation and theoretical support showing why the very first timestep concentrates the diversity bottleneck in distilled models. Our code and data are available at distillation.baulab.info

1. Introduction

Distilled diffusion models generate images in far fewer timesteps but lack the sample diversity of their original base model counterparts. In this paper we ask: *Why do distilled diffusion models collapse in sample diversity?*

This limitation fundamentally constrains practical applications. When users generate multiple images from the same prompt, they expect varied outputs [24] that explore different creative interpretations. Figure 1 illustrates this

*Correspondence to gandikota.ro@northeastern.edu

problem: while base models produce diverse structural compositions across random seeds, distilled variants converge to visually similar results despite their computational advantages. This diversity collapse limits the utility of efficient models in creative workflows, design exploration, and applications requiring multiple candidate generations [24].

The challenge is particularly puzzling given recent advances in distillation quality. Diffusion models demonstrate unprecedented generation quality [4, 14, 19, 27, 28], yet their computational demands create deployment barriers. Modern distillation techniques [19, 20, 23, 32, 38, 39] have successfully maintained image quality while reducing inference steps from 50-100 to just 1-4. Some distilled models even achieve better distributional diversity than their base counterparts. However, distributional diversity is different from sample diversity. Distributional diversity is the ability to cover the full spectrum of training data across varied prompts, while sample diversity is the ability to generate diverse images for a single prompt under different seeds. SDXL-DMD2 [38] shows superior distributional diversity with lower FID [13] scores on COCO [21] datasets but exhibits poor sample diversity, making the collapse even more mysterious.

To solve this puzzle, we first investigate whether distillation damages the model’s concept representations by analyzing the top variation directions [10] in base models and examining if these directions exist in distilled models. Surprisingly, we find that distilled models retain the variational directions needed for diversity—they simply fail to activate them during generation. This leads us to hypothesize that timestep dynamics could be the reason, supported by theoretical analysis showing how distillation collapses the output variance at early timesteps. We test this by visualizing what models predict at intermediate timesteps [14, 36], revealing that distilled models commit to their final structure almost immediately at the first step while base models distribute decisions across many steps. To causally test our hypothesis, we introduce a hybrid inference technique and demonstrate that sample diversity can be drastically improved by simply modifying the first timestep of distilled models.

Our investigation reveals that the problem lies not in *what* distilled models learn, but in *how* they generate. Through careful analysis of model representations and generation dynamics, we identify the root cause and demonstrate a simple solution that restores sample diversity without sacrificing efficiency. Our findings challenge conventional assumptions about the diversity-efficiency tradeoff and provide actionable insights for both researchers developing distillation methods and practitioners deploying efficient diffusion models.

2. Related Works

Diffusion Distillation: While diffusion models [14, 33, 34] excel at high-quality image synthesis, their requirement for 20-100 sampling steps creates significant computational bottlenecks. Diffusion distillation techniques address this limitation by finetuning base models that maintain quality with fewer steps. Progressive distillation [30] established the foundation by iteratively training student models to match teacher outputs with half the sampling steps. Recent approaches have further improved efficiency through distinct methodologies: Adversarial Diffusion Distillation [32], implemented in SDXL-Turbo, integrates score distillation with adversarial training to enable high-fidelity generation in just 1-4 steps, effectively combining diffusion guidance with GAN-like discriminators. Distribution Matching Distillation [38], featured in SDXL-DMD2, takes a different approach by focusing on matching output distributions rather than specific trajectories, eliminating regression loss and implementing a two time-scale update rule that significantly improves training stability. For balancing quality and mode coverage, Progressive Adversarial Diffusion Distillation [20] in SDXL-Lightning employs staged training with specialized latent-space discriminators, offering flexibility through checkpoints optimized for 1-8 step inference. Latent Consistency Models [23], applied in SDXL-LCM, ensure consistency in latent representations across noise levels for distillation, reducing steps to 4-8 while preserving generation quality. Despite these advances in efficiency, the relationship between model distillation and sample diversity has remained largely unexplored.

Concept Representation: Research in concept representation for diffusion models has evolved from basic personalization to sophisticated control mechanisms [2, 3, 8, 25, 37, 40]. Textual Inversion [6] captures the semantics of a concept with learnable embeddings in text space without modifying model weights, allowing personalization with just a few images. DreamBooth [29] advanced this approach by fine-tuning models with unique identifiers and a specialized prior preservation loss. Custom Diffusion [17] streamlined this process by optimizing only cross-attention layers, reducing storage requirements to just 3% of model weights while enabling multi-concept customization simultaneously. For precise attribute manipulation, Concept Sliders [8] introduced low-rank adaptors that create interpretable controls over specific visual attributes like age or weather conditions. This technique was expanded in SliderSpace [10], which decomposes a model’s visual variation capabilities, i.e. sample diversity, into multiple controls from a single prompt, enhancing creative exploration. Recent works have addressed the issue of suboptimal mode following in finetuned models by implementing an inference-time guidance annealing [15]. Complementary to

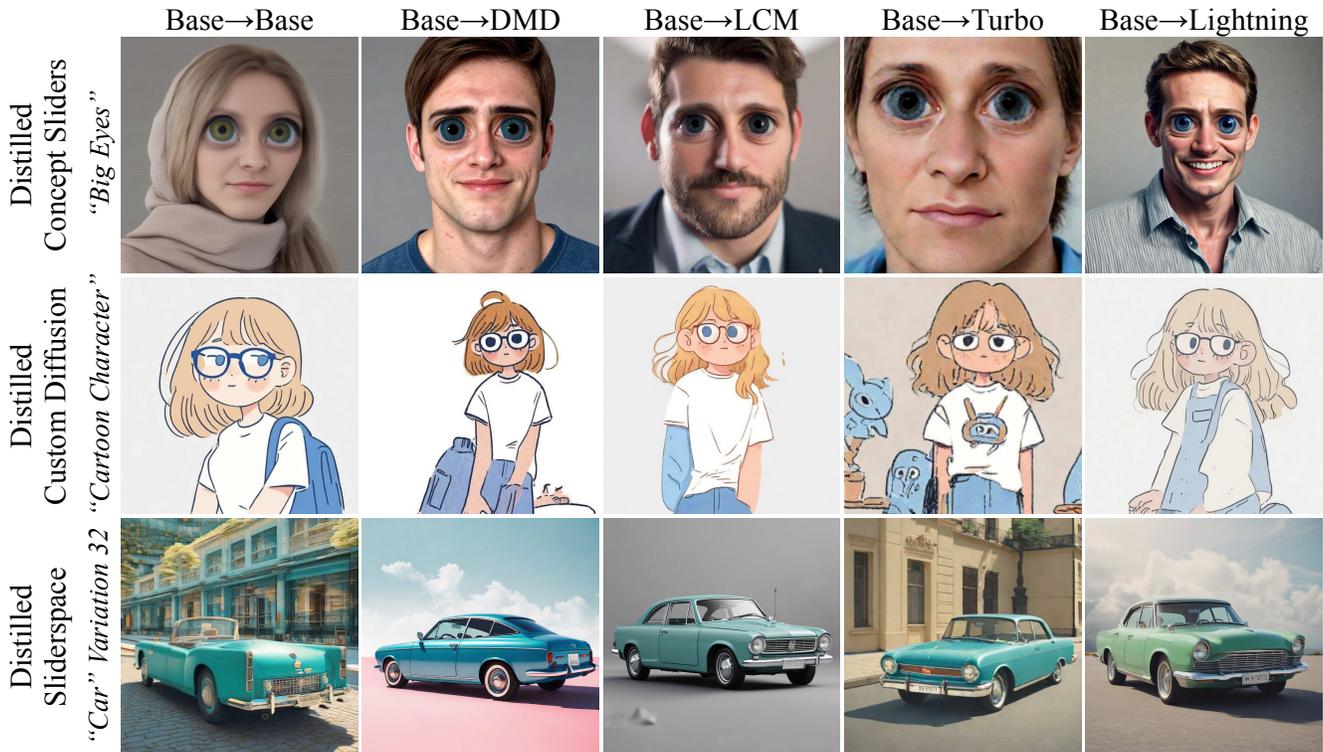


Figure 2. Control directions (Sliders [17]), customization adapters (Custom Diffusion [8]), and variational directions (Sliderspace [10]) trained on SDXL-Base transfer to all distilled models without additional finetuning. SliderSpace results suggest that top variation directions that capture sample diversity in the base model exist in distilled models but are not spontaneously activated during generation.

these control mechanisms, hierarchical concept trees [1, 35] were developed to enable intuitive exploration of related visual concepts. Recent work has also addressed ethical concerns through targeted concept removal techniques by editing selective weights [7, 9, 22], redirecting concept representations [16, 26]. Since distillation modifies the UNet model of diffusion, in this work, we mainly focus on custom concept and control representations that are captured in UNet modules. Our work uniquely explores whether such control mechanisms can be distilled from base to efficient models without additional training.

3. Control Distillation

Having established that distilled models suffer from reduced sample diversity (Figure 1), we investigate the underlying cause. Our first hypothesis is that distillation erodes the model’s concept space: that the distilled model fails to learn or retain the representations necessary for diverse generation. If this were true, the model would naturally produce less varied outputs because the required conceptual building blocks would be missing or damaged.

To test this hypothesis, we examine whether the representational structure needed for diversity is preserved in distilled models. We pose this as *representational compat-*

ibility: do the variation directions and control mechanisms present in base models transfer to distilled models without retraining?

3.1. Experimental Setup

We test three complementary families of controls that probe different aspects of model representations:

Concept Sliders [8, 10] provide fine-grained control over visual attributes (e.g., age, weather, eye size) through low-rank adaptations. These test whether distilled models preserve the semantic understanding needed to manipulate specific visual properties.

Customization mechanisms including Custom Diffusion [17] and DreamBooth [29] capture nuanced, user-defined concepts through specialized training procedures. These mechanisms test whether distilled models retain the capacity to encode and recall complex, personalized concepts.

SliderSpace [10] decomposes the seed-induced variations of a base diffusion model into interpretable, continuous directions—each corresponding to a controllable visual factor (e.g., texture, lighting, layout). This directly analyzes whether the variational components responsible for sample diversity are present in distilled models.

Method	Concept	Base→Base	Base→DMD	Base→LCM	Base→Turbo	Base→Lightning
Concept Sliders [8]	Age	20.4	17.8	27.1	19.0	24.8
	Smile	19.7	21.4	19.5	33.5	14.0
	Muscular	34.6	26.7	33.8	39.0	33.2
Customization [17, 29]	Lego	32.2	26.8	26.0	30.3	29.7
	Watercolor style	34.3	31.4	29.6	27.5	39.2
	Crayon style	32.7	27.8	24.7	29.5	32.5
Sliderspace [10] (Top n th Variation)	Direction 1	29.3	28.1	24.1	27.7	22.4
	Direction 16	32.9	32.0	33.4	28.1	29.9
	Direction 32	30.7	28.9	29.3	30.5	31.4

Table 1. We show the percentage change in CLIP score from the original image and the edited image. Higher values indicate stronger attribute change or style transfer. Control effectiveness is largely preserved when transferring from base to different distilled models, with only minor variations across distillation techniques. Importantly, SliderSpace directions for the concept “car” which capture the base’s natural variational structure transfer. This demonstrates that the representational components needed for diversity exist in distilled models.

For each mechanism, we perform bidirectional transfer experiments: training on base models and applying to distilled models, and vice versa. We study four SDXL distilled families—SDXL-Turbo [32], SDXL-Lightning [20], SDXL-LCM [23], and SDXL-DMD2 [38]—which span diverse distillation procedures. Additional training details and experiments on additional base models (SD 2.1 [28] and DiT-backed PixArt [4]) are provided in the Appendix D.

3.2. Results

Our experiments reveal surprising evidence that contradicts our initial hypothesis. Figure 2 demonstrates seamless transfer of control mechanisms across model variants. For example, a “comical big eyes” slider trained on SDXL effectively controls SDXL-Turbo’s generations, despite Turbo requiring only 1-4 steps compared to SDXL’s 20-100 steps.

Table 1 quantifies this compatibility through CLIP scores [12]. Transfer effectiveness remains consistently high across all tested combinations, with positive score changes indicating successful attribute manipulation. While absolute scores vary due to different guidance scales across models, the consistent positive changes confirm that concept representations are preserved during distillation. We show more experiments for the Distillation→Base evidence in Appendix B.

Most tellingly, SliderSpace [10] analysis reveals that variation directions learned on base models transfer perfectly to distilled variants without retraining (Tab 1 and Fig. 2). This indicates that the factors needed for diverse generation are present and accessible in distilled models. However, during free generation with different random seeds, distilled models rarely activate these factors.

These results directly contradict our initial hypothesis that distillation destroys concept representations. The concepts remain encoded and controllable in distilled models; what fails is their spontaneous activation during few-step sampling. This finding redirects our investigation from

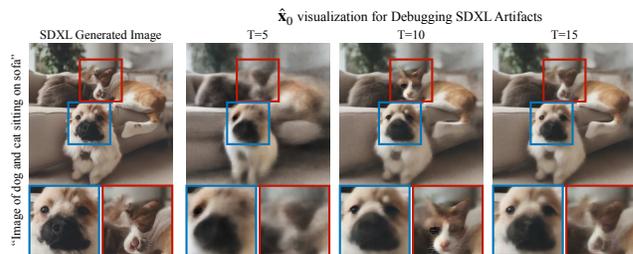


Figure 3. \hat{x}_0 visualization reveals generation inconsistencies. When prompted with “Image of dog and cat sitting on sofa,” the SDXL model produces an image with only a dog. However, \hat{x}_0 visualization at $T = 10$ shows the model initially conceptualizing a cat face (red box) before abandoning this element in the final generation. This demonstrates how diffusion models can discard semantic elements during the denoising process.

what is missing to why existing diversity mechanisms fail to activate. If the required variational directions exist but are not used, the problem must lie in the generation dynamics rather than the representational capacity.

4. \hat{x}_0 : Visualizing Intermediate Latents

Having established that distilled models possess the representational components needed for diversity, we must understand why these components fail to activate during generation. Our investigation shifts from *what* distilled models learn to *how* they generate. To analyze the generation dynamics, we employ \hat{x}_0 trajectory visualization as a diagnostic tool.

In the standard diffusion formulation [14, 33], \mathbf{x}_0 represents the clean image that is progressively corrupted with noise over T timesteps. During reverse generation, at each timestep t , the model predicts noise $\epsilon_\theta(\mathbf{x}_t, t)$ to compute the next denoising step:

Let \mathbf{x}_0 be an initial image and \mathbf{x}_T be pure Gaussian noise. The forward diffusion process gradually adds noise

according to a variance schedule β_t , with corresponding noise level parameters $\alpha_t = 1 - \beta_t$ and cumulative parameters $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The generative process aims to reverse this diffusion, starting from \mathbf{x}_T and progressively denoising to reconstruct \mathbf{x}_0 . At timestep t , the model predicts noise $\epsilon_\theta(\mathbf{x}_t, t)$ to compute the next step:

$$\mathbf{x}_{t-1} = \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}} \quad (1)$$

Using this predicted noise ϵ_θ , we can estimate what the model believes the final clean image will be at any intermediate timestep t , by extrapolating the same noise prediction all the way to \mathbf{x}_0 :

$$\hat{\mathbf{x}}_{0|t} = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \quad (2)$$

Visualizing these $\hat{\mathbf{x}}_{0|t}$ trajectories across timesteps reveals when structural decisions are made and how they evolve during generation [36]. This technique allows us to peer into the model’s “thought process” and understand the temporal dynamics of image formation.

4.1. $\hat{\mathbf{x}}_0$ for Investigating Generation Artifacts

Before applying this technique to our diversity investigation, we demonstrate its utility for understanding generation artifacts. In Figure 3, when prompted with “Image of dog and cat sitting on sofa,” the SDXL model produces a final image containing only a dog’s face. However, $\hat{\mathbf{x}}_{0|t}$ visualization at early timesteps ($T = 10$) reveals that the model initially conceptualized a cat face (highlighted in red box) before abandoning this element in later steps. This insight exposes how diffusion models can “change their mind” during generation, sometimes discarding semantic elements present in the prompt.

4.2. $\hat{\mathbf{x}}_0$ for Investigating Sample Diversity

We now apply $\hat{\mathbf{x}}_0$ visualization to compare base and distilled models under identical conditions: same seed and prompt (“image of a dog”). Standard visualizations of intermediate latents \mathbf{x}_t (Fig 4.a) show only subtle differences across base and distilled models. In stark contrast, the $\hat{\mathbf{x}}_0$ trajectories (Fig 4.b) expose a dramatic pattern that could potentially explain the diversity collapse.

Distilled models commit to their final image structure almost immediately after the first timestep. Their $\hat{\mathbf{x}}_{0|t}$ predictions quickly converge to the final output. Base models, conversely, distribute structural decision-making across many timesteps, with their $\hat{\mathbf{x}}_{0|t}$ predictions gradually refining and evolving toward the final image.

Figure 5 quantifies this phenomenon by measuring DreamSim distances [5] between intermediate $\hat{\mathbf{x}}_{0|t}$ predictions and final outputs across COCO-10k [21] prompts. The data reveals that distilled models achieve a large fraction of

their final structure after a single timestep, while base models require approximately 30% of their total inference steps to reach comparable structural definition. We provide more qualitative evidence in Appendix F.

Our discovery suggests a testable hypothesis: the first timestep is the primary culprit behind diversity loss in distilled models. We provide theoretical analysis (detailed in Appendix A) showing why the first timestep concentrates the diversity bottleneck. The analysis demonstrates how timestep compression in distillation amplifies decision-making pressure early in the process, with the amplification factor being largest at initial timesteps.

5. Diversity Distillation

Our $\hat{\mathbf{x}}_0$ analysis identified a clear hypothesis: the first timestep is the primary bottleneck causing diversity collapse in distilled models. To test this causally, we design experiments that directly manipulate the first timestep and measure the impact on sample diversity. If our hypothesis is valid, targeted interventions at this critical step should restore diversity without affecting efficiency.

We develop two approaches to test this hypothesis. First, we replace the first timestep of distilled model inference with the corresponding step from the base model. Second, we examine whether simply skipping the problematic first timestep can improve diversity. Both approaches target the identified bottleneck while preserving the computational advantages of distilled models.

Algorithm 1 implements our hybrid inference strategy. The approach uses the base model for the critical first timestep(s) to establish diverse structural foundations, then transitions to the distilled model for efficient completion. This design directly tests whether the first timestep controls diversity while maintaining computational efficiency.

Algorithm 1 Hybrid Inference for Diversity Distillation

Require: Base model f_{base} , distilled model f_{distil} , total timesteps T , transition point k

Ensure: Generated image \mathbf{x}_0

```

1: Initialize  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ 
2: for  $t = T, T - 1, \dots, 1$  do
3:   if  $t > T - k$  then  $\triangleright$  Critical timesteps for diversity
4:      $\mathbf{x}_{t-1} \leftarrow f_{\text{base}}(\mathbf{x}_t, t, \text{prompt})$ 
5:   else  $\triangleright$  Efficient refinement timesteps
6:      $\mathbf{x}_{t-1} \leftarrow f_{\text{distil}}(\mathbf{x}_t, t, \text{prompt})$ 
7:   end if
8: end for
9: return  $\mathbf{x}_0$ 

```

5.1. Experimental Results

We evaluate our approach across two complementary diversity metrics that capture different aspects of model per-

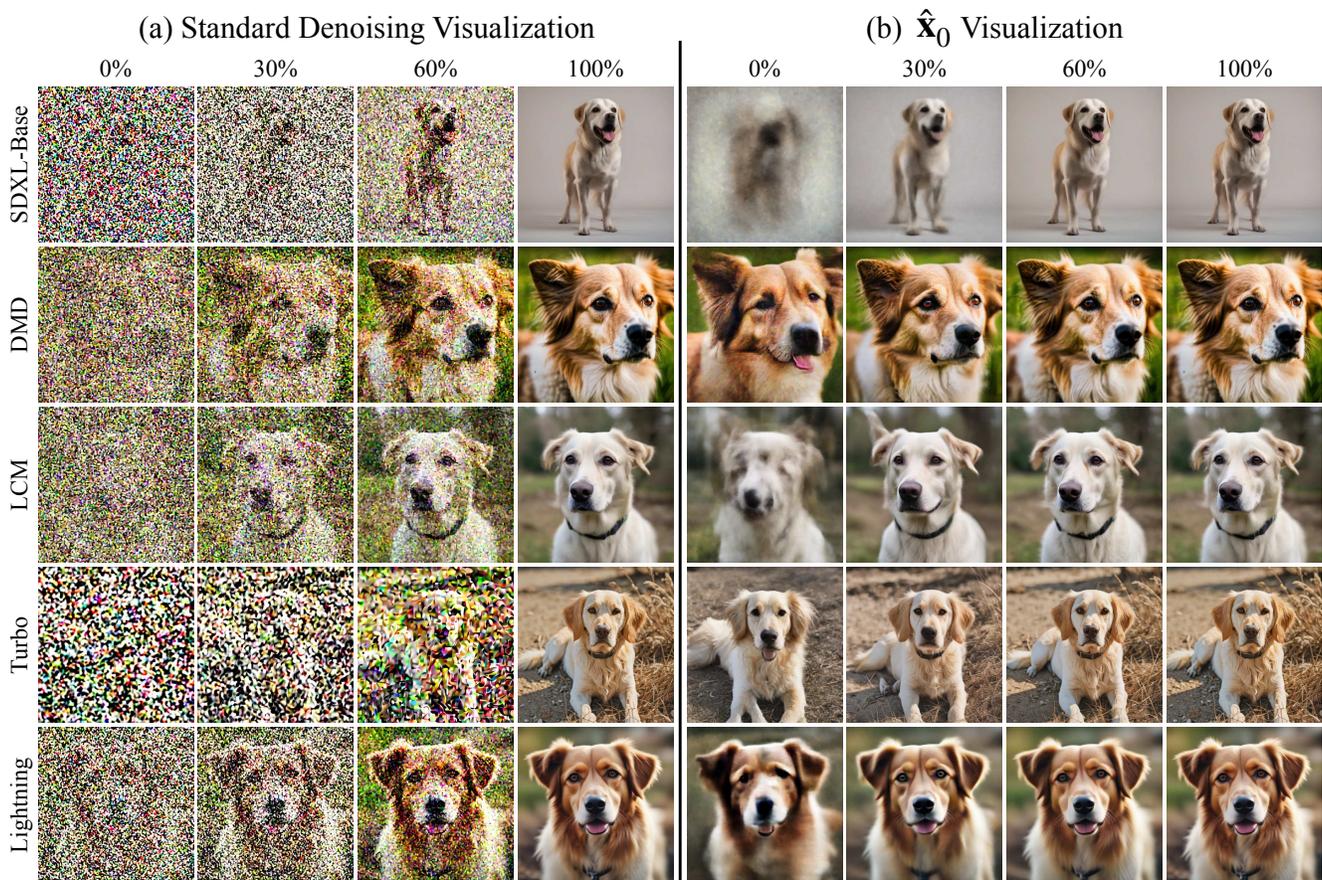


Figure 4. Comparison of standard diffusion visualization vs. \hat{x}_0 visualization. (a) Standard visualization of intermediate latents shows subtle differences between base and distilled models. (b) \hat{x}_0 visualization reveals dramatic differences in how models predict the final output. Distilled models commit to final image structure in the first timestep, while base models gradually refine structure across multiple steps, explaining the observed mode collapse in distilled models.

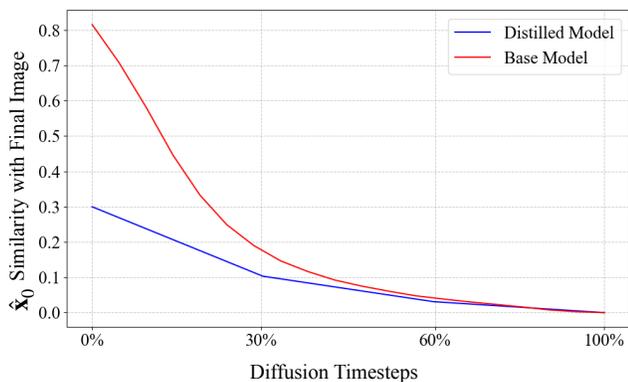


Figure 5. Measuring the dreamsim distance between intermediate \hat{x}_0 visualization and final generated image reveals that distilled models establish structural image composition within the initial diffusion step, whereas base models require approximately 30% of steps to achieve comparable structural definition.

formance. **Distributional diversity** measures how well generated samples match the real training data distribution across varied prompts, assessed through FID [13] (lower is better). **Sample diversity** measures variation among outputs generated from the same prompt with different random seeds, quantified by average pairwise DreamSim distance [5] (higher is better).

Distributional Diversity. Table 3 demonstrates that our hybrid approach achieves superior distributional diversity compared to both base and distilled models. Measuring FID against the COCO-30k dataset, our method achieves better scores than the base model while maintaining the computational efficiency of distilled inference. This confirms that our intervention does not compromise the model’s ability to handle diverse prompts. We hypothesize the improvement in FID compared to the base model is due to the usage of the distilled model in final timesteps. Prior work suggests that varied guidance levels in final timesteps can affect FID

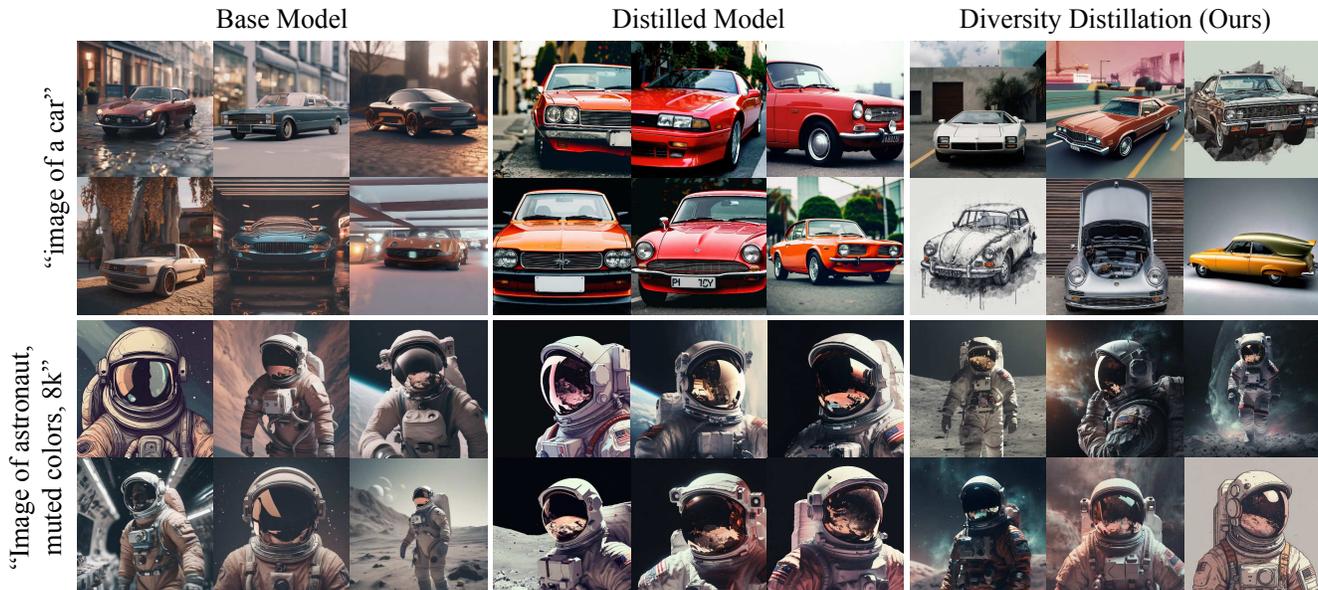


Figure 6. **Visual comparison of generation diversity.** Each row shows three different generations (different random seeds) for the same prompt using: (left) base model, (middle) distilled model, and (right) our diversity distillation approach. Note how the distilled model produces similar car types, poses, and astronaut orientations across seeds, while our approach restores diversity in colors, contexts, and compositions comparable to the base model while maintaining similar inference speed.

Prompt	Base	Distilled	Hybrid (Ours)
Sunset beach	0.396	0.271	0.373
Cute puppy	0.233	0.199	0.265
Futuristic city	0.237	0.198	0.283
Person	0.484	0.347	0.461
Van Gogh art	0.337	0.305	0.366
Average	0.337	0.264	0.350

Table 2. Sample diversity measured by average pairwise DreamSim distance (higher is more diverse). Our hybrid approach restores the lost diversity in distillation with a simple intervention.

metrics [18].

Sample Diversity. Table 2 reveals that our approach dramatically restores sample-level diversity. For each prompt, we generate 100 images with different random seeds and calculate average pairwise DreamSim distances. The hybrid approach restores the diversity lost during distillation.

Figure 6 provides visual confirmation of these quantitative results. The distilled model clearly exhibits reduced structural variety across random seeds, producing similar compositions and layouts. Our hybrid approach successfully restores this diversity, generating varied structural arrangements comparable to the base model while maintaining fast inference speeds.

Method	Steps	FID(↓)	IS(↑)	CLIP(↑)	Time (s)(↓)
Base	50	12.74	24.74	31.83	9.22
Distilled	4	15.52	27.20	31.69	0.64
Hybrid (Ours)	4	10.79	26.13	32.12	0.64
Skip First Timestep	3	10.12	24.69	31.71	0.53

Table 3. Measuring distributional diversity using FID shows our approach achieves comparable or better diversity than SDXL-Base while maintaining SDXL-DMD [38] efficiency. While our focus is sample diversity (Table 2), this confirms our intervention preserves dataset coverage. Skipping first timestep achieves superior FID but lower generative quality (CLIP [12], IS [31]).

Causal Evidence. To causally close the loop, we also provide an experiment in the Appendix E where keeping the first timestep with the distilled model and replacing the later timesteps with the base model does not help with improvement in sample diversity. This demonstrates that the first timestep is indeed the critical bottleneck, validating our mechanistic understanding derived from \hat{x}_0 visualization.

5.2. Hyperparameter Analysis and Efficiency

Figure 7 analyzes the key parameters of our approach. Most importantly, using the base model for just the first timestep ($k=1$) provides substantial diversity gains with minimal computational overhead (Fig 7.b), confirming our hypothesis about first-step criticality. The guidance scale analysis shows optimal performance around zero guidance from the base model, suggesting that natural diversity is preserved

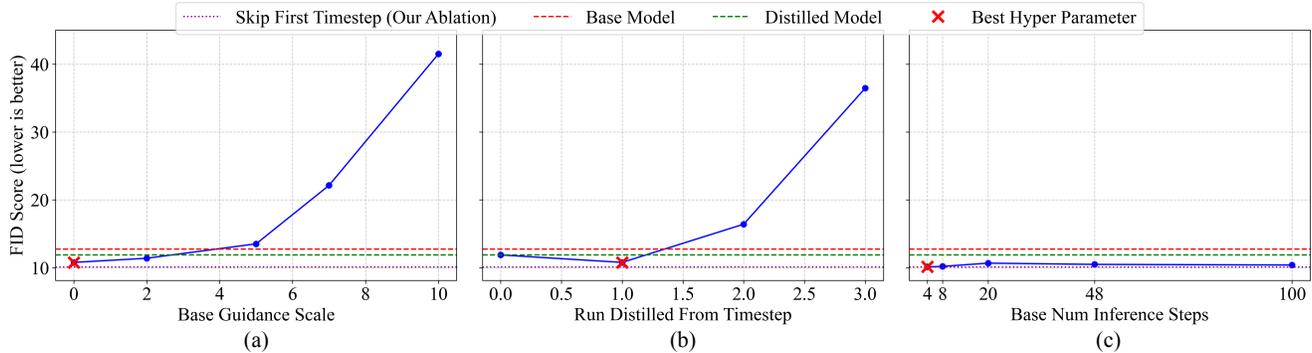


Figure 7. (a) Impact of guidance scale from the base model on diversity shows optimal performance around 0 guidance. (b) Effect of the number of distilled model steps (k) being replaced by base model inference. Running distilled model from first timestep ($k = 1$) provides diversity gains with minimal computational overhead. (c) Comparing the total timesteps of base model when replacing the first timestep of distilled model shows that replacing 1-1 timesteps of distilled with base is most ideal.

without additional steering (Fig 7.a). Finally, setting the base model’s noise schedule similar to distilled ($n=4$) has optimal results (Fig 7.c). We find that using exact inference conditions for distilled models where the model weights are swapped with base for the first timestep is ideal.

For scenarios where loading both models simultaneously is not feasible, we explore skipping the first timestep entirely. Table 3 shows this approach provides significant diversity improvements, though our hybrid method achieves superior quality as measured by CLIP and Inception scores. We provide qualitative examples in the Appendix G.

Our hybrid method demonstrates that the conventional efficiency-diversity tradeoff in distilled models can be resolved through targeted intervention. By identifying and addressing the specific timestep responsible for diversity collapse, we restore sample variety without sacrificing computational advantages. Our theoretical support in Appendix A provides a principled explanation for our empirical observations and suggests that distributing first-timestep decisions across multiple steps during training could be a promising direction for future distillation methods.

6. Limitations

While our approach significantly improves diversity without substantial computational overhead, some limitations remain. First, our method requires loading both base and distilled models in memory, increasing resource requirements compared to traditional inference. Future distillation work could explore our insights to design diversity-preserving mechanisms directly into a distilled model.

Second, our analysis focuses primarily on visual diversity metrics. Further investigation is needed to understand the impact on semantic diversity: the range of concepts and compositions a model can generate. Developing more diversity metrics that capture both visual and semantic varia-

tions can provide deeper insights into distillation.

Finally, our approach treats all prompts uniformly, but different concepts may benefit from different base/distilled step allocations. Adaptive inference strategies that dynamically adjust the transition point based on prompts could further optimize the quality-efficiency trade-off.

7. Conclusion

This work addresses a fundamental limitation of distilled diffusion models: the trade-off between computational efficiency and sample diversity. Our contributions are three-fold: (1) We demonstrate that distilled models retain all the variational directions needed for diversity, contradicting the hypothesis that distillation damages concept representations—the required diversity mechanisms exist but fail to activate during generation; (2) We identify the root cause using \hat{x}_0 trajectory visualization, revealing that distilled models concentrate structural decision-making in the first timestep while base models distribute decisions across many steps, and provide theoretical support showing why the first timestep creates a diversity bottleneck; and (3) We validate this theory by developing diversity distillation, demonstrating that targeted intervention at a single timestep restores full base model sample diversity as measured by DreamSim scores with strong empirical validation.

Our experimental results challenge the conventional diversity-efficiency trade-off. Diversity distillation restores the diversity of the original base model while maintaining the computational efficiency of distilled inference (0.64s vs. 9.22s per image). By providing both mechanistic understanding and theoretical grounding for why diversity collapse occurs, our approach eliminates this traditional trade-off without additional training or model modifications, opening new possibilities for deploying efficient yet diverse generative models in creative applications.

Acknowledgment

RG and DB are supported by Open Philanthropy and NSF grant #2403304.

References

- [1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3
- [2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaoohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2
- [3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2
- [4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 4
- [5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 5, 6
- [6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [7] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 3
- [8] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 2, 3, 4
- [9] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 3
- [10] Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. *arXiv preprint arXiv:2502.01639*, 2025. 2, 3, 4
- [11] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024. 2
- [12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 7
- [13] 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. *Advances in neural information processing systems*, 30, 2017. 2, 6
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4
- [15] Rohit Jena, Ali Taghibakhshi, Sahil Jain, Gerald Shen, Nima Tajbakhsh, and Arash Vahdat. Elucidating optimal reward-diversity tradeoffs in text-to-image diffusion models. *arXiv preprint arXiv:2409.06493*, 2024. 2
- [16] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 3
- [17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2, 3, 4
- [18] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024. 7
- [19] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2
- [20] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2, 4
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2, 5
- [22] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 3
- [23] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2, 4
- [24] ManBearScientist. Shop for the right seed. https://www.reddit.com/r/StableDiffusion/comments/x6kjdh/steps_for_getting_better_images/, 2022. [Accessed 16-09-2025]. 1, 2

- [25] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2
- [26] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024. 3
- [27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2
- [28] 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*, pages 10684–10695, 2022. 2, 4
- [29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3, 4
- [30] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2
- [31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7
- [32] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2, 4
- [33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 2, 4
- [34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2
- [35] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6): 1–13, 2023. 3
- [36] Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023. 2, 5
- [37] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2
- [38] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 2, 4, 7
- [39] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 2
- [40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2