Bringing Adversarial Loss to the Text Stream of Multimodal Diffusion Transformers

Anonymous authors
Paper under double-blind review

Abstract

Adversarial diffusion distillation has conventionally relied on image features for discrimination. We revisit this choice for multimodal diffusion transformers (MMDiTs), whose text stream becomes image-dependent through joint attention. We introduce *Text-Stream Adversarial Distillation* (TSAD), which exploits intermediate text-token states as a new adversarial representation. Because the text stream has already read the image while remaining tied to the text condition, changing that condition changes what information it extracts from the image. Building on this property, we introduce *Aspect-Cued Discrimination* (ACD), which conditions the discriminator on different aspects of the sample, such as typography, objects, or style, to steer which properties are emphasized during adversarial training. Building on the expert-discriminator design of NitroFusion, we integrate TSAD and ACD in **NitroFusion2**.

NitroFusion2 brings few-step distillation from the conventional 4–8-step regime down to just 1–2 steps at comparable quality. At one step, it produces sharp rendered text and rich fine-grained details, making one-step generation practical rather than merely a stress test for distillation. At two steps, it remains competitive across the evaluated benchmarks. These results establish the MMDiT text stream as a powerful and steerable representation for adversarial diffusion distillation.



Figure 1: **Two-step samples from NitroFusion2.** Best viewed with zoom.

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

Adversarial diffusion distillation is an effective approach to few-step generation [75]: discriminator heads distinguish reference samples from student samples using intermediate features of a pretrained backbone, and the student is trained to fool them [6, 32, 58, 59]. The backbone may be an external encoder or the diffusion model itself. Nearly every part of that recipe has been revisited, from the objectives [74, 75] and training strategies [8, 53] to the discriminator design [6, 33, 58], except one: what the discriminator heads read. They read image features. We revisit this choice for multimodal diffusion transformers (MMDiTs).

This choice was natural for earlier diffusion architectures. In a cross-attention UNet [2, 49, 54, 55], the text embeddings used for conditioning remain image-independent throughout the denoising network. The architecture therefore does not maintain text-token hidden states that depend on the sample being processed. MMDiTs [9] change this premise. Through joint attention, intermediate text-token states repeatedly interact with image tokens and become dependent on the particular sample being processed [1, 11, 64]. The MMDiT text stream therefore exposes a representation that conventional adversarial distillation has largely left unused.

We exploit this representation with *Text-Stream Adversarial Distillation* (TSAD), which attaches lightweight discriminator heads to intermediate text-token states of the discriminator backbone. These heads operate on text-token representations that have already interacted with the image. This turns the text stream itself into a discriminator representation.

In joint attention, text tokens query image features, and what they gather is written back into their own states [11, 30, 64], so changing the text condition changes what the text stream extracts from the image. If the reference faithfully presents a queried property, such as clear rendered text, the correct object colour, or a particular visual style, the student must reproduce it well enough to fool the discriminator under the same emphasis. Building on this property, we introduce *Aspect-Cued Discrimination* (ACD). For each reference image, we construct an image-grounded structured caption with six fields: typography, object, style, background, action, and layout. At each training step, the discriminator is conditioned on either one valid aspect field or the full structured caption, while the student keeps its original prompt. The cue therefore steers what the text stream queries from the image and which aspects are emphasized by the adversarial signal.

Building on the expert-discriminator design of NitroFusion [6], we introduce **NitroFusion2**, distilled from Z-Image-Turbo [77], extending adversarial discrimination to the MMDiT text stream through TSAD and ACD.

Controlled ablations test these two design choices. Text-stream discrimination outperforms conventional image-stream discrimination across all evaluated benchmarks, while combining the two streams performs best on most metrics. Aspect cues further turn the discriminator into an expert on the selected aspect, improving performance on that property. Together, these designs enable a substantial leap in one-step generation. At two steps, NitroFusion2 is competitive with alternatives [37] that use separate parameter sets for their two-step trajectories.

In summary, our contributions are:

- We identify the text stream of MMDiTs as a discriminative representation for adversarial diffusion distillation and introduce *Text-Stream Adversarial Distillation*, which applies discriminator heads to intermediate text-token states.
- We introduce *Aspect-Cued Discrimination*, which uses the discriminator's text conditioning to steer which properties are emphasized by the text stream.
- We show that text-stream discrimination outperforms conventional image-stream discrimination across all evaluated benchmarks, and that aspect cues yield aspect-specific gains. Together, these advances bring few-step distillation from the conventional 4–8-step regime down to 1–2 steps at comparable quality.

If this paper has one message, it is this: in an MMDiT, the text stream is no longer only conditioning. Once it reads the image, it becomes a discriminative representation, opening a new direction for adversarial diffusion distillation and allowing that discrimination to be steered toward specific aspects.

2 Related Work

Few-Step Diffusion Models. Reducing diffusion sampling to a few network evaluations has motivated a large body of distillation work. Progressive Distillation [56] pioneered a curriculum that halves the step count iteratively, later generalised to multistep targets [16]. Consistency models [61] and their variants [7, 24, 41, 42, 53, 60, 65] enforce self-consistency along the sampling trajectory, mapping any point on it directly to the endpoint. Distribution Matching Distillation (DMD) [43, 75, 76] minimises a divergence between student and teacher outputs, and has since been extended in several directions [3, 10, 20, 25, 29, 38, 40, 66, 68, 69, 74]. MeanFlow [12, 13] learns the average velocity across an interval. Self-adversarial fine-tuning has the model supply its own adversarial target [8, 36]. Despite this progress, one- and two-step generation remains challenging [8, 37].

Adversarial Diffusion Distillation. Adversarial objectives recover fidelity lost at low step counts. ADD [59] attaches lightweight discriminator heads to a frozen DINOv2 encoder [47] in pixel space. LADD [58] moves discrimination into latent space, applying heads to the token sequence the teacher diffusion model emits after each attention block. NitroFusion [6] organises the critic as an ensemble of specialised expert heads in place of a single scorer. Subsequent methods develop these designs across backbones and head architectures [3, 6, 7, 22, 32, 33, 66, 73, 75], and several train the critic on teacher generations in place of real images [3, 6, 22, 32, 37, 58]. Text-conditioned discrimination was developed earlier in text-to-image GANs [28, 45, 52, 57, 63, 72, 78]. GigaGAN [21] brings it to large-scale models, conditioning its discriminator on a pooled global descriptor while its generator uses word-level cross-attention. Across these designs the text conditions the judgement while the features the discriminator reads come from the image.

Text-Image Representations in Diffusion Models. Diffusion models develop rich text-image correspondences. In cross-attention UNets, word-level attention maps support attribution [35, 62], editing [18, 48], inference-time correction [5, 51], and segmentation [46, 70], while the text representation itself does not vary with the image. Joint attention in multimodal diffusion transformers [9] changes this. ConceptAttention [17] localizes concepts in attention outputs, while recent analyses show that text-stream positions absorb image content [30, 64], including in image-conditioned in-context generation [11], causally steer outputs [27], and support concept prediction from partially denoised states [1]. These representations have been used for analysis and inference-time control, attention-map supervision [23], and preference rewards [39], but not directly as the feature space for adversarial supervision.

3 Approach

3.1 Preliminaries: Adversarial Diffusion Distillation

A few-step student $G_\theta(x_t, t, p)$ maps a latent x_t at noise level t and original prompt p to a clean latent. For one-step generation, $\hat{x} = G_\theta(z, 1, p)$ with $z \sim \mathcal{N}(0, I)$. To compare a reference sample x and a student sample $\hat{x}$ at the same noise level, we use the flow-matching interpolant [34]

$$x_t = (1-t)x + t\epsilon, \qquad \hat{x}_t = (1-t)\hat{x} + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

We denote the resulting reference and student distributions by p_{real} and p_{fake}.

Following adversarial diffusion distillation [58, 59], we reuse a pretrained text-conditioned diffusion model as a frozen discriminator backbone and attach lightweight trainable heads to its intermediate states. Writing D for the composition of the frozen backbone with a trainable head, the adversarial objective separates reference from student samples,

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x_t \sim p_{\text{real}}}\big[\log D(x_t)\big] + \mathbb{E}_{\hat{x}_t \sim p_{\text{fake}}}\big[\log\big(1 - D(\hat{x}_t)\big)\big], \tag{2}$$

and the student is trained to raise the value D assigns to its own samples.

3.2 Text-Stream Adversarial Distillation

We refer to the evolving image-token and text-token hidden states in an MMDiT as the image and text streams. Existing adversarial distillation reads representations from the image stream [3, 6, 33, 58, 75]. *Text-Stream Adversarial Distillation* (TSAD) instead extends adversarial discrimination to the text-stream attention outputs.

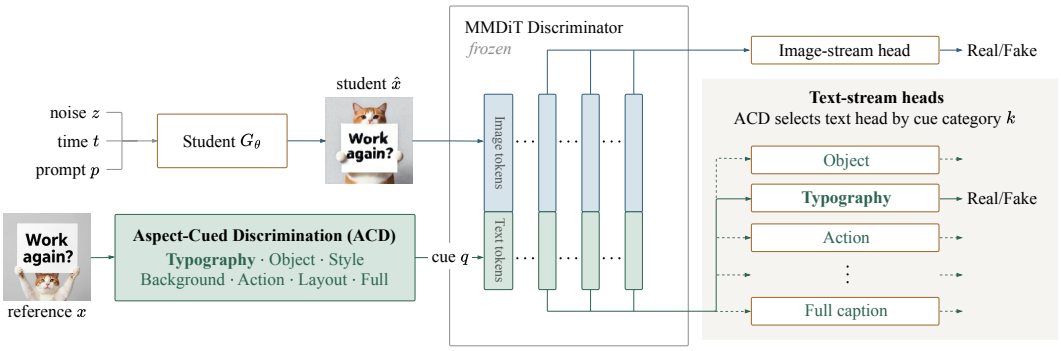

Figure 2: **Architecture of Text-Stream Adversarial Distillation (TSAD) and Aspect-Cued Discrimination (ACD).** TSAD adds discriminator heads to intermediate text-token states alongside conventional image-stream heads. ACD builds image-grounded aspect fields, then samples one cue per training step. The cue conditions the MMDiT discriminator backbone and selects the corresponding text-stream head, steering discrimination toward that aspect.

In a cross-attention UNet [54], the text representation $E(p)$ is computed independently of the image and supplied to the image network as keys and values. Although the resulting cross-attention weights depend on image features, the text representation itself remains fixed, with $\partial E(p)/\partial \hat{x}_t = 0$. A discriminator attached directly to these text states therefore cannot provide an adversarial gradient to the student.

In an MMDiT [9], text and image tokens instead interact through joint attention, whether the two streams carry separate weights or share them as in Z-Image [77]. Consider one layer with incoming text states $H_T \in \mathbb{R}^{m\times d}$. Let $Q_T = H_T W_Q$ denote the text queries, and let $K = [K_I; K_T]$ and $V = [V_I; V_T]$ collect the image and text keys and values. The text-query attention weights are

$$A = \text{softmax}\left(Q_T K^\top/\sqrt{d}\right), \qquad A = [A_I\ A_T]. \tag{3}$$

The corresponding attention output is $U_T = A_I V_I + A_T V_T$, and U_I denotes the image-stream attention output of the same block, followed by $H'_T = \psi(H_T + U_T W_O)$, where ψ denotes the remaining token-wise operations of the block.

The image contribution $A_I V_I$ makes the text attention output directly dependent on image features [30, 64]. Intuitively, each text token acts as a query into the image: it selects image information through A_I, which contributes directly to its attention output. This gives sample-dependent text representations whose positions remain associated with the conditioning text [1]. Where image tokens describe local appearance at fixed spatial positions, the text stream summarizes the sample along the axes the prompt names, so the adversarial signal is organized by those axes rather than by local texture alone. TSAD attaches discriminator heads directly to these attention outputs; image-to-text attention maps A_I serve for visualization and for the layer selection in Appendix A.2.

Equation (3) establishes the property required for discrimination, namely that the text attention outputs depend on the image. Section 4 tests whether they suffice as an adversarial signal.

As illustrated in Fig. 2, we attach lightweight discriminator heads [33] to text-stream attention outputs at selected layers of the frozen MMDiT backbone, with image-stream heads attached to the corresponding image attention outputs. We denote their logits by $\ell_q^{\text{text}}(x_t)$ and $\ell_q^{\text{img}}(x_t)$, where q is the text conditioning supplied to the discriminator backbone. Both sets of heads read attention outputs from the same backbone pass and provide adversarial signals to the student.

3.3 Aspect-Cued Discrimination

Because the discriminator reuses a text-conditioned diffusion backbone, its conditioning prompt can be chosen independently of the original prompt p used by the student. We exploit this flexibility with *Aspect-Cued Discrimination* (ACD), which conditions the discriminator on one structured aspect of the sample at a time. Because the backbone is frozen and the cue enters only as its text input, each



Figure 3: **Qualitative cue-dependent image-to-text attention maps.**

cue turns the same network into a different critic at no additional cost: rather than using a separate discriminator per property, we re-query one shared backbone for the property we want to emphasize.

For each reference image, we construct six image-grounded aspect fields indexed by

$$\mathcal{A} = \{\text{typography}, \text{object}, \text{style}, \text{background}, \text{action}, \text{layout}\}.$$

We denote the text associated with aspect k by a_k. Their combination forms the full structured caption c_{full}. Fields whose content is absent from the image are left empty. The valid cue categories for a sample are

$$\mathcal{K}(a) = \{\text{full}\} \cup \{k \in \mathcal{A} : a_k \neq \emptyset\}.$$

At each training step, we sample $k \sim \text{Uniform}(\mathcal{K}(a))$. The discriminator cue q is c_{full} when $k = \text{full}$ and a_k otherwise. The student remains conditioned on the original prompt p, while the frozen discriminator backbone is conditioned on q. The same cue is used for the corresponding reference and student samples. Following the expert-head design of NitroFusion [6], the text-stream discriminator uses heads specific to cue categories. A typography cue is scored by typography heads, an object cue by object heads, and likewise for the remaining aspects and the full caption. The image-stream heads remain shared across cues, while their input states still depend on q through the conditioned MMDiT backbone.

The sampled cue therefore affects text-stream discrimination in two coupled ways. It changes the text queries and hence the image information written into the text-token states, and it selects the corresponding aspect-specific text-stream heads. A typography cue, for example, conditions the backbone on the rendered-text description and routes the resulting states to the typography heads, while an object cue uses the object description and object heads.

Figure 3 qualitatively illustrates this cue-dependent attention. With the same image fixed, different aspect cues produce different image-to-text attention patterns; for example, the typography and object cues emphasize different regions associated with rendered text and foreground objects, the background cue shifts attention away from the foreground toward the surrounding scene, and the full caption spreads attention more broadly across the image. These patterns are consistent with ACD steering image-to-text attention through the discriminator cue. ACD requires no additional backbone evaluation and only selects the corresponding lightweight text-stream heads.

3.4 Dataset Construction

We construct a synthetic dataset of 100k reference images [3, 6, 22, 32, 37, 58] using Z-Image-Turbo [77] with twelve sampling steps, split evenly between English and Chinese prompts. Images are generated at a fixed budget of roughly 1024×1024 pixels across a range of aspect ratios. We denote the original prompt by p. For each reference image, the multimodal model Qwen3.5-35B-A3B [50] receives only the image and generates six image-grounded fields describing typography, object, style, background, action, and layout. These fields are combined to form the full structured caption c_{full}.

A field is left empty when the corresponding content is absent from the image. For example, the typography field is populated only when visible text is present. We additionally verify that the resulting structured caption is consistent with the original prompt p. During training, the student remains conditioned on p, while the discriminator cue is sampled uniformly from c_{full} and the available non-empty aspect fields. The cue categories are listed in Appendix A.1.

3.5 Training Objective

We alternate between updating the discriminator heads and the student while keeping the discriminator backbone frozen. For each stream $s \in \{\text{text}, \text{img}\}$ we use a hinge loss [31] for adversarial

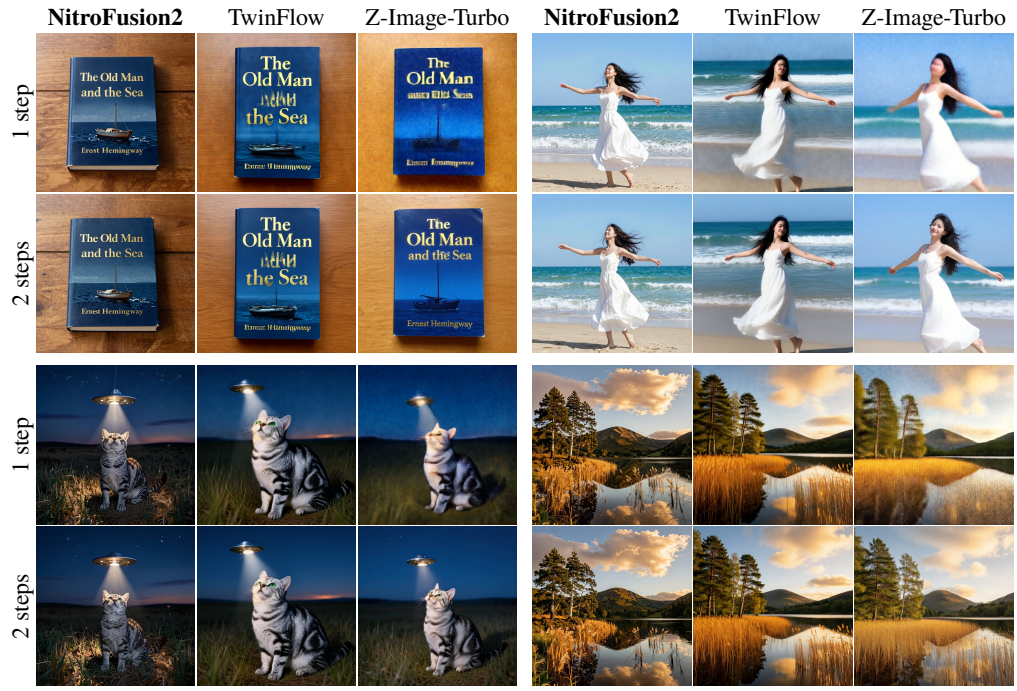

Figure 4: **Qualitative comparison.** NitroFusion2 preserves rendered text and fine image detail markedly better than Z-Image-Turbo and TwinFlow.



Figure 5: **Qualitative results across multiple steps.**

training,

$$\mathcal{L}_{\mathrm{D}}^{s} = \frac{1}{2}\mathbb{E}_{q}\Big[\mathbb{E}_{x_t\sim p_{\mathrm{real}}}\max\big(0, 1-\ell_q^s(x_t)\big) + \mathbb{E}_{\hat{x}_t\sim p_{\mathrm{fake}}}\max\big(0, 1+\ell_q^s(\hat{x}_t)\big)\Big], \tag{4}$$

where q is sampled according to the ACD procedure in Section 3.3.

The student is trained to increase the discriminator logits assigned to its own samples. The text-stream and image-stream adversarial losses are

$$\begin{aligned}\mathcal{L}_{\mathrm{text-adv}} &= -\mathbb{E}_q\mathbb{E}_{\hat{x}_t\sim p_{\mathrm{fake}}}\left[\ell_q^{\mathrm{text}}(\hat{x}_t)\right],\\ \mathcal{L}_{\mathrm{img-adv}} &= -\mathbb{E}_q\mathbb{E}_{\hat{x}_t\sim p_{\mathrm{fake}}}\left[\ell_q^{\mathrm{img}}(\hat{x}_t)\right].\end{aligned} \tag{5}$$

At low noise levels the student is additionally regressed toward the reference sample,

$$\mathcal{L}_{\mathrm{fm}} = \mathbb{E}_{t\le t_{\mathrm{fm}}}\left[w(t)\,\|G_\theta(x_t,t,p)-x\|_2^2\right], \tag{6}$$

Table 1: **Few-step students of Z-Image.**

Method	Params	NFE	Text rendering		Overall		
			LongText-EN	LongText-CN	GenEval	DPG-Bench	OneIG
NitroFusion2 (ours)	6B	1	**43.34**	**47.78**	**75.86**	**86.48**	**47.63**
Z-Image-Turbo [77]	6B	1	11.72	12.70	57.08	77.23	32.43
TwinFlow [8]	6B	1	29.18	36.85	66.34	85.17	42.20
NitroFusion2 (ours)	6B	2	76.03	81.95	74.46	**86.21**	51.45
Z-Image-Turbo [77]	6B	2	72.89	78.48	**76.53**	85.78	50.94
TwinFlow [8]	6B	2	71.99	78.65	72.41	85.98	51.38
Z-Image-Turbo++ [37]	6B+6B	2	**89.88**	**91.62**	75.70	85.86	**52.50**
Z-Image-Turbo [77]	6B	8	91.74	92.56	75.01	84.86	52.84

Table 2: **Per-task breakdown on GenEval.**

Method	Params	NFE	Overall	Single	Two	Counting	Colors	Position	Color attr.
NitroFusion2 (ours)	6B	1	**75.86**	**100.00**	**96.97**	**62.19**	**91.22**	32.25	**72.50**
Z-Image-Turbo [77]	6B	1	57.08	97.81	51.26	55.31	76.86	31.75	29.50
TwinFlow [8]	6B	1	66.34	97.50	77.02	55.31	82.45	**36.75**	49.00
NitroFusion2 (ours)	6B	2	74.46	**100.00**	**96.46**	56.88	**90.16**	31.25	**72.00**
Z-Image-Turbo [77]	6B	2	**76.53**	99.38	89.39	**74.38**	87.77	**50.50**	57.75
TwinFlow [8]	6B	2	72.41	99.69	87.12	67.81	85.11	40.00	54.75
Z-Image-Turbo++ [37]	6B+6B	2	75.70	99.69	87.63	69.06	84.57	48.75	64.50
Z-Image-Turbo [77]	6B	8	75.01	99.69	88.89	71.88	85.11	45.75	58.75

where t_{fm} specifies the regression range and $w(t)$ weights the corresponding noise levels. The student objective combines the three terms,

$$\mathcal{L}_{\text{G}} = (1 - \lambda_{\text{img}})\mathcal{L}_{\text{text-adv}} + \lambda_{\text{img}}\mathcal{L}_{\text{img-adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}, \tag{7}$$

where λ_{img} controls the balance between text-stream and image-stream adversarial discrimination. We use $\lambda_{\text{img}} = 0.1$, $\lambda_{\text{fm}} = 0.1$ and $t_{\text{fm}} = 0.75$.

4 Experiments

Both our student and our discriminator backbone are initialized from Z-Image-Turbo [77], and the discriminator backbone is kept frozen throughout training. We evaluate NitroFusion2 on LongText-Bench [14] in English and Chinese, OneIG-Bench [4], GenEval [15], and DPG-Bench [19]. We compare against few-step students of Z-Image: Z-Image-Turbo [77], TwinFlow [8], and Z-Image-Turbo++ [37]. Z-Image-Turbo++ uses step-decoupled weights, with a separate model for each of its two steps. Its weights are not released, so we report its published two-step metrics but cannot evaluate it at one step or include it in our qualitative comparisons. Training details are given in Appendix A and per-dimension benchmark breakdowns in Appendix B. All qualitative samples in this section are generated at 1024×1024.

4.1 Qualitative Results

Figure 4 compares NitroFusion2 with Z-Image-Turbo and TwinFlow at one and two sampling steps. The clearest difference is in rendered text. At one step, NitroFusion2 already produces a sharp, legible title, "The Old Man and the Sea," while both baselines blur or distort the lettering. At two steps, NitroFusion2 also renders the much smaller author name, "Ernest Hemingway," clearly. Because the discriminator is conditioned on the described text, the student must reproduce the glyphs it names, not merely produce sharp lettering.

Beyond typography, NitroFusion2 already produces realistic, fine-grained details at one step, including sharp structure in hair, fur, foliage, water, and reflections. Figure 5 further shows that high-quality generation is available from the first step and remains stable as the sampling budget increases. One-step generation can still struggle with complex scenes, but these errors are progressively alleviated with more steps; in these examples the visual differences beyond eight steps are small.

Table 3: **Which stream the heads read.** Trained for 1,000 steps and evaluated at 1 NFE.

Setting	LongText-EN	LongText-CN	GenEval	DPG-Bench	OneIG
Text	22.31	24.37	**66.71**	84.39	38.45
Image	17.86	19.26	66.00	82.87	35.79
Text + image	**24.47**	**26.63**	64.39	**84.86**	**40.02**

Table 4: **Which cue conditions the discriminator?** Trained for 1,000 steps and evaluated at 1 NFE.

	Text rendering		Style	Object		Alignment	
Cue	LT-EN	LT-CN	OneIG-Style	GenEval-Single	DPG-Entity	OneIG-Align.	DPG-Global
Typography	**45.06**	**53.01**	32.66	99.38	92.52	78.48	79.33
Style	16.09	19.28	33.75	98.44	92.36	76.62	79.60
Object	20.82	25.19	32.67	**100.00**	92.94	79.14	78.72
Full caption	17.29	19.55	30.88	99.06	92.64	76.96	79.33
All cues	24.47	26.63	**35.79**	98.44	**93.21**	**79.69**	**80.85**

4.2 Quantitative Results

Table 1 compares few-step students of Z-Image at one and two sampling steps. At one step, NitroFusion2 leads all evaluated baselines across all five benchmarks. Compared with TwinFlow, it substantially improves LongText-EN and LongText-CN by 14.2 and 10.9 points, respectively, while also improving GenEval, DPG-Bench, and OneIG. The GenEval breakdown in Table 2 further shows that this advantage extends across compositional skills. NitroFusion2 leads five of six categories, with particularly strong results on single- and two-object generation, colors, and color attribution.

At two steps, NitroFusion2 remains strong overall. It leads four of six GenEval categories and outperforms TwinFlow on both LongText splits. Z-Image-Turbo++ achieves higher scores on some metrics and uses two step-specific 6B parameter sets, 12B parameters in total, whereas a single 6B NitroFusion2 model is used at any sampling step.

Most notably, NitroFusion2 leads all evaluated benchmarks at one step. At that single step it also surpasses the eight-step Z-Image-Turbo reference on DPG-Bench (86.48 against 84.86) and is on par with it on GenEval (75.86 against 75.01), at one eighth the sampling cost, while a substantial gap to that reference remains on text rendering. One step is also where the adversarial signal matters most, since the student cannot revise its output over later steps and must place all structure in a single pass.

4.3 Ablation Study

We ablate the two components: first, which stream provides the adversarial representation; second, which cue conditions the shared discriminator backbone. All ablation runs are trained for 1,000 steps.

Which stream should the discriminator read? Table 3 compares image-stream, text-stream, and joint discrimination. Text-stream discrimination alone outperforms conventional image-stream discrimination on every benchmark. Combining both streams further improves four of the five metrics, including both LongText splits, DPG-Bench, and OneIG. GenEval is the exception, where the text-only model performs best.

These results isolate the contribution of extending adversarial discrimination to the text stream. Text-stream supervision is stronger than image-stream supervision across all five metrics. Our approach retains both because the image-stream signal provides an additional gain on most metrics.

Which cue should condition the discriminator? We next fix the joint text- and image-stream discriminator and vary only the cue supplied to its shared backbone. Table 4 compares models trained with a single typography, style, or object cue, the full structured caption, and ACD with all cues sampled uniformly. All settings use the same training data and discriminator architecture. Where Fig. 3 shows that the cue changes where the text stream reads, Table 4 asks whether that shows up in what the student learns.



Figure 6: **Two-step ablative samples.** Left: image-stream, text-stream, and joint discrimination. Right: models trained with individual typography, style, or object cues compared with the full seven-cue ACD mixture.

Among the single-aspect models, each cue performs best on the evaluation category associated with the aspect it names. The typography cue gives by far the strongest LongText scores, the style cue gives the strongest OneIG-Style score, and the object cue gives the strongest single-aspect results on both object-focused metrics. The full structured caption does not lead any targeted category. This targeted pattern indicates that changing the cue changes which properties receive the strongest adversarial supervision. Typography alone, for instance, exceeds the final model on LongText, because all of its adversarial supervision is spent on rendered text.

Sampling across all cues produces a broader effect. Rather than specializing in a single property, the mixed cue policy balances supervision across different aspects. It does not maximize every targeted category, but gives the strongest OneIG-Style, DPG-Entity, and both alignment scores in the table. More importantly, the full ACD mixture outperforms the full structured prompt on six of seven metrics, supporting the benefit of decomposing the prompt into specialized aspect cues rather than conditioning on the full caption alone. In particular, the two alignment metrics are highest under the mixed cue policy rather than any single-aspect specialization.

The qualitative ablations in Fig. 6 illustrate the same trends. On the left, image-stream discrimination struggles with rendered text and complex structure: it duplicates "KYOTO" and introduces artifacts around the cat, while text-stream and joint discrimination produce cleaner results. On the right, the effect of different ACD cues is visible across aspects. The typography cue produces the clearest menu text, while the object cue best preserves the requested sequence of colored balloons. Sampling across all cues performs well in both cases, illustrating the broader balance of the full ACD mixture.

5 Conclusion and Limitations

Our experiments show that MMDiT text-token states can serve as an effective adversarial representation. **Text-Stream Adversarial Distillation** (TSAD) exploits these states for supervision, with particularly strong gains in text rendering, while **Aspect-Cued Discrimination** (ACD) provides targeted control and a better overall balance than full-prompt conditioning. Together, they enable NitroFusion2 to substantially narrow the gap between one-step and multi-step generation. The text stream is useful here not only because it carries image information, but because what it carries can be redirected by the conditioning, making it a steerable representation rather than a fixed one.

Several limitations remain. Complex scenes are still challenging, particularly at one step, and sample diversity is reduced. ACD also exhibits a seesaw effect across cues, where emphasizing one aspect can improve that property while weakening others. Our layer selection is supported empirically rather than established as optimal.

These observations suggest several directions for future work. Better balancing the seesaw effect across cues may further improve ACD, while studying where and how text-stream information evolves across depth could lead to more effective discriminator designs. Which aspects make effective cues also remains an open question; the six fields used here are a design choice. More broadly, text-stream states may provide useful supervision beyond adversarial distillation, opening a wider direction for training and controlling multimodal diffusion models.

AI USE STATEMENT

In this work, we used generative AI tools and LLMs to generate synthetic data sets, implement methods, provide feedback on research methodology and experiments, and interpret results. We also used them to draft and edit parts of this paper and to create and modify its figures. We did not use them to propose or refine the hypotheses of this work or to develop its conceptual framework. We have reviewed all AI-assisted work: every reported number comes from a run the authors launched and verified. The authors are responsible for all content in this paper, including anything produced with AI assistance.

REPRODUCIBILITY STATEMENT

The method is defined in Sections 3.2 and 3.3 and the training objective in Section 3.5. Appendix A gives the training setup. Appendix A.1 lists the seven cue categories and Appendix A.2 the layers each set of heads reads. Appendix C sets out one full training step as pseudo-code. Dataset construction, including the annotation model and the six structured-caption fields, is described in Section 3.4. NitroFusion2 will be released publicly upon acceptance.

REFERENCES

[1] Kanghyun Baek, Jaihyun Lew, Chaehun Shin, Jungbeom Lee, and Sungroh Yoon. Diagnosing and correcting concept omission in multimodal diffusion transformers. In *ICML*, 2026. 2, 3, 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Hmrishav Bandyopadhyay, Rahim Entezari, Jim Scott, Reshinth Adithyan, Yi-Zhe Song, and Varun Jampani. SD3.5-Flash: Distribution-guided distillation of generative flows. *arXiv preprint arXiv:2509.21318*, 2025. 3, 5

[4] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. In *NeurIPS*, 2025. 7

[5] 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. In *SIGGRAPH*, 2023. 3

[6] Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. NitroFusion: High-fidelity single-step diffusion through dynamic adversarial training. In *CVPR*, 2025. 2, 3, 5, 15

[7] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. SANA-Sprint: One-step diffusion with continuous-time consistency distillation. In *ICCV*, 2025. 3

[8] Zhenglin Cheng, Peng Sun, Jianguo Li, and Tao Lin. TwinFlow: Realizing one-step generation on large models with self-adversarial flows. In *ICLR*, 2026. 2, 3, 7

[9] 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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 3, 4

[10] Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased DMD: Few-step distribution matching distillation via score matching within subintervals. In *CVPR*, 2026. 3

[11] Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. In *ICML Workshops*, 2026. 2, 3

[12] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, 2025. 3

[13] Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *CVPR*, 2026. 3

[14] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Di Wang, and Jie Jiang. X-Omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. 7

[15] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 7

[16] Jonathan Heek, Emiel Hoogeboom, and Tim Salimans. Multistep consistency models. *arXiv preprint arXiv:2403.06807*, 2024. 3

[17] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *ICML*, 2025. 3

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 3

[19] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 7

[20] Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, Bo Zhang, Mengmeng Wang, Steven Hoi, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning. In *ECCV*, 2026. 3

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, 2023. 3

[22] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional GANs. In *ECCV*, 2024. 3, 5

[23] Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4Diff: Unveiling open-vocabulary semantic segmentation in text-to-image diffusion transformers. In *NeurIPS*, 2025. 3

[24] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *ICLR*, 2024. 3

[25] Youngjoong Kim, Deokyeong Lee, and Jaesik Park. Distribution matching distillation without fake score network. *arXiv preprint arXiv:2605.19256*, 2026. 3

[26] Krea AI. Krea 2 Turbo. `https://huggingface.co/krea/Krea-2-Turbo`, 2026. 16

[27] Binglei Li, Mengping Yang, Zhiyu Tan, Junping Zhang, and Hao Li. TexTailor: Inference-time textual guidance tailoring for multimodal diffusion transformers. In *ECCV*, 2026. 3

[28] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In *NeurIPS*, 2019. 3

[29] Haoyu Li, Tingyan Wen, Lin Qi, Zhe Wu, Yihuang Chen, Xing Zhou, Lifei Zhu, Xueqian Wang, and Kai Zhang. 1.x-distill: Breaking the diversity, quality, and efficiency barrier in distribution matching distillation. *arXiv preprint arXiv:2604.04018*, 2026. 3

[30] Maohua Li, Qirui Li, Yanke Zhou, Yiduo Li, Zhaosheng Chi, Chao Xu, Cuifeng Shen, Yixuan Xu, Hanlin Tang, Kan Liu, Tao Lan, Lin Qu, and Shao-Qun Zhang. Text template tokens are implicit semantic registers in diffusion transformers. *arXiv preprint arXiv:2607.19139*, 2026. 2, 3, 4

[31] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017. 5

[32] Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2, 3, 5, 16

[33] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In *ICML*, 2025. 2, 3, 4, 15

[34] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[35] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *CVPR*, 2024. 3

[36] Deyuan Liu, Peng Sun, Yansen Han, Zhenglin Cheng, Chuyan Chen, and Tao Lin. Self-adversarial one step generation via condition shifting. *arXiv preprint arXiv:2604.12322*, 2026. 3

[37] Dongyang Liu, Ruoyi Du, David Liu, Dengyang Jiang, Liangchen Li, Qilong Wu, Zhen Li, Steven C. H. Hoi, Hongsheng Li, and Peng Gao. High-fidelity two-step image generation via teacher-aligned end-to-end distillation. *arXiv preprint arXiv:2606.12575*, 2026. 2, 3, 5, 7

[38] Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. In *ICLR*, 2026. 3

[39] Gongye Liu, Bo Yang, Yida Zhi, Zhizhou Zhong, Lei Ke, Didan Deng, Han Gao, Yongxiang Huang, Kaihao Zhang, Hongbo Fu, and Wenhan Luo. Beyond VLM-Based rewards: Diffusion-native latent reward modeling. In *ICML*, 2026. 3

[40] Tao Liu, Hao Yan, Mengting Chen, Taihang Hu, Zhengrong Yue, Zihao Pan, Jinsong Lan, Xiaoyong Zhu, Ming-Ming Cheng, Bo Zheng, and Yaxing Wang. Continuous-time distribution matching for few-step diffusion distillation. *arXiv preprint arXiv:2605.06376*, 2026. 3

[41] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *ICLR*, 2025. 3

[42] 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. 3

[43] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *NeurIPS*, 2023. 3

[44] lvladikov. Krea 2 Turbo distill 2-step LoRA. `https://huggingface.co/lvladikov/Krea2-Turbo-Distill-2step-LoRA`, 2026. 16

[45] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018. 3

[46] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation. In *NeurIPS*, 2023. 3

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 3

[48] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 3

[49] 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. In *ICLR*, 2024. 2

[50] Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. 5, 15

[51] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *NeurIPS*, 2023. 3

[52] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 3

[53] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory segmented consistency model for efficient image synthesis. In *NeurIPS*, 2024. 2, 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[56] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 3

[57] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In *ICML*, 2023. 3

[58] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, 2024. 2, 3, 5

[59] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024. 2, 3

[60] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *ICLR*, 2024. 3

[61] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 3

[62] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *ACL*, 2023. 3

[63] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *CVPR*, 2022. 3

[64] Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in T2I models. In *NAACL*, 2025. 2, 3, 4

[65] Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *NeurIPS*, 2024. 3

[66] Xu Wang, Zexian Li, Litong Gong, Tiezheng Ge, and Zhijie Deng. AdvDMD: Adversarial reward meets DMD for high-quality few-step generation. *arXiv preprint arXiv:2604.28126*, 2026. 3

[67] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, and Kun Yan. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 16

[68] Junyi Wu, Weijian Luo, Haoyang Zheng, Ruizhe Zhang, and Guang Lin. Diff-instruct with diffused reward: Towards principled one-step generator RL. *arXiv preprint arXiv:2605.24001*, 2026. 3

[69] Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. In *ICML*, 2026. 3

[70] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 3

[71] Wuli-art. Qwen-Image-2512 turbo LoRA, 2 steps. https://huggingface.co/Wuli-art/Qwen-Image-2512-Turbo-LoRA-2-Steps, 2026. 16

[72] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 3

[73] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. UFOGen: You forward once large scale text-to-image generation via diffusion GANs. In *CVPR*, 2024. 3

[74] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025. 2, 3

[75] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024. 2, 3

[76] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 3

[77] Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, and Aiming Hao. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. 2, 4, 5, 7

[78] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021. 3

A Implementation Details

We use Adam with $\beta_1 = 0.0$, $\beta_2 = 0.9$, no weight decay, a learning rate of 5×10^{-5} and a batch size of 256, on eight H100 GPUs. The final model is trained for 20,000 steps in two stages. The first 5,000 steps are off-policy. The remaining 15,000 are on-policy, where the student's input is a state rolled out along its own trajectory, and we set $\lambda_{\text{fm}} = 0$, dropping the flow-matching term.

The discriminator heads follow the design of APT [33]. Following NitroFusion [6], the image stream and each cue carry a pool of three heads rather than a single head, and one head is drawn at random per step. We also refresh the heads periodically by re-initializing their weights, which reduces overfitting.

A.1 The Cue List

Table 5 lists the seven cue categories and the text each one supplies to the discriminator. The structured caption they are drawn from is written by Qwen3.5-35B-A3B [50] from the reference image itself, so each field is detailed and closely aligned with that image.

Table 5: **The seven cue categories.**

Cue	Text the discriminator is conditioned on
Full caption	Every field concatenated: the complete specification of the image.
Typography	The rendered text, transcribed line by line with each line's position and lettering. Populated only when the image contains visible text.
Object	The subject: what is depicted, with the counts and identities of the entities preserved.
Style	The palette, artistic medium and rendering treatment, describing how the image is rendered rather than what it shows.
Background	The setting the subject sits in: surroundings, surfaces, weather and ambient detail, excluding the subject itself.
Action	What the subject is doing, or an explicit statement that the scene is static.
Layout	Where elements sit in the frame and how they are placed relative to one another.

A.2 Choosing Where the Heads Read

We score each layer of the frozen backbone by how far a cue's attention map departs from the full caption's. The full caption itself is scored by its mean departure from the other six cues. Cues separate more after layer 20, so the four best-scoring layers all fall in one band. We keep the two strongest and replace the others with the best-separating layer of the mid and of the early range, so a head also reads coarser shared structure (Table 6). The image stream gets no such selection: its states are not indexed by the cue, so there is no per-cue attention map to rank. We spread its taps evenly across depth instead, at layers 5, 13, 21 and 29.

Table 6: **Read layers per cue.** Indices into the backbone's 30 joint-attention layers.

Cue	Top four, best first	Selected
Full caption	[22, 21, 23, 20]	[22, 21, 17, 10]
Typography	[22, 21, 20, 24]	[22, 21, 17, 10]
Object	[21, 20, 22, 24]	[21, 20, 17, 10]
Style	[21, 22, 20, 25]	[21, 22, 17, 10]
Background	[21, 22, 20, 23]	[21, 22, 17, 10]
Action	[22, 21, 20, 24]	[22, 21, 17, 10]
Layout	[22, 21, 20, 25]	[22, 21, 17, 10]

Table 7 compares the selected taps against the four best-scoring band layers and against the even spacing the image stream uses. The selected taps lead on four of the five metrics, most clearly on

text rendering; GenEval is the exception. We did not sweep it further, so the choice is supported empirically rather than established as optimal.

Table 7: **Where the text-stream heads read.** Trained for 1,000 steps and evaluated at 1 NFE.

Text-stream taps	LongText-EN	LongText-CN	GenEval	DPG-Bench	OneIG
Selected (Table 6)	**24.47**	**26.63**	64.39	**84.86**	**40.02**
Top four	21.26	25.17	64.91	83.40	36.53
Evenly spaced, $[5, 13, 21, 29]$	22.56	24.67	**67.62**	84.74	38.71

B Additional Results

Table 8 breaks OneIG-Bench down by dimension and Table 9 breaks DPG-Bench down by category. Figs. 8 and 9 show further two-step and single-step samples.

On the OneIG diversity dimension (Table 8), NitroFusion2 scores below both baselines at one and two steps. This suggests a trade-off between diversity and fidelity [32].

Table 8: **Per-dimension breakdown on OneIG-Bench.** Results on the five dimensions that form the overall score.

Method	Params	NFE	Overall	Alignment	Text	Diversity	Style	Reasoning
NitroFusion2	6B	1	**47.63**	**85.13**	**76.28**	8.50	**43.00**	**25.24**
Z-Image-Turbo	6B	1	32.43	65.57	29.79	12.68	31.37	22.75
TwinFlow	6B	1	42.20	81.68	54.13	**14.97**	36.10	24.12
NitroFusion2	6B	2	51.45	85.41	91.34	9.63	**44.20**	26.68
Z-Image-Turbo	6B	2	50.94	83.92	92.85	13.85	36.25	27.86
TwinFlow	6B	2	51.38	85.26	90.61	**15.49**	38.03	27.52
Z-Image-Turbo++	6B+6B	2	**52.50**	**85.65**	**97.09**	11.75	37.87	**30.12**
Z-Image-Turbo	6B	8	52.84	84.03	99.32	13.85	36.81	30.19

Table 9: **Per-category breakdown on DPG-Bench.** Results on the five categories that form the overall score.

Method	Params	NFE	Overall	Entity	Attribute	Relation	Global	Other
NitroFusion2	6B	1	**86.48**	**92.36**	**89.62**	**94.74**	**84.19**	**88.00**
Z-Image-Turbo	6B	1	77.23	84.85	83.08	91.03	76.90	82.80
TwinFlow	6B	1	85.17	90.99	88.56	94.43	83.89	87.20
NitroFusion2	6B	2	**86.21**	**92.47**	88.94	93.81	83.28	86.00
Z-Image-Turbo	6B	2	85.78	91.73	88.78	94.16	**84.19**	**89.60**
TwinFlow	6B	2	85.98	91.83	89.02	94.16	81.46	87.60
Z-Image-Turbo++	6B+6B	2	85.86	91.84	**89.30**	**94.58**	83.89	**89.60**
Z-Image-Turbo	6B	8	84.86	89.59	90.14	92.16	91.29	88.68

Beyond the few-step students of Z-Image, we compare with two recent open-weight models using their published two-step Turbo LoRAs: Krea 2 Turbo [26, 44] at 12B parameters and Qwen-Image-2512 Turbo [67, 71] at 20B.

Figure 7 shows the qualitative comparison. At one step, only NitroFusion2 produces consistently sharp outputs across all three examples, preserving fine detail in the cherry-blossom branches, facial hair, and ornate headdresses. At two steps, the differences become smaller, while NitroFusion2 continues to produce fine-grained detail with a 6B model, half the size of Krea 2 Turbo and less than a third of Qwen-Image-2512 Turbo. Table 10 reports the quantitative results. At one step, NitroFusion2 leads four of the five benchmarks, trailing only on LongText-CN. At two steps, Qwen-Image-2512 Turbo leads on text rendering and Krea 2 Turbo on GenEval, while the three models are close on DPG-Bench and OneIG, leaving NitroFusion2 competitive with both larger models.

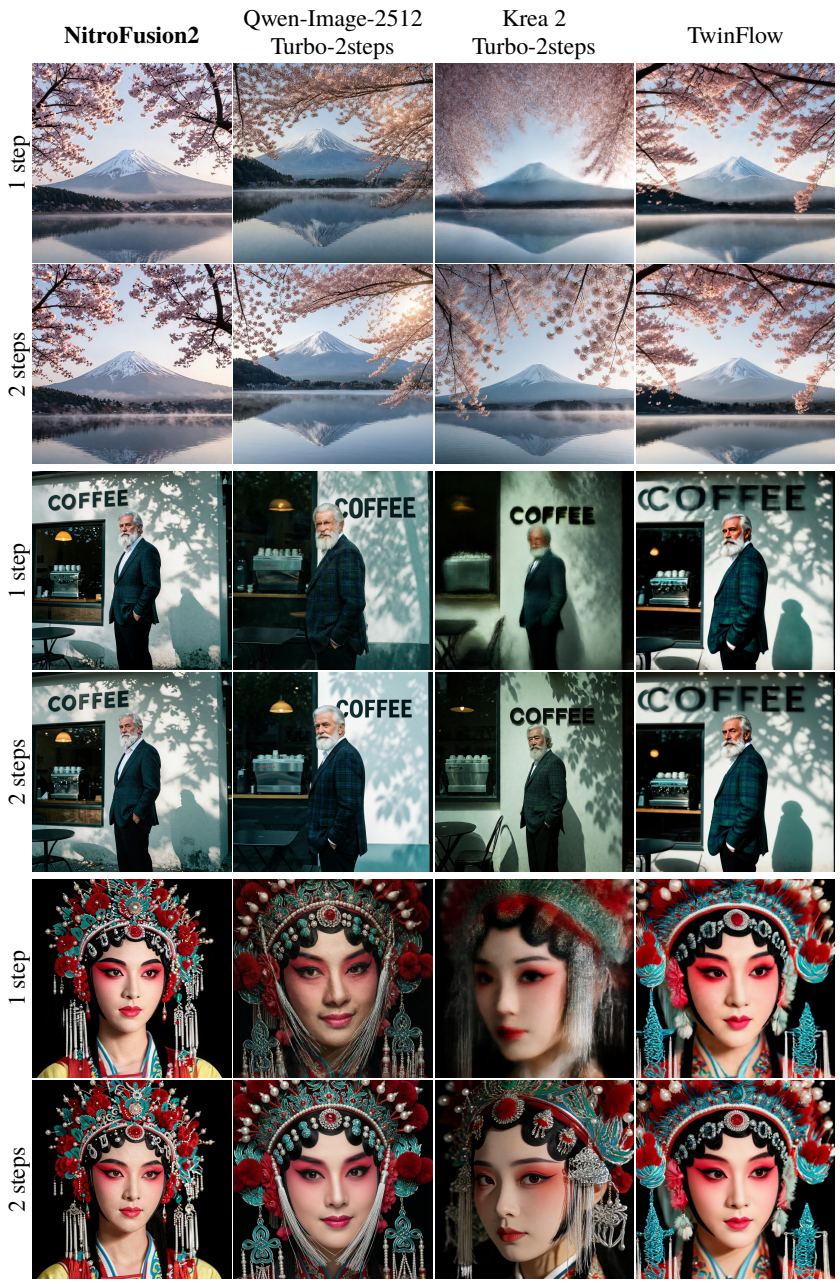

Figure 7: **Additional qualitative comparison at one and two steps.**

Table 10: **Comparison against recent open-weight models at one and two steps.** Every row was sampled by us under identical settings. Parameter counts are denoiser weights.

			Text rendering		Overall		
Method	Params	NFE	LongText-EN	LongText-CN	GenEval	DPG-Bench	OneIG
NitroFusion2 (ours)	6B	1	**43.34**	47.78	**75.86**	**86.48**	**47.63**
Qwen-Image-turbo-2steps	20B	1	40.44	**51.50**	69.34	86.06	40.03
Krea2-turbo-2steps	12B	1	13.86	3.67	51.24	74.62	30.47
TwinFlow	6B	1	29.18	36.85	66.34	85.17	42.20
NitroFusion2 (ours)	6B	2	76.03	81.95	74.46	86.21	51.45
Qwen-Image-turbo-2steps	20B	2	**89.42**	**93.00**	75.28	86.26	50.79
Krea2-turbo-2steps	12B	2	86.78	50.28	**85.52**	**86.67**	**52.59**
TwinFlow	6B	2	71.99	78.65	72.41	85.98	51.38



Figure 8: Additional two-step samples from NitroFusion2, at 1024×1024.

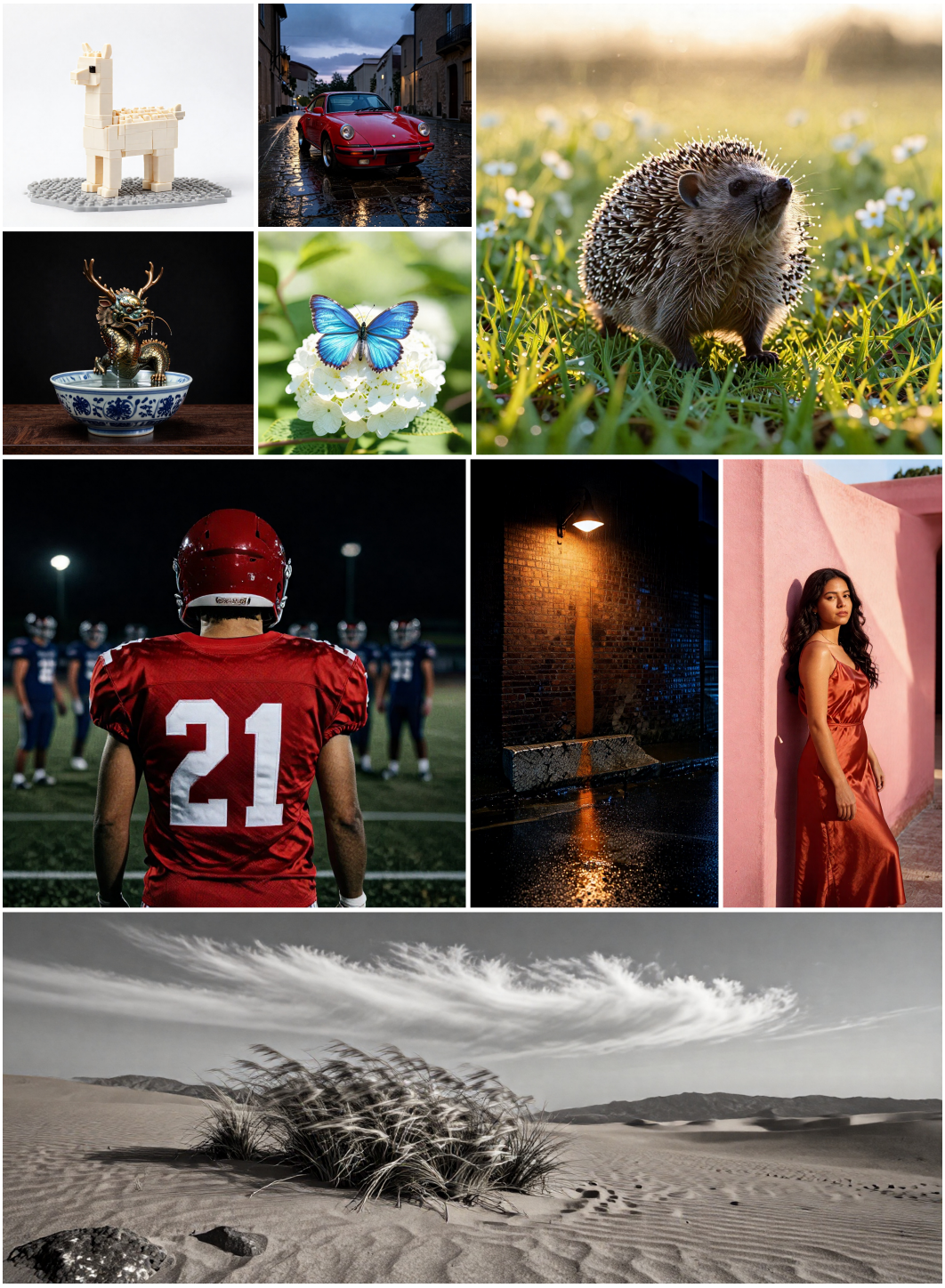

Figure 9: Additional single-step samples from NitroFusion2, at 1024×1024.

C PSEUDO-CODE

Algorithm 1 One training step of NitroFusion2. ACD draws the cue and routes the heads (lines 4–5). TSAD reads the text stream (line 9). Everything else is a standard adversarial distillation step.

Require: Frozen MMDiT backbone F; student G_θ; text-stream head groups $\{D^k_{\text{txt}}\}$, one per cue category; image-stream heads D_{img}; reference set $\mathcal{D}$ of pairs (x, c_{full}) with original prompt p

1: **while** not converged **do**
2: $(x, c_{\text{full}}, p) \sim \mathcal{D}, \quad z \sim \mathcal{N}(0, \mathbf{I})$
3: **Student output:** $\hat{x} \leftarrow G_\theta(z, 1, p)$
4: **ACD, draw the cue:** $k \sim \text{Uniform}(\mathcal{K}(a)), \quad q \leftarrow$ field k of c_{full}
5: **ACD, route the heads:** draw D_{txt} from the group D^k_{txt}
6: $t \sim \mathcal{U}, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad$ form x_t and $\hat{x}_t$
7: **One frozen pass per sample, both conditioned on q:**
8: $U_T, U_I \leftarrow F(x_t, q), \quad \hat{U}_T, \hat{U}_I \leftarrow F(\hat{x}_t, q)$
9: **TSAD, read the text stream:** $\ell^{\text{text}}_q \leftarrow D_{\text{txt}}(U_T), \quad \hat{\ell}^{\text{text}}_q \leftarrow D_{\text{txt}}(\hat{U}_T)$
10: **Image stream, as usual:** $\ell^{\text{img}}_q \leftarrow D_{\text{img}}(U_I), \quad \hat{\ell}^{\text{img}}_q \leftarrow D_{\text{img}}(\hat{U}_I)$
11: **Update the heads:** descend $\mathcal{L}_{\text{D}}$ on $D_{\text{txt}}, D_{\text{img}}$ with $\hat{x}_t$ detached
12: **Update the student:** descend $\mathcal{L}_{\text{G}} = (1 - \lambda_{\text{img}})\mathcal{L}_{\text{text-adv}} + \lambda_{\text{img}}\mathcal{L}_{\text{img-adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}$ on G_θ
13: **end while**
14: **return** trained student G_θ