TeEFusion: Blending Text Embeddings to Distill Classifier-Free Guidance

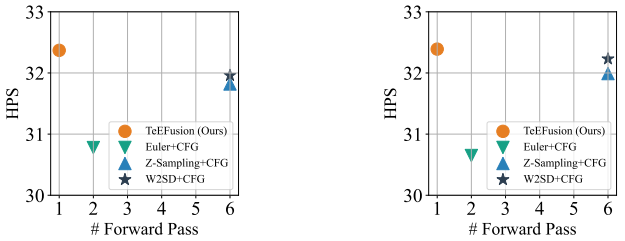
Minhao Fu*[1,2,3] Guo-Hua Wang†[3] Xiaohao Chen[3] Qing-Guo Chen[3]
Zhao Xu[3] Weihua Luo[3] Kaifu Zhang[3]
[1] School of Artificial Intelligence, Nanjing University
[2] National Key Laboratory for Novel Software Technology, Nanjing University
[3] Alibaba International Digital Commerce Group
fumh@lamda.nju.edu.cn, {wangguohua, xiaohao.cxh, qingguo.cqg, changgong.xz, weihua.luowh, kaifu.zkf}@alibaba-inc.com

Abstract

*Recent advances in text-to-image synthesis largely benefit from sophisticated sampling strategies and classifier-free guidance (CFG) to ensure high-quality generation. However, CFG's reliance on two forward passes, especially when combined with intricate sampling algorithms, results in prohibitively high inference costs. To address this, we introduce TeEFusion (**Te**xt **E**mbeddings **Fusion**), a novel and efficient distillation method that directly incorporates the guidance magnitude into the text embeddings and distills the teacher model's complex sampling strategy. By simply fusing conditional and unconditional text embeddings using linear operations, TeEFusion reconstructs the desired guidance without adding extra parameters, simultaneously enabling the student model to learn from the teacher's output produced via its sophisticated sampling approach. Extensive experiments on state-of-the-art models such as SD3 demonstrate that our method allows the student to closely mimic the teacher's performance with a far simpler and more efficient sampling strategy. Consequently, the student model achieves inference speeds up to 6× faster than the teacher model, while maintaining image quality at levels comparable to those obtained through the teacher's complex sampling approach. The code is publicly available at github.com/AIDC-AI/TeEFusion.*

1. Introduction

Models based on the Flow Matching [14, 16] pipeline, such as SD3 [5] and FLUX [11], have emerged as leading generative frameworks by directly synthesizing structured images from textual prompts. In the realm of text-to-image generation, their versatility is evidenced by a broad spectrum of applications, ranging from image editing [4] to creative concept generation for digital art [21].



(a) SD3 [5] (b) In-house T2I

Figure 1. Comparison of different sampling strategy on two DiT [22] models. "# Forward Pass" specifies the number of forward passes required in one denoising step.

Classifier-Free Guidance (CFG) [7] is a key technique for ensuring high-quality image synthesis by utilizing conditional and unconditional predictions to steer the generation toward regions with higher text-conditioned density. However, sampling with CFG demands two forward passes, one conditioned on the provided textual prompt and another on a null (or background) prompt, which introduces substantial computational overhead and slows down inference.

To mitigate this issue, guidance distillation [19] is commonly employed in state-of-the-art models, such as variants like FLUX.1-dev [11]. This approach integrates the guidance as an input parameter, reducing the number of required forward passes back to one. Consequently, guidance distillation has become both popular and crucial in state-of-the-art models, as it allows models to be scaled up without concern for cost or to employ more complex inference sampling algorithms. As illustrated in Fig. 1, advanced sampling strategies such as Z-Sampling [1]+CFG and W2SD [2]+CFG require almost 3× inference burden of the traditional Euler [16]+CFG counterpart, even though they yield higher image quality (as evidenced by HPS [34]), thereby highlighting the significant potential for efficiency

*Work done during the internship at Alibaba International Digital Commerce Group.
†G. Wang is the corresponding author.

Figure 2. Qualitative results by different prompting strategies. The leftmost three columns display images generated from different prompt inputs, while the final column illustrates the outcome of additively fusing semantic information from both the preserved and masked components.

optimization.

Several recent studies [13, 20, 24, 36–38, 41] attempt to advance distillation for image synthesis [19]. However, these approaches are confined to scenarios where both teacher and student models utilize the same sampling algorithm. In other words, they do not focus on distilling from particularly complex sampling algorithms while enabling the student model to rely solely on a simple sampling strategy. Moreover, these methods are overly complex, posing significant obstacles for implementation and adaptation. This issue is further amplified by the large-scale nature of modern state-of-the-art models, where tuning hyperparameters becomes increasingly challenging, ultimately hindering the deployment of these intricate methods.

As a result, these methods currently lack robust and large-scale validation, particularly in two critical aspects: 1) the inclusion of general benchmarks, such as DPG-Bench [8] or aesthetic scoring metrics like HPS [34]; and 2) testing on architectures that match the scale of the large DiT [22] used in state-of-the-art models like SD3 [5] and FLUX [11]. Consequently, it remains an open question whether these distillation techniques can be effectively applied to the latest, most advanced text-to-image generation models.

Motivated by these challenges, in this paper we explore more effective and efficient approaches for CFG distillation. We find that linear operations on different text features in the embedding space can effectively fuse or suppress aspects of the original text (cf. Fig. 2). Based on this observation, we propose TeEFusion (**Te**xt **E**mbeddings **Fusion**), a novel yet remarkably simple distillation method that directly leverages the underlying sampling formulation of CFG and integrates the guidance magnitude into the text embeddings. Specifically, our approach moves the linear combination of predictions from the conditional and unconditional model outputs forward, applying it directly to combine the corresponding text embeddings. The student model is optimized to learn the teacher model's denoised outputs, generated via a complex sampling strategy, thereby concurrently distilling both guidance signals and sampling procedures.

TeEFusion offers two primary advantages. First, compared to existing distillation techniques, our approach is exceptionally simple, with clear and easy-to-implement algorithmic principles. Unlike [20], which employs a feed-forward network to bridge the guidance magnitude, our method introduces no additional architecture—the structure of the distilled model remains identical to the undistilled model. Moreover, TeEFusion does not introduce any extra hyperparameters, ensuring that the distillation process remains efficient and easy to integrate into existing pipelines.

Secondly, we conduct a comprehensive evaluation of TeEFusion on state-of-the-art text-to-image models, including SD3 [5] and our In-house T2I, which demonstrates performance on par with SD3. Empirically results demonstrate that TeEFusion achieves significant improvements over baseline methods in terms of aesthetic scoring [34], object composition [8] and prompt-following ability [23].

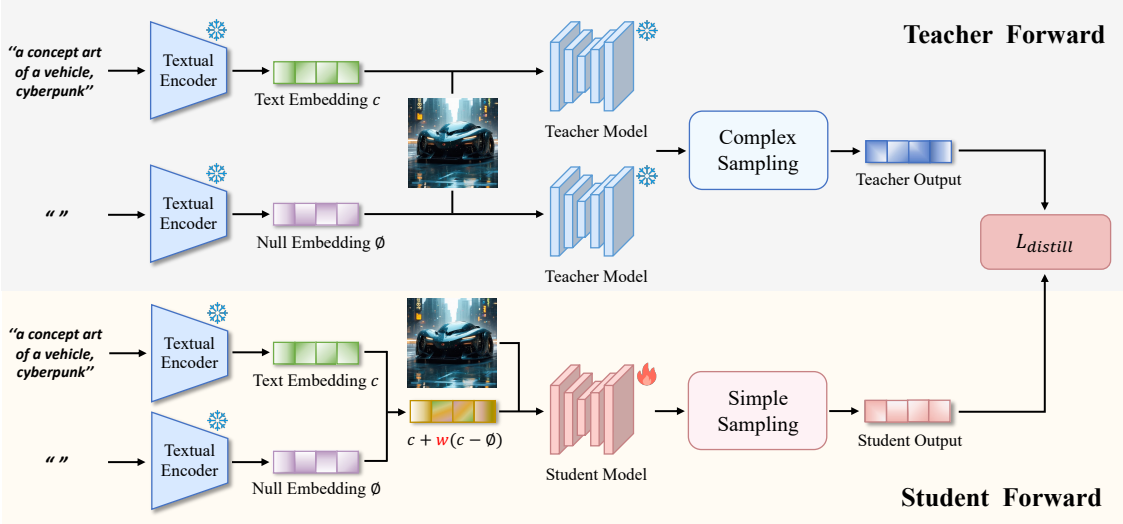


Figure 3. Illustration of TeEFusion. Our approach shifts the fusion of the DiT [22] model's outputs directly into the text embedding space, as specified in Eq. 7. This design is compatible with various sampling methods employed by the teacher model, offering significant potential for test-time scaling.

Moreover, our approach seamlessly accommodates the diverse samplers used by the teacher model. We hope that our evaluations will provide valuable insights for assessing distillation techniques in state-of-the-art, large-scale text-to-image models.

The contributions of this paper can be summarized as:

- We empirically find that for current state-of-the-art text-to-image generation models, adding or subtracting text embeddings from specific prompts effectively merges or alleviates certain visual concepts in generated images.
- We propose TeEFusion, a novel and simple distillation strategy that incorporates guidance from CFG directly into the text embeddings without introducing any extra parameters. Our approach is seamlessly compatible with a variety of complex sampling strategies employed by the teacher model.
- We evaluate TeEFusion on industrial-grade, large-scale text-to-image models, including state-of-the-art frameworks such as SD3 [5] and our In-house T2I model, using standard benchmarks like DPG-Bench [8], aesthetic metrics (HPS [34]), and prompt-following assessments [23]. Empirical results demonstrate that TeEFusion significantly outperforms baseline methods in terms of aesthetic quality, object composition, and prompt adherence.

2. Related Work

2.1. Test-Time Scaling

Recent research has increasingly focused on the scaling behavior of inference in diffusion models [7, 18]. RePaint [18] iteratively refined latent variables by reintroducing Gaussian noise to restore previous noise levels, establishing a reflection-based sampling paradigm that has been widely adopted in subsequent works [1–3, 33]. For example, Z-Sampling [1] changed denoised outputs in a zigzag manner by alternating forward and reverse passes with varying guidance magnitudes, thereby steering the random input toward semantically meaningful regions. W2SD [2] further extended this approach by explicitly incorporating both strong and weak models.

2.2. Diffusion Distillation

Early studies aimed to enhance the sampling efficiency of diffusion models by leveraging advanced numerical solvers [15, 28, 40]. Despite these improvements, significant room for further efficiency gains remains. Diffusion distillation [19] addresses this challenge by training a simpler student model to replicate the behavior of a more complex teacher model, primarily through step distillation [9, 24–26, 29, 39], thereby reducing the required inference steps. Some approaches [25, 26] incorporated adversarial training by introducing an additional discriminator to facilitate generalization of the score function. However, the complexity inherent to these methods limits their practical deployment, especially given the scale and complexity of contemporary diffusion models.

Among these methods, DistillCFG [20] is the most relevant, using an extra MLP to encode the guidance scale. DICE [42] also adds an attention-enhanced MLP but distills only a single fixed scale, which limits flexibility at inference. More variants such as MG [30] remain unevaluated on industrial-scale datasets and models. Additionally, existing distillation frameworks, such as progressive distil-

lation [24], have not explicitly addressed scenarios involving teacher models that adopt complex test-time sampling strategies, a situation increasingly prevalent in state-of-the-art diffusion models.

To bridge this critical gap, our proposed framework, TeEFusion, advances guidance distillation by explicitly incorporating the guidance magnitude into the linear combination of conditional and unconditional text embeddings. Furthermore, TeEFusion effectively generalizes across diverse complex sampling strategies employed by teacher models, addressing an essential yet largely overlooked aspect of diffusion model distillation.

3. Preliminaries

3.1. Classifier-Free Guidance

Classifier-Free Guidance (CFG) [7] enhances the alignment between generated images and textual descriptions by adjusting the sampling distribution:

$$\tilde{p}_\theta(x_t|c) \propto p_\theta(x_t|c)^{1+w} p_\theta(x_t)^{-w}, \tag{1}$$

where c represents the text prompt, $p_\theta(x_t|c)$ is the conditional distribution, and $p_\theta(x_t)$ denotes the unconditional counterpart. CFG refines the generation process by estimating the diffusion score[1] as:

$$\tilde{\epsilon}_\theta(x_t, c) = (1+w)\epsilon_\theta(x_t, c) - w\epsilon_\theta(x_t), \tag{2}$$

The scalar w determines the strength of guidance: larger values steer the synthesis more strongly towards the textual prompt, thereby promoting outputs that faithfully reflect the input text while reducing dependence on unconditioned predictions. Note that two forward passes are required to get the prediction from $\epsilon_\theta(x_t, c)$ and $\epsilon_\theta(x_t)$.

3.2. Sampling with Reflection.

Some methods [1, 2] employ sampling strategies that extend beyond the standard Euler method with reflection, refining the starting points of the denoising process with richer semantic information.

Weak-to-Strong Diffusion (W2SD) [2] leverages the discrepancy between a weak model $\mathcal{M}^w$ and a strong model $\mathcal{M}^s$ to bridge the gap toward an ideal generative distribution. In W2SD, a reflective operation updates the latent variable x_t as follows:

$$\tilde{x}_t = \mathcal{M}^w_{\text{inv}}(\mathcal{M}^s(x_t, t), t) \tag{3}$$

$$x_{t-1} = \mathcal{M}^s(\tilde{x}_t, t), \tag{4}$$

thereby guiding the sampling trajectory toward regions of higher text density. Z-Sampling [1] is a simpler version of W2SD by sharing the weights between $\mathcal{M}^s$ and $\mathcal{M}^k$, but using different guidance signals.

However, the extra inversion step where the weak model processes the denoised output of the strong models introduces two additional forward passes per sampling iteration. Consequently, the time cost of reflection sampling is $3\times$ compared to standard sampling (one for standard denoising, one for inversion, and one for subsequent denoising). When combined with CFG, the overall time cost escalates to nearly $6\times$.

[1] For notational simplicity, we denote the score prediction for both diffusion and flow matching models as ϵ_θ.

Algorithm 1 The distillation pipeline of TeEFusion.

Require: Teacher model $\epsilon_{\theta_\text{T}}$, Dataset D
$\epsilon_{\theta_\text{S}} \leftarrow \epsilon_{\theta_\text{T}}$ ▷ Initialize student using teacher weights
while not converged **do**
  $x_0, c \sim \mathcal{D}$ ▷ Sample data
  $t \sim U[0,1]$ ▷ Sample time
  $w \sim U[w_\text{min}, w_\text{max}]$ ▷ Sample guidance
  $\epsilon \sim N(0, I)$ ▷ Sample noise
  $x_t = (1-t)x_0 + t\epsilon$ ▷ Add noise to data
  $\tilde{\epsilon}_{\theta_\text{T}}(x_t, w, c) = (1+w)\epsilon_{\theta_\text{T}}(x_t, t, c) - w\epsilon_{\theta_\text{T}}(x_t, t, \varnothing)$ ▷ Compute target using CFG
  if Reflection **then**
    $\tilde{\epsilon}_{\theta_\text{T}}(x_t, w, c) \leftarrow$ Refine $\tilde{\epsilon}_{\theta_\text{T}}(x_t, w, c)$ using Eq. 3 and 4
  end if
  Compute $\hat{z}_{t,c,\varnothing,w}$ using Eq. 7
  $L_\text{distll} = \|\epsilon_{\theta_\text{S}}(x_t, \hat{z}_{t,c,\varnothing,w}) - \tilde{\epsilon}_{\theta_\text{T}}(x_t, w, c)\|_2^2$ ▷ Loss
  $\epsilon_{\theta_\text{S}} \leftarrow \epsilon_{\theta_\text{S}} - \gamma \cdot \nabla_{\epsilon_{\theta_\text{S}}} L_\text{distill}$ ▷ Optimization
end while
return $\epsilon_{\theta_\text{S}}$ ▷ Output the trained student model

However, the extra inversion step where the weak model processes the denoised output of the strong models introduces two additional forward passes per sampling iteration. Consequently, the time cost of reflection sampling is $3\times$ compared to standard sampling (one for standard denoising, one for inversion, and one for subsequent denoising). When combined with CFG, the overall time cost escalates to nearly $6\times$.

4. An In-Depth Analysis of CFG

We begin by examining the mechanism underlying CFG. As indicated by Eq. 2, the output of CFG can be expressed as a linear combination of the conditional prediction $\epsilon_\theta(x_t|c)$ (with coefficient $1+w$) and the unconditional prediction $\epsilon_\theta(x_t)$ (with coefficient $-w$). This formulation motivates the idea of moving this linear combination earlier in the processing pipeline, thereby reducing the need for two separate forward passes to just one. A natural approach is to integrate this combination directly into the text embeddings to efficiently incorporate the guidance magnitude into the model as:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \hat{c}), \tag{5}$$

where $\hat{c} = (1+w)c - w\varnothing = c + w(c - \varnothing)$ is marked as the *fusion text embedding* with $\varnothing$ from the null prompt to provide the background for generation. However, the success of this approach depends on the assumption that linearly combining input conditions is meaningful.

Therefore, we perform a preliminary test to verify

whether merging the conditional and unconditional text embeddings into a single text embedding can yield the necessary semantic representations. In this experiment, we randomly masked certain components of the prompt, replacing them with a *[mask]* token. We then generated images under three different conditions:

1. Using only the unmasked portion of the prompt.
2. Using only the masked components (i.e., the content that was replaced by *[mask]*).
3. Using the original, unaltered prompt.
4. Both the unmasked and masked components are converted into text embeddings, which are then merged via element-wise addition. The resulting fused embedding is subsequently used for image generation.

This setup enables us to evaluate whether the additive (or subtractive) fusion of semantic information from both preserved and masked prompt components yields meaningful features that can guide high-quality image synthesis via a linear combination (cf. Eq. 5) in the text embedding space. As shown in Fig. 2, images generated by adding the text embeddings of the preserved and masked parts match (or even surpass) the quality of those produced from the original, unaltered prompt. These findings indicate that additive operations in the text embedding space can effectively merge diverse prompt patterns, thereby highlighting the potential of our sampling technique to use Eq. 5 as a surrogate to integrate guidance magnitudes while keeping high efficiency during sampling (see Sec. A.1 for quantitative results).

5. TeEFusion to Distill CFG

The analysis in Sec. 4 demonstrates that simple addition or subtraction of text embeddings from different prompts can effectively reconstruct or eliminate specific semantic components of the original prompt, indicating that such operations in the embedding space yield features with clear semantics. Building on this observation, we hypothesize that any linear combination of these text embeddings—with appropriately moderate coefficients can similarly produce semantically meaningful representations. This insight motivates the use of Eq. 5 for guidance distillation with a properly configured w.

Motivated by this hypothesis, we propose TeEFusion (**Te**xt **E**mbeddings **Fusion**), a simple distillation framework that effectively distills both classifier-free guidance and complex reflection-based sampling techniques simultaneously. In TeEFusion, guidance magnitude w is incorporated into the student model by linearly combining conditional and unconditional text embeddings, with w serving as the combination coefficient. Subsequently, the student model is trained to mimic the denoised output of teacher models that employ reflection-based sampling (see Fig. 3).

However, directly applying Eq. 5 to TeEFusion poses challenges. As the guidance scale w increases, the variance of the fused prompt $\hat{c}$ grows on the order of $\mathcal{O}(w^2)$, which, when w becomes too large, leads to poor numerical stability. To address this issue, we project w into a vector space using a sine and cosine time embedding [32]. This approach not only ensures that different values of w remain distinguishable, but also mitigates the numerical instability arising from excessive variance.

Specifically, the joint embedding of the text prompt and timestep in the text-to-image model [5, 11] $\epsilon(x_t, t, c)$ is formulated as:

$$z_{t,c} = \mathcal{G}(\psi(t)) + \mathcal{F}(c), \tag{6}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ denote two distinct multi-layer perceptions (MLPs), and $\psi(\cdot)$ represents the sinusoidal encoding of the timestep. This formulation effectively decouples the processing of textual and temporal information, thereby enabling a more structured integration of the guidance magnitude.

Our TeEFusion distills the guidance magnitude w by incorporating it into $z_{t,c}$ as:

$$\hat{z}_{t,c,\varnothing,w} = \mathcal{G}(\psi(t)) + \mathcal{F}(c) + \underbrace{\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)}_{\text{extra term}}, \tag{7}$$

where the term $c - \varnothing$ is computed after projection into the text embedding space. Assume that the teacher and student models are denoted by $\epsilon_{\theta_\mathrm{T}}$ and $\epsilon_{\theta_\mathrm{S}}$, respectively. After obtaining $\hat{z}_{t,c,\varnothing,w}$, it is used to compute the guided output of student model. The distillation objective then lets the student mimic the CFG-derived output from the teacher model with complex sampling (cf. Algorithm 1).

Comparing Eq. 7 with the definition of $\hat{c}$, the term $\mathcal{F}(c)+\mathcal{G}(\psi(w))\mathcal{F}(c-\varnothing)$ exactly corresponds to the term $c+w(c-\varnothing)$ in $\hat{c}$, but computed in the embedding space.

TeEFusion offers several notable advantages. First, its w-injection method is exceptionally simple. Unlike other approaches [20, 25, 26] that require extra network structures and parameters, our method is very easy to implement. Second, it demonstrates strong scalability. By employing more robust teacher sampling strategies or larger network architectures, our approach can further empower the student model. Finally, it is easy to train and converges quickly, as evidenced by the results in Fig. 5b. These features make TeEFusion an attractive and practical solution for efficient guidance distillation in state-of-the-art text-to-image generation systems.

6. Experiments

In this section, we begin by describing the experimental setups. Next, we present the main results, demonstrating the effectiveness of TeEFusion through both quantitative and qualitative analyses. Finally, we conduct comprehensive ablation studies to investigate the contributions of each component in our approach and to evaluate its robustness under various configurations.

Figure 4. Qualitative comparison among various methods across different guidance scales w. Prompts: 1) *An egirl with pink hair and extensive makeup.* 2) *A cozy tavern with a retro video game vibe and cinematic lighting, featuring a cartoon-style animation and detailed background art.*

Table 1. HPS results (↑) on aesthetic benchmarks. Higher scores indicate that the generated images better conform to human aesthetic standards. "Cost" refers to the inference time required, measured and compared separately for each model. "I.-T2I" is the abbreviation for "In-house T2I". Results of the student model are from the teacher model using W2SD+CFG. The best results are in bold face.

Model	Method	Cost	Prompt Group			
			Anime	Concept-Art	Paintings	Photo
SD3	*Teacher*					
	CFG	2×	30.78	30.06	30.28	27.93
	W2SD+CFG	6×	31.96	30.65	30.67	29.76
	Student					
	DistillCFG	1×	31.14	29.52	30.03	29.04
	TeEFusion	1×	**32.37**	**30.88**	**30.74**	**29.84**
I.-T2I	*Teacher*					
	CFG	2×	30.65	29.27	28.94	27.99
	W2SD+CFG	6×	32.23	31.29	31.22	29.93
	Student					
	DistillCFG	1×	30.80	29.41	29.14	28.79
	TeEFusion	1×	**32.39**	**31.34**	**31.27**	**29.96**

6.1. Experimental Settings

Models. We employ two text-to-image generation models that both leverage the DiT [22] architecture and Flow Matching [16] framework: SD3 [5] and our In-house T2I. SD3 [5] is a publicly available open source model with 2B parameters, offering superior visual quality and enhanced compositional consistency. In-house T2I, is optimized for photorealistic images in e-commerce scenarios and consists of 1B parameters.

Training Details. We employ the subset of LAION-5B [27] for distillation. In particular, high-resolution images with an aesthetic score [27] exceeding 6 are selected as the training set. The student model is optimized using AdamW [17] with an effective batch size of 128 and a default learning rate of 5×10^{-6}. For the teacher model, we adopt W2SD [2]+CFG [7] as the sampling strategy, given its superior performance as demonstrated in Fig. 1. Since W2SD leverages two models of differing strengths during sampling, these models are obtained using CHATS [6]. All distilled models in our experiments are trained within the W2SD+CFG framework, with w_{min} and w_{max} in Algorithm 1 set as 2 and 14, respectively.

Baselines. Given the limited research on directly applying guidance distillation to large text-to-image models, we adopt DistillCFG [20] as our baseline. DistillCFG integrates the guidance scale through an additional MLP embedding as in recent FLUX.1-dev [11]. Moreover, we compare TeEFusion with the teacher model to further validate its effectiveness.

Evaluation Details. *HPS* [34] provides a broad benchmark for aesthetic evaluation, featuring 3,200 prompts distributed equally among four distinct styles: *Anime*,

Table 2. Quantitative results on CLIP score and DPG-Bench (%). Higher scores indicate that the generated images exhibit superior overall performance, reflecting enhanced visual quality, finer details, and better alignment with the input prompt. †: measured using prompt from Anime.

Model	Method	CLIP†	DPG-Bench					
			Global ↑	Entity ↑	Attribute ↑	Relation ↑	Other ↑	Overall ↑
SD3	*Teacher*							
	CFG	34.93	85.71	90.84	87.79	93.58	86.40	85.09
	W2SD+CFG	34.96	82.98	92.13	88.60	94.24	91.60	86.56
	Student							
	DistillCFG	34.53	81.46	90.47	87.85	**93.42**	84.40	84.13
	TeEFusion	**34.63**	**81.76**	**90.60**	**88.11**	93.15	**86.80**	**84.56**
In-house T2I	*Teacher*							
	CFG	34.25	83.89	88.08	88.07	91.57	76.40	81.88
	W2SD+CFG	34.62	82.37	91.41	89.08	93.73	78.80	85.20
	Student							
	DistillCFG	33.38	**84.50**	**90.21**	88.24	93.54	80.00	83.52
	TeEFusion	**33.81**	84.19	90.08	**88.56**	**93.73**	**81.60**	**84.13**

Table 3. Modular ablation of TeEFusion using In-house T2I with prompts from Anime. The "Config" of TeEFusion specifies the configuration of "*extra term*" in Eq. 7.

Config	Metrics			
	HPS ↑	IR ↑	PickScore ↑	CLIP ↑
DistillCFG	30.80	113.12	22.42	33.38
$\mathcal{G}(\psi(w))$	31.05	116.39	22.49	33.59
$\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)$	32.39	128.50	22.68	33.81

Concept-Art, *Paintings* and *Photo* (800 prompts each). This extensive prompt collection supports robust and consistent evaluation outcomes. *DPG-Bench* [8] includes 1,065 prompts, with each prompt describes several objects with different features and explains how these objects relate to each other. It is designed to test how well text-to-image models can combine these elements into a coherent image. For aesthetic evaluation, we employ three state-of-the-art evaluators: *HPS* [34] (with its v2 version), *ImageReward (IR)* [35], and *PickScore* [10]. We also use the *CLIP* [23] score to assess how well the generated images align with their corresponding input prompts.

6.2. Main Results.

As demonstrated in Table 1, our TeEFusion method consistently achieves the best scores on aesthetic evaluations, surpassing the DistillCFG baseline and even the teacher model across all prompt categories for both SD3 and In-house T2I models. Table 2 further highlights TeEFusion's superior performance on the DPG-Bench, where it not only excels in object attributes, spatial relationships, and overall composition, but also achieves higher CLIP scores compared to DistillCFG, underscoring its ability to faithfully capture textual semantics.

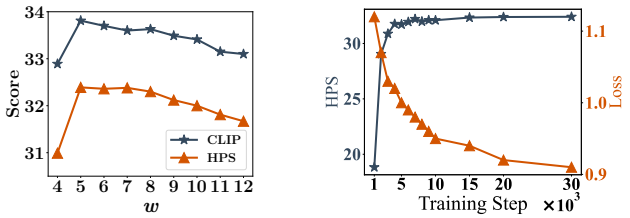
These results clearly indicate that TeEFusion effectively integrates the guidance magnitude through linear fusion in the text embedding space, thereby preserving semantic structure and compositional quality while significantly improving sampling efficiency. Notably, TeEFusion achieves an inference speed that is 6× faster than that of the teacher model. Overall, TeEFusion successfully distills complex sampling strategies into a streamlined inference process, delivering high image quality with minimal trade-offs, and making it a promising approach for real-world applications.

Table 4. Ablation study on distilling with different sampling strategies. In each group, the first row shows the teacher's sampling strategy, while the second row displays TeEFusion distilled with it. Model: In-house T2I. Prompt: Anime.

Sampling	Cost	Metrics			
		HPS ↑	IR ↑	PickScore ↑	CLIP ↑
Euler+CFG	2×	30.65	118.05	22.79	34.25
TeEFusion	1×	30.61	117.29	22.78	33.58
Z-Sampling+CFG	6×	31.99	125.42	22.61	34.47
TeEFusion	1×	32.01	125.07	22.65	33.62
W2SD+CFG	6×	32.23	127.56	22.49	34.62
TeEFusion	1×	32.39	128.50	22.68	33.81

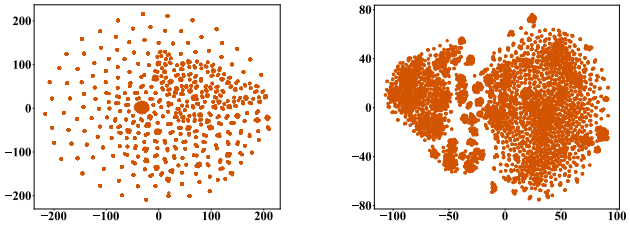
6.3. Ablation Studies

Modular Ablation. We conduct modular ablation on In-house T2I using prompts from the Anime (see Table 3). By replacing the additional MLP used in DistillCFG with $\mathcal{G}(\psi(w))$, we observe an improvement in all scores. Furthermore, the complete TeEFusion configuration achieves the best results, consistently outperforming all other variants. These findings demonstrate that each component within TeEFusion meaningfully contributes to the effectiveness of guidance distillation. Notably, the student model employs simple Euler sampling without CFG.



(a) Guidance scale w (b) Training Step

Figure 5. Ablation study on varying (a) guidance scale w and (b) the number of training steps in In-house T2I.



(a) DistillCFG (b) TeEFusion

Figure 6. Illustration of embedded w, with values randomly sampled from the range [2, 14], visualized using t-SNE [31].

Switching Teacher Sampling. We also investigate the effectiveness of TeEFusion under different sampling strategies used by teacher models (cf. Table 4). The results indicate that TeEFusion achieves nearly lossless distillation regardless of the teacher sampling method. Although we observe minor performance reductions when distilling from simpler sampling methods in Euler+CFG and Z-Sampling+CFG, these slight differences are expected, as perfectly matching the teacher model is challenging.

Importantly, when distilling from the more robust sampling method (W2SD+CFG), TeEFusion not only achieves near-lossless distillation but slightly surpasses the teacher's performance across several metrics. This demonstrates that TeEFusion effectively preserves model quality during distillation. Moreover, these results indicate a promising property of TeEFusion: its performance is expected to further improve as stronger sampling strategies or larger teacher models become available in the future, making our method highly scalable and adaptable for test-time scaling.

Is w-distillation successful? We further examine the influence of varying the guidance scale w on TeEFusion's performance. As illustrated in Fig. 4, the images produced by TeEFusion under different w settings are both diverse and visually appealing. Moreover, the model achieves the highest HPS and CLIP scores at $w = 5$ (see Fig. 5a), indicating that an optimal guidance scale effectively enhances both aesthetic quality and prompt adherence. The distinct score variations across different w values further confirm that TeEFusion maintains sensitivity to guidance scale changes after distillation. These findings validate the capability of our approach to preserve meaningful guidance scale characteristics, thereby enabling users to effectively balance text-image consistency and visual aesthetics.

How fast does TeEFusion converge? We further examine the convergence speed of TeEFusion in Fig. 5b. It reveals that the HPS value increases rapidly during the early stages of training (below 5k steps) and then gradually saturates. Similarly, the loss curve drops quickly at first and levels off once the training steps exceed 10k. These findings clearly indicate that TeEFusion converges remarkably fast. This rapid convergence can be attributed to the fact that all encoding operations related to the guidance scale w utilize modules directly inherited from a pretrained model, without any randomly initialized components. In fact, for In-house T2I, TeEFusion converges in under 4 hours when trained on 16 A100 GPUs, demonstrating an exceptionally efficient training process.

Qualitative Analysis. In Fig. 4, we present the qualitative results of various methods across different w. We observe that the images generated by the teacher model using W2SD+CFG do not consistently achieve optimal aesthetic quality, particularly when w is excessively high. Notably, at high w values, W2SD+CFG is more prone to collapse. For example, the left image appears less appealing, while the right image exhibits pronounced periodic artifacts [12]. In contrast, the distilled models do not suffer from these issues, which explains why TeEFusion attains higher aesthetic scores compared to W2SD+CFG due to its enhanced stability. Moreover, we consistently observed that TeEFusion outperforms DistillCFG, especially at lower w values.

Additionally, in Fig. 6 we illustrate the feature distributions of embedded w. An interesting observation is that the embedded w obtained by TeEFusion exhibit a more coherent distribution, whereas those of DistillCFG appear markedly more discrete. This can be attributed to TeEFusion's additional utilization of information from the $c - \varnothing$, which results in a smoother guidance signal and, consequently, improved generation quality.

7. Conclusions and Limitations

In this paper, we introduced TeEFusion, an efficient distillation method that fuses guidance magnitudes into text embeddings, streamlining the inference process for text-to-image synthesis. By linearly combining conditional and unconditional embeddings, TeEFusion replicates the teacher model's performance, achieving comparable image quality while enabling up to 6× faster inference. However, TeEFusion sometimes produces semantic inconsistencies (e.g., mismatched attributes) and its outputs do not always precisely align with the teacher model (cf. Fig. 4). Future work will focus on mitigating these issues.

References

[1] Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *International Conference on Learning Representations*, pages 1–14, 2025. 1, 3, 4

[2] Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. *arXiv:2502.00473*, 2025. 1, 3, 4, 6

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 843–852, 2023. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, pages 12606–12633, 2024. 1, 2, 3, 5, 6

[6] Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. CHATS: Combining human-aligned optimization and test-time sampling for text-to-image generation. *arXiv:2502.12579*, 2025. 6

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 1, 3, 4, 6

[8] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. EllA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv:2403.05135*, 2024. 2, 3, 7

[9] Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, pages 26565 – 26577, 2022. 3

[10] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 36652–36663, 2023. 7

[11] Black Forest Labs. FLUX.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1, 2, 5, 6

[12] Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv:2502.01993*, 2025. 8

[13] Yanyu Li, Huan Wang, Qing Jin, et al. SnapFusion: Text-to-image diffusion model on mobile devices within two seconds. In *Advances in Neural Information Processing Systems*, pages 20662–20678, 2023. 2

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, pages 1–13, 2023. 1

[15] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, pages 1–11, 2022. 3

[16] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, pages 1–15, 2023. 1, 6

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, pages 1–18, 2019. 6

[18] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11451–11461, 2022. 3

[19] Zhiyuan Ma, Yuzhu Zhang, Guoli Jia, et al. Efficient diffusion models: A comprehensive survey from principles to practices. *arXiv:2410.11795*, 2024. 1, 2, 3

[20] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2, 3, 5, 6

[21] OpenAI. DALL·E 2, 2022. https://openai.com/dall-e-2/. 1

[22] William Peebles and Saining Xie. Scalable diffusion models with Transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 4172–4182, 2023. 1, 2, 3, 6

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 7

[24] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, pages 1–21, 2022. 2, 3, 4

[25] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia*, pages 1–11, 2024. 3, 5

[26] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103, 2024. 3, 5

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in neural information processing systems*, pages 25278–25294, 2022. 6

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, pages 1–12, 2022. 3

[29] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211 – 32252, 2023. 3

[30] Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv:2502.12154*, 2025. 3

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 8

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 6000–6010, 2017. 5

[33] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, et al. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE/CVF International Conference on Computer Vision*, pages 7589–7599, 2023. 3

[34] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv:2306.09341*, 2023. 1, 2, 3, 6, 7

[35] Jiazheng Xu, Xiao Liu, Yuchen Wu, et al. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 15903–15935, 2024. 7

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

[37] Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.

[38] Bowen Zheng and Tianming Yang. Diffusion models are innate one-step generators. *arXiv:2405.20750*, 2024. 2

[39] Hongkai Zheng, Weilie Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International Conference on Machine Learning*, pages 42390 – 42402, 2023. 3

[40] Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. DPM-Solver-v3: Improved diffusion ode solver with empirical model statistics. In *Advances in Neural Information Processing Systems*, pages 55502 – 55542, 2024. 3

[41] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity Distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, pages 62307–62331, 2024. 2

[42] Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. DICE: Distilling classifier-free guidance into text embeddings. *arXiv preprint arXiv:2502.03726*, 2025. 3

Figure 7. Generation examples of failure cases. Prompt: *1) not a cat. 2) liquid glass. 3) cold fire.*

A. More Experimental Results and Analyses

A.1. Quantitative Analysis of Additive Text Embeddings

To validate the effectiveness of additive text embeddings, we conducted quantitative experiments across different text-to-image models. The cosine similarity between original and fused embeddings (Cos Sim._{txt}) and their corresponding generated images (Cos Sim._{img}) are summarized in the table below:

Metric	SD3	In-house T2I	FLUX.1-dev
Cos Sim._{txt}	0.8073	0.8192	0.8286
Cos Sim._{img}	0.8732	0.9137	0.9318

These results confirm that additive embedding operations preserve over 80% cosine similarity in text space and over 90% in image space, demonstrating their ability to merge diverse semantic patterns effectively.

A.2. Operational Boundaries and Failure Cases

Our fusion mechanism $\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)$ operates within the encoder's linear regime through bounded sine-cosine positional encodings ($\|\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)\|_2 \leq \delta$). However, failure cases arise when:

- Semantic vectors exhibit non-orthogonality (e.g., contradictory phrases like "cold fire")
- Contextual interference occurs in composite prompts (e.g., "not a cat")

These limitations are visualized in Figure 7.