TPSO: Training-Free Diverse Image Generation via Semantic Prompt Embedding Optimization

Debin Meng[1], Chen Jin[2], Zheng Gao[1], Yanran Li[3], Ioannis Patras[1], Georgios Tzimiropoulos[1]
[1]Queen Mary University of London, [2]Centre for AI, AstraZeneca, UK, [3]University of Bedfordshire
{debin.meng, z.gao, i.patras, g.tzimiropoulos}@qmul.ac.uk, chen.jin@astrazeneca.com
Yanran.Li@beds.ac.uk

***Abstract*—Image diversity remains a fundamental challenge for text-to-image diffusion models. Low-diversity generation often leads to repetitive outputs, increasing sampling redundancy and hindering both creative exploration and downstream applications. A key factor is the tendency of diffusion models to collapse toward *strong modes in the learned distribution*. Existing attempts to improve diversity, such as steering-based guidance, often introduce distortions that degrade image quality. To address this issue, we propose Token-Prompt Embedding Space Optimization (TPSO), a training-free and model-agnostic module. TPSO introduces learnable parameters to explore underrepresented regions of the token embedding space, reducing the tendency to repeatedly sample from strong modes of the distribution. Meanwhile, a prompt-level semantic constraint regulates distribution shifts, preventing quality degradation while preserving semantic fidelity. Extensive experiments on MS-COCO across three representative diffusion backbones demonstrate that TPSO substantially improves diversity, boosting performance from 1.10 to 4.18, while maintaining image quality with only a modest inference-time overhead of 3.6% to 8.9%. *Code is available at: https://github.com/Open-Debin/TPSO* .**

***Index Terms*—Text-to-Image Generation, Generative Diversity, Diffusion Models, Prompt Embedding Optimization, Training-Free Methods.**

I. Introduction

Powered by recent developments in diffusion models [1, 2] and contrastively trained text encoders [3], modern text-to-image systems are now capable of producing visually compelling and semantically faithful images from diverse natural-language prompts [4]. A fundamental attribute of generative models is their ability to produce diverse outputs under the same condition. Ideally, a generative model should generate samples that are both consistent with the conditioning input and diverse in appearance. For example, a prompt describing "a wooden chair" can produce many valid outcomes that differ in shape, material texture, or stylistic design. However, Astolfi et al. [5] point out that current diffusion models often suffer from limited diversity: state-of-the-art Text-to-Image models tend to improve consistency and/or realism by sacrificing representation diversity. This lack of diversity leads to highly similar outputs under the same conditions, resulting in redundant resampling and increased computational cost, while also constraining creative exploration and impairing downstream applications such as data generation for model training, where repetitive samples reduce dataset utility. CADS [6] and Particle Guidance [7] attribute this phenomenon largely to mode collapse induced by classifier-free guidance (CFG) [8], a standard practice in diffusion models for enforcing fidelity. Under the strong guidance of CFG [8] tends to push the mini-batch samples towards a strong mode within the learned distribution to enhance fidelity, at the expense of diversity [7].

To address the limited diversity caused by CFG [8], Particle Guidance [7] optimizes intermediate latents within a batch by repelling samples from each other to enforce separation. Sadat et al. [6] further point out that the limited diversity of CFG stems from strong conditional guidance, which pushes samples toward strong modes of the learned distribution. Based on this insight, they propose Dynamic Classifier-Free Guidance (CFG) and CADS. Dynamic CFG [6] reduces the effect of CFG by dynamically modifying the guidance strength, while CADS [6] adjusts the conditional strength by interpolating Gaussian noise with the conditional signal. Although these approaches can improve diversity, they lack semantic constraints and therefore often push samples toward semantically implausible regions, which frequently introduces unintended distortions and degrades image quality (See Fig. 5).

To address these limitations, we introduce Token–Prompt Embedding Space Optimization (TPSO), which focuses on solving two core problems: generation sampled in strong modes of the learned distribution and quality degradation when promoting diversity. The core idea of TPSO is to jointly operate in two complementary representation spaces. First, it optimizes learnable offsets in the token embedding space, enabling exploration of underutilized token representations, reducing the tendency to repeatedly sample from strong modes of learned distribution, thereby encouraging diverse generation. Second, this exploration is regulated by semantic alignment losses in a high-level prompt embedding space, which preserves semantic consistency and prevents the structural distortions commonly observed in prior methods. Importantly, TPSO is an inference-stage optimization module, making it fully training-free, lightweight, and model-agnostic. This design allows TPSO to seamlessly integrate into existing text-to-image pipelines to enhance generative diversity without compromising image quality.

Our key contributions are summarized as follows:

- We introduce a generic, lightweight, training-free, and model-agnostic module for enhancing generative diversity in representative diffusion pipelines.
- We present TPSO, a dual-space optimization module equipped with diversity and semantic alignment losses.

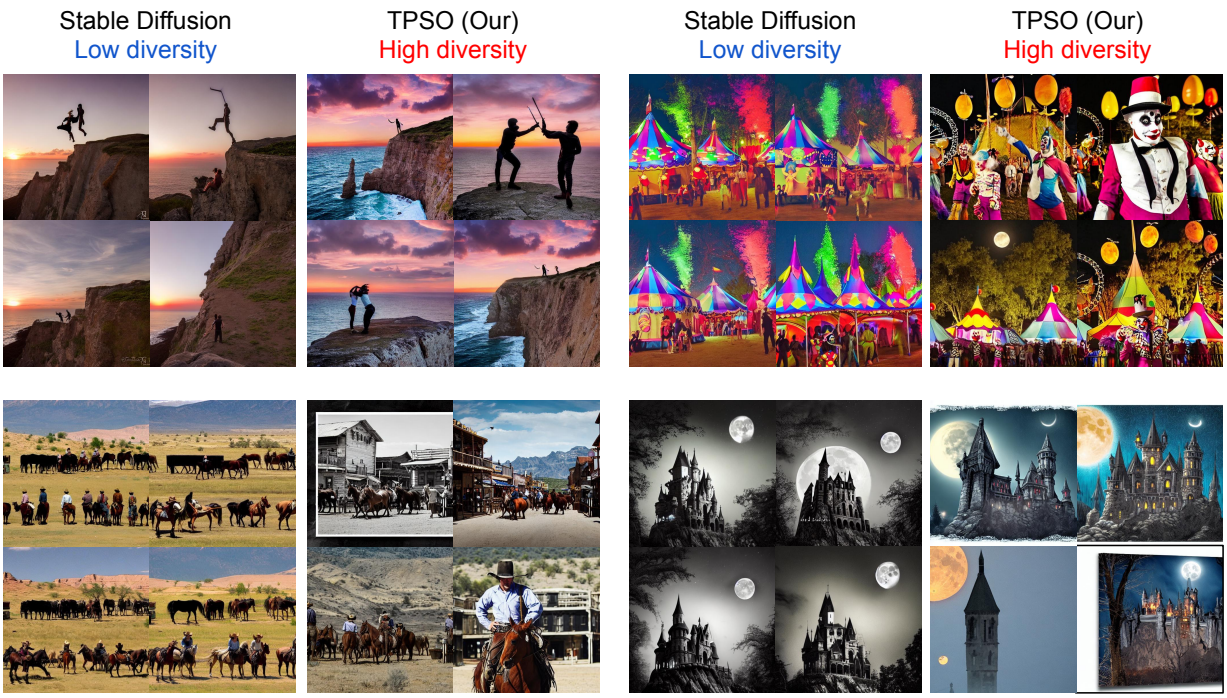

Fig. 1. Baseline diffusion models tend to produce low-diversity or nearly repeated outputs across sampling runs under the same condition, leading to redundant sampling efforts. In contrast, our approach enhances output diversity, yielding a broader range of plausible and distinct visual outcomes.

Guided by these losses, TPSO explores underutilized token-level representations via learnable offsets to mitigate strong-mode sampling in the learned distribution, thereby encouraging diverse generation, while regulating semantic deviation in the prompt embedding space to preserve image fidelity. A coarse-to-fine scheduling strategy is further introduced to inject diversity at early denoising stages while maintaining high-quality refinement in later steps.

- Extensive experiments on MS-COCO across three representative diffusion backbones demonstrate that TPSO substantially improves sample diversity (from 1.10 to 4.18) while maintaining image quality with minimal inference-time overhead.

II. Related work

Conditional image generation has attracted substantial attention from both academia and industry due to its ability to synthesize realistic and aesthetically compelling images. This technology supports a wide range of applications in design, entertainment, advertising, and creative content generation [4, 9, 10]. To improve controllability and alignment with user intent, conditional generative models guide the generation process using explicit input signals such as text [4, 11], reference images [12, 13, 14, 15], spatial layouts [16, 17, 18] or combinations [19, 20]. Although these approaches improve semantic control and fidelity, most existing efforts focus mainly on image quality and conditional alignment, often overlooking diversity, a fundamental attribute of generative models [5].

Diversity in Image generation Diversity reflects a generative model's ability to produce multiple varied outputs under the same conditions and is essential for creative exploration and downstream applications. [5]. However, models often suffer from mode collapse, with samples concentrating around dominant modes [21]. Recently, diffusion models [4, 2] have largely replaced GANs [22] as the mainstream generative paradigm, with prior studies [23, 24] showing that GANs capture less diversity than diffusion models. Existing approaches attempt to enhance diversity in diffusion models by perturbing intermediate latents [7] or adjusting text embeddings [6]. While effective at separating samples, these techniques often push the generation trajectory into unstable regions of the latent space, resulting in degraded fidelity. A concurrent approach [25] trains a prompt encoder to generate multiple alternative prompts to increase diversity. However, it requires additional training and is not explicitly designed to address strong-mode sampling in the learned distribution. In contrast, our TPSO module is entirely training-free and directly explores underutilized token-embedding subspaces to mitigate strong-mode sampling in the learned distribution, while enforcing semantic constraints in the prompt-embedding space to preserve fidelity. As a result, TPSO achieves increased sample diversity without sacrificing image quality.

Prompt Feature Optimization for Diffusion Models Textual Inversion [26] and MCPL [27] optimize new token embeddings in the Stable Diffusion text encoder to represent novel visual concepts. Similarly, MinorityPrompt [28] optimizes token embeddings to improve the generation of minority samples,

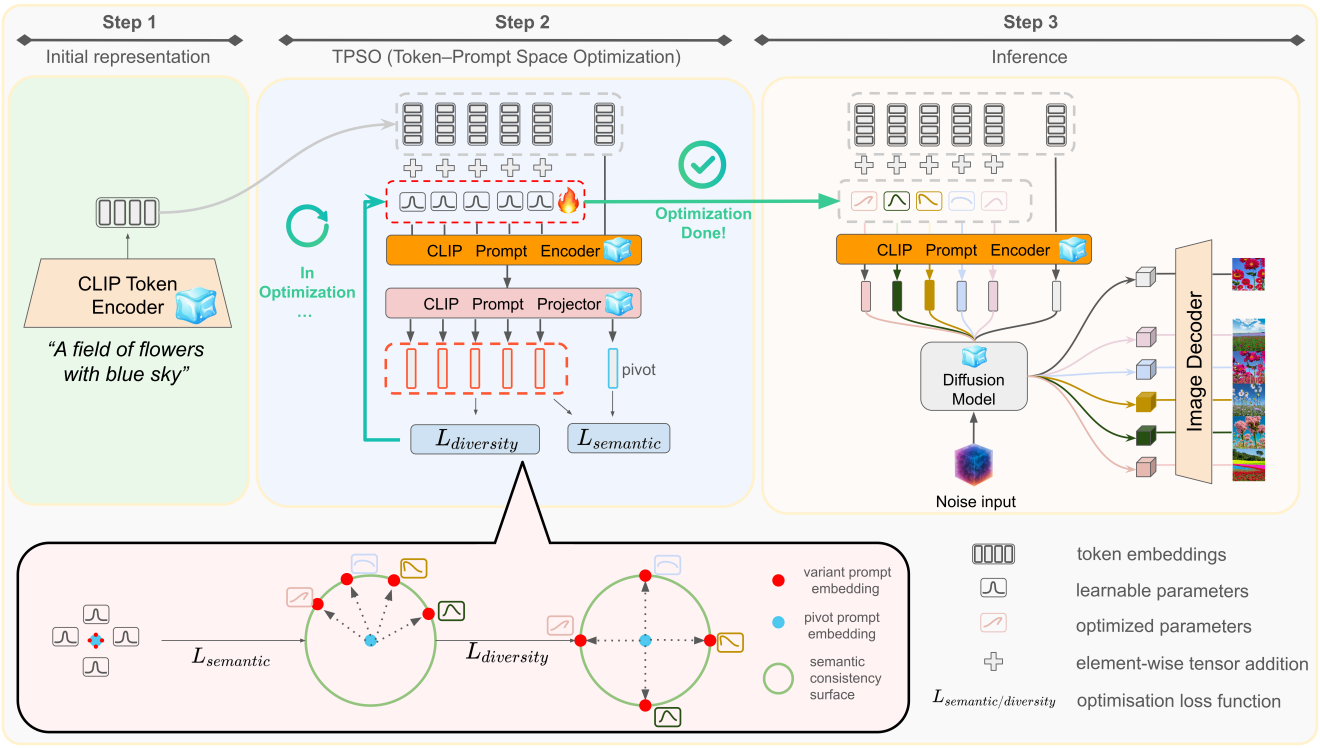

Fig. 2. Overview of the TPSO pipeline. TPSO optimizes lightweight learnable offsets in the token embedding space to produce multiple semantically aligned prompt variants, which induce diverse conditional representations and yield different CFG guidance directions. This enables substantially richer generative diversity, all without modifying the diffusion model or introducing any additional components.

while ToMe [29] merges prompt tokens and optimizes prompt embeddings to address attribute blindness in image generation. These works indicate that the token embedding space encodes rich semantic information and can be effectively manipulated to influence generative behavior. Inspired by these approaches, our work introduces token embedding space $\mathcal{S}_{\mathbf{e}} := \{e_i\}_{i=1}^{N}$ as a new direction for promoting diversity in image generation. Although CADS [6] also perturbs representations within the prompt encoder, it operates on the contextualized embeddings $\mathcal{S}_y := \{\mathbf{y}_i\}_{i=1}^{N}$. In contrast, our method applies optimization directly at the token embedding level $\mathcal{S}_{\mathbf{e}}$, which is more amenable to optimization by leveraging the CLIP text encoder E and the image–text aligned prompt space $\mathcal{S}_{\mathbf{v}} := \{\mathbf{v}_i\}_{i=1}^{N}$ during optimization. See Sections III-A and III-B for details.

III. Preliminary

A. CLIP Text Encoder

The CLIP text encoder [3] is a core component in modern text-to-image generation systems, including Stable Diffusion [4, 10] and DALL·E [9]. Given a text prompt c with tokens $\{i_1, \ldots, i_N\}$, each token is first mapped to an embedding via a deterministic lookup, $\mathbf{e}_n = M[i_n]$, where $M \in \mathbb{R}^{V \times d}$ is the token embedding matrix. We refer to this stage as the *token encoder*. The token embeddings are then contextualized by a Transformer-based *prompt encoder* E, yielding $\{y_1, \ldots, y_n\} = E(\mathbf{e}_1, \ldots, \mathbf{e}_n)$, which serve as the textual conditioning for the diffusion model. Finally, a lightweight *prompt projector* f maps the hidden state corresponding to the <EOS> token, denoted as y_{eos}, into a high-level image–text aligned semantic space, $\mathbf{v} = f(y_{\text{eos}})$, this semantic space serves as the target space for our optimization.

B. Quality Degradation in Diversity-Targeted Methods

Modern diffusion models often suffer from limited sample diversity [5]. To mitigate this issue, Particle Guidance [7] and CADS [6] introduce diversity-promoting perturbations in the latent space and conditioning space of diffusion, respectively. However, Particle Guidance operates primarily in the latent space at early diffusion steps, where the signal-to-noise ratio is low and semantic information is weak. Perturbations introduced at this stage are therefore prone to drifting toward low-semantic regions, resulting in degraded image fidelity (see Fig. 5). CADS promotes diversity by injecting random noise into the diffusion conditioning; although effective, such noise can easily corrupt the conditioning signal and degrade generation quality (Fig. 5). Consequently, both methods lack an explicit mechanism to preserve fidelity while encouraging diversity. To address this limitation, we introduce a fidelity-aware diversity module. Unlike CADS, which random perturbs the contextualized embeddings y_i, our TPSO optimizes a *fidelity- and diversity-aware offset* using semantic and diversity losses, which is added at the token embedding level $\mathbf{e}_i$ prior to contextualization (see $\{y_1, \ldots, y_n\} = E(\mathbf{e}_1, \ldots, \mathbf{e}_n)$, where $\mathbf{e}_i$ denotes the token embedding and y_i its contextualized rep-

Fig. 3. TPSO produces a wide range of visually distinct yet semantically faithful variants, demonstrating its ability to increase diversity without sacrificing image quality. In contrast, baseline diffusion models generate highly repetitive outputs across different sampling runs, revealing limited diversity.

resentation). This design allows gradients from the projected text representation $\mathbf{v} = f(y_{\text{eos}})$ to propagate through both the prompt encoder E and the projector f, enabling effective optimization. In contrast, perturbations applied directly to y_i in CADS do not affect y_{eos} and therefore cannot be optimized via objectives defined on $\mathbf{v}$.

IV. Method

Motivated by the effectiveness of token-level representations in influencing generative behavior (Section II), we leverage the token embedding space $\mathcal{S}_{\mathbf{e}}$ to promote diversity in image generation. Specifically, we optimize multiple learnable offsets on the deterministic token embeddings. Each offset is added to the original token embedding to produce a perturbed embedding, which is then processed by the CLIP prompt encoder to form a variant. These variants are fed into the generative model, resulting in diverse outputs under the same condition. To ensure that these variants maintain both diversity and fidelity, the offsets are jointly optimized using a diversity loss and a semantic alignment loss. Details of the token and prompt representations are given in Section IV-A, followed by the semantic regularization (Section IV-B), the diversity objective (Section IV-C), and the overall optimization and scheduling strategy across diffusion timesteps (Sections IV-D and IV-E).

A. Variant Token and Prompt Embedding

As illustrated in Figure 2, our method operates on the text encoder of Stable Diffusion. Let $\{i_1, \ldots, i_n\}$ denote the input prompt tokens, and let

$$\boldsymbol{e}_{\text{token}} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}, \qquad \mathbf{e}_i \in \mathbb{R}^d, \tag{1}$$

be their corresponding fixed CLIP token embeddings obtained via deterministic matrix indexing $\mathbf{e}_n = M[i_n]$, where $M \in \mathbb{R}^{V \times d}$ (Section III-A).

Algorithm 1 Optimization of Learnable Token Offsets

Input: Token embeddings $\boldsymbol{e}_{\text{token}}$, learnable offsets $\boldsymbol{\varepsilon}$
Output: Optimized learnable offsets $\boldsymbol{\varepsilon}$
Initialize $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 10^{-4})$
Compute $\mathbf{v} = f(E(\boldsymbol{e}_{\text{token}}))$
Compute $\mathbf{v}' = f(E(\boldsymbol{e}_{\text{token}} + \boldsymbol{\varepsilon}))$ (Eq. 3 and 4)
repeat
  Update $\boldsymbol{\varepsilon} \leftarrow \text{Adam}(\boldsymbol{\varepsilon}, \nabla_{\boldsymbol{\varepsilon}} \mathcal{L}_{\text{joint}})$ (Eq. 7)
  Recompute $\mathbf{v}' = f(E(\boldsymbol{e}_{\text{token}} + \boldsymbol{\varepsilon}))$
until $\mathcal{L}_{\text{semantic}}$ convergence (Eq. 8)
return $\boldsymbol{\varepsilon}$

To enable controllable variation, we introduce a set of learnable token-level offset parameters

$$\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\}, \qquad \varepsilon_i \in \mathbb{R}^d. \tag{2}$$

and define the resulting variant token embeddings as

$$\boldsymbol{e}_{\text{token}} + \boldsymbol{\varepsilon} = \{\mathbf{e}_1 + \varepsilon_1,\ \mathbf{e}_2 + \varepsilon_2,\ \ldots,\ \mathbf{e}_n + \varepsilon_n\}. \tag{3}$$

These variant embeddings are then passed through the CLIP prompt encoder E and projector network f to obtain prompt-level embeddings:

$$\mathbf{v} = f(E(\boldsymbol{e}_{\text{token}})), \qquad \mathbf{v}' = f(E(\boldsymbol{e}_{\text{token}} + \boldsymbol{\varepsilon})), \tag{4}$$

where both $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^d$ represent the original and offset-induced prompt embeddings, respectively. These embeddings constitute the basis for the semantic alignment and diversity objective.

B. Semantic Optimization Target

To ensure that the generated variants preserve the semantic meaning of the original prompt, we define a semantic align-

TABLE I
MAIN COMPARISON RESULTS BASED ON 50K SAMPLES. PRIMARY EVALUATION METRICS ARE HIGHLIGHTED IN **BOLD**.

	Quality			Consistency	**Diversity**		
Method	**FID** ↓	Precision ↑	PickScore ↑	CLIP ↑	**Recall** ↑	**MSS** ↓	**Vendi** ↑
Stable Diffusion 1.5 [4]	20.093	**0.610**	**21.48**	**31.700**	0.145	0.231	5.558
Dynamic CFG [6]	20.405	0.606	21.04	31.170	0.090	0.359	2.560
CADS [6]	19.750	0.602	21.09	30.940	0.313	0.194	5.933
Particle Guidance [7]	19.990	0.572	21.34	31.320	0.218	0.194	6.387
TPSO	**18.191** (-1.902)	0.608 (-0.002)	21.18 (-0.3)	31.100 (-0.6)	**0.374** (+0.229)	**0.158** (-0.073)	**6.705** (+1.147)
Stable Diffusion 2.1 [4]	19.871	0.599	**21.74**	**31.880**	0.117	0.247	5.237
Dynamic CFG [6]	24.823	**0.601**	21.12	30.960	0.033	0.355	2.608
CADS [6]	19.891	0.599	21.22	31.370	0.176	0.221	5.524
Particle Guidance [7]	19.757	0.595	21.73	31.840	0.125	0.243	5.298
TPSO	**17.569** (-2.302)	0.571 (-0.028)	21.34 (-0.4)	31.100 (-0.78)	**0.430** (+0.313)	**0.137** (-0.11)	**7.299** (+2.062)
Stable Diffusion 3.5 [30]	24.567	**0.645**	21.87	32.070	0.006	0.369	2.828
Dynamic CFG [6]	23.943	0.574	21.48	31.310	0.037	0.350	2.739
CADS [6]	26.934	0.639	22.17	31.760	0.066	0.277	4.359
Particle Guidance [7]	24.275	0.638	22.06	**32.190**	0.006	0.385	2.562
TPSO	**23.623** (-0.944)	0.576 (-0.069)	22.10(-0.77)	31.930 (-0.14)	**0.287** (+0.281)	**0.156** (-0.213)	**7.010** (+4.182)

ment loss. Since the CLIP prompt embedding space is explicitly structured by cosine similarity (e.g., CLIP Score [3]), we measure semantic deviation using the cosine similarity between the original prompt embedding $\mathbf{v}$ and each offset-induced embedding $\mathbf{v}'_i$:

$$\mathcal{L}_{\text{semantic}} = \sum_{i=1}^{N} \big|\cos(\mathbf{v}, \mathbf{v}'_i) - \kappa\big|. \tag{5}$$

We use a summation rather than an average since each variant has independent learnable offsets, and the gradient from $\mathcal{L}_{\text{semantic}}$ is not shared across variants.

C. Diversity Optimization Target

To encourage the generated image variants to be diverse rather than repeat, we introduce a diversity loss applied to the set of variant prompt embeddings $\mathbf{v}' = \{\mathbf{v}'_1, \ldots, \mathbf{v}'_N\}$. This loss penalizes high similarity between any pair of variants:

$$\mathcal{L}_{\text{div}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \cos(\mathbf{v}'_i, \mathbf{v}'_j). \tag{6}$$

D. Convergence Condition

The overall objective combines semantic preservation and diversity enhancement:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{semantic}} + \lambda \cdot \mathcal{L}_{\text{div}}, \tag{7}$$

where λ controls the strength of the diversity term. The complete optimization procedure is summarized in Algorithm 1.

For each learnable offset, the optimization terminates when either the convergence criterion in Eq 8 is satisfied with tolerance $\sigma = 10^{-2}$, or a maximum of 50 iterations is reached:

$$\big|\cos(\mathbf{v}, \mathbf{v}'_i) - \kappa\big| < \sigma. \tag{8}$$

TABLE II
IMPACT OF SEMANTIC RETENTION DEGREE κ

κ	FID ↓	Precision ↑	CLIP ↑	Recall ↑	MSS ↓	Vendi ↑
0.70	49.69	0.59	30.61	**0.48**	**0.12**	**7.56**
0.80	**49.58**	0.60	31.14	0.37	0.16	6.65
0.90	52.30	**0.62**	**31.56**	0.22	0.23	5.21

E. Progressive Embedding Scheduler.

Diffusion sampling follows a coarse-to-fine process [31]: early steps determine the global structure and benefit most from optimized embeddings to promote diversity, whereas later steps refine texture and appearance and are best guided by the original embedding to preserve visual quality [6]. We exploit this property using a simple ratio-based scheduler. Diffusion sampling proceeds from timestep T (e.g., 1000) down to 1. Given a proportion parameter $r \in [0, 1]$, the early timesteps (from $t = T$ down to $T(1 - r)$) use the optimized embedding with a progressive interpolation back to the original embedding, while the remaining timesteps rely solely on the original embedding. Let y and y' denote the original and optimized embeddings, respectively:

$$y_t^* = \alpha_t \, y' + (1 - \alpha_t) \, y, \tag{9}$$

where

$$\alpha_t = \begin{cases} 0, & t \leq T(1-r), \\ \dfrac{t - T(1-r)}{rT}, & T(1-r) < t \leq T. \end{cases} \tag{10}$$

We find that setting $r = 0.4$ provides high diversity while preserving image quality.

Fig. 4. Effect of the diversity loss on generation diversity.

TABLE III
COMPARISON OF COARSE-TO-FINE AND FINE-TO-COARSE SCHEDULING FOR OPTIMIZED EMBEDDINGS.

Scheduling Strategy	Setting	FID ↓	CLIP ↑	Recall ↑
Coarse-to-Fine ($r > 0$)	$r = 0.4$	**49.58**	31.14	0.374
	$r = 0.6$	50.06	30.96	**0.433**
Fine-to-Coarse ($r < 0$)	$r = -0.4$	63.53	**31.63**	0.165
	$r = -0.6$	61.55	31.59	0.169

V. EXPERIMENTS

A. Experimental Settings

Implementation Details. To verify that TPSO is model-agnostic, we evaluate it on three representative text-to-image generators: (1) Stable Diffusion 1.5 (SD1.5) and Stable Diffusion 2.1 (SD2.1), both based on the denoising diffusion framework [4, 1, 2]; and (2) Stable Diffusion 3.5 (SD3.5) [30], which adopts a rectified-flow formulation [32, 33]. We follow the default public configurations for all models. Notably, SD3.5 [30] employs two CLIP text encoders, which we optimize independently. All experiments are conducted with fixed random seeds to ensure fair and model-consistent comparisons on an A100 GPU using PyTorch 2.6.

Dataset. Following the experimental protocol of CADS [6], we conduct all experiments on the MSCOCO validation set [34]. We randomly sample 5,000 captions and generate 10 images per caption, resulting in 50k samples for each model–method pair. For ablation studies, we use a reduced subset of 1,000 captions (10k samples).

B. Evaluation Metrics

We follow CADS [6] and evaluate our method across three complementary aspects: **quality**, **conditional alignment**, and **diversity**.

Diversity Metrics. We report Recall [35] as a set-level diversity metric between synthetic and real images. Sample-level diversity is measured using the Mean Similarity Score (MSS) and the Vendi Score [36], which assess diversity among samples generated under the same condition. Both metrics are computed with SSCD [37] features following CADS [6].

TABLE IV
EFFECT OF DIFFERENT INITIALIZATION DISTRIBUTIONS FOR THE LEARNABLE PARAMETER ε (LINE 3 OF ALGORITHM 1).

	Quality		Consistency	Diversity		
Initialization	FID ↓	Precision ↑	CLIP ↑	Recall ↑	MSS ↓	Vendi ↑
Normal(0, 1e-4)	**50.69**	0.560	**30.06**	0.508	0.121	7.76
Normal(0, 1)	52.82	0.535	29.02	0.552	0.117	7.87
Laplace(0, 1)	59.97	0.494	27.84	0.551	0.097	8.30
Laplace(0, $\sqrt{1.5}$)	67.22	0.459	27.02	**0.554**	**0.087**	**8.52**
Gamma(1,1)	54.73	0.539	28.69	0.532	0.109	8.05
Gamma(4,$\sqrt{4}$)	53.83	0.543	28.86	0.533	0.111	8.00
Uniform(-1,1)	51.10	**0.563**	29.83	0.51	0.125	7.70
Uniform(-3,3)	63.98	0.466	27.44	0.547	0.092	8.39

TABLE V
IMPACT OF THE DIVERSITY-LOSS WEIGHT λ.

λ	FID ↓	Precision ↑	CLIP ↑	Recall ↑	MSS ↓	Vendi ↑
0.0	49.58	0.60	**31.14**	0.37	0.16	6.65
5.0	**49.38**	**0.62**	31.04	0.43	0.14	7.13
10.0	48.54	**0.62**	30.89	**0.47**	**0.12**	**7.54**

Quality Metrics. Fréchet Inception Distance (FID) [38] serves as our primary quality metric. We also report Precision [35] as a complementary measure of sample fidelity. However, as noted in CADS, Precision does not reliably evaluate models that produce highly diverse outputs.

Alignment Metric. Text–image consistency is evaluated using the CLIP Score [3], defined as the cosine similarity between CLIP embeddings of the generated image and its corresponding prompt.

C. Qualitative Results.

Comparison with Baseline Methods. Our objective is to increase image diversity while preserving visual fidelity. As shown in Fig. 3, baseline methods tend to produce visually similar results, often repeating comparable layouts or structures even when different noise seeds are used. In contrast, our method generates a broader range of plausible variations, exhibiting differences in geometry, color, and fine-grained details while remaining faithful to the input prompt.

Importantly, TPSO achieves this diversity while maintaining image fidelity, unlike Dynamic CFG, CADS, and Particle Guidance, which often improve diversity at the expense of text alignment or visual quality.

Quality Preservation As shown in Fig. 5, CADS [6] and Particle Guidance [7] introduce greater structural distortions or semantic drift than TPSO when increasing diversity, highlighting the challenge of improving diversity without degrading image fidelity.

D. Quantitative Results

Table I summarizes the quantitative performance across all baselines and diffusion backbones. TPSO consistently improves all diversity metrics (Recall, MSS, and Vendi) while also achieving better FID, our primary quality indicator. These results demonstrate that TPSO effectively enhances diversity without sacrificing overall image fidelity.

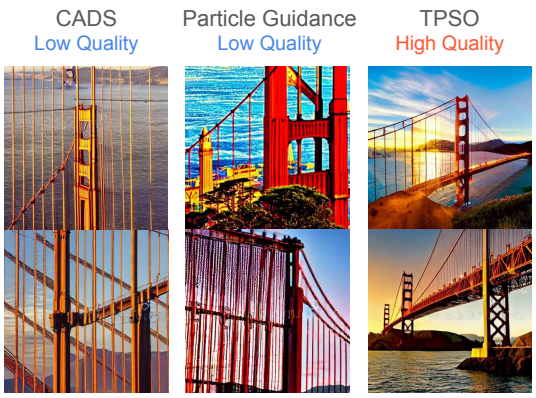

Fig. 5. Quality preservation under increased diversity. TPSO maintains higher visual quality and semantic alignment than CADS and Particle Guidance.

TABLE VI
ABLATION STUDY OF $\mathcal{L}_{\text{SEMANTIC}}$ AND $\mathcal{L}_{\text{DIVERSITY}}$. THE FIRST ROW ADDS A LEARNABLE OFFSET $\epsilon \sim \mathcal{N}(0, 10^{-4})$ WITHOUT OPTIMIZATION.

Optimization	$\mathcal{L}_{\text{semantic}}$	$\mathcal{L}_{\text{diversity}}$	FID ↓	CLIP ↑	Recall ↑	Vendi ↑
✗	✗	✗	58.11	**31.61**	0.03	1.98
✓	✗	✓	55.42	28.47	**0.62**	**9.21**
✓	✓	✗	49.58	31.14	0.37	6.65
✓	✓	✓	**49.38**	31.04	0.43	7.13

Although Precision exhibits a slight decrease, this reflects a known limitation of the metric, as it can favor high-quality yet less diverse samples [6]. This should not be interpreted as a degradation in image quality.

TPSO shows minor reductions in CLIP Score. This is because our method introduces learnable offsets in the prompt embedding space, which may induce small but controlled semantic shifts. However, these changes are marginal (less than 0.78 in absolute difference) and negligible compared to the substantial gains in diversity and FID.

E. Ablation Study

Semantic Retention Degree κ. κ controls the semantic alignment between variant and original prompt embeddings. Lower κ permits larger token-level deviations, resulting in reduced fidelity but increased diversity (Table II). We find $\kappa = 0.80$ provides a strong balance between diversity and semantic fidelity.

Effect of Diversity Loss and Weight λ. As shown in Table V, larger diversity-loss weights λ induce stronger intra-prompt variation, leading to more diverse generations.

Effect of $\mathcal{L}_{\text{semantic}}$ and $\mathcal{L}_{\text{diversity}}$. Tab. VI shows that using $\mathcal{L}_{\text{diversity}}$ alone increases diversity at the cost of fidelity, while using $\mathcal{L}_{\text{semantic}}$ alone yields limited diversity. Joint optimization achieves the best diversity–fidelity trade-off.

Effect of the Coarse-to-Fine Scheduler. The ratio r controls the portion of the sampling trajectory where optimized embeddings are applied. Positive values (e.g., $r = 0.4$) apply optimized embeddings in early denoising stages, whereas negative values (e.g., $r = -0.4$) apply them in later refinement stages, serving as a control for the coarse-to-fine design.

As shown in Table III, applying variations at early stages ($r > 0$) significantly improves diversity while preserving visual quality, whereas late-stage variations ($r < 0$) yield limited diversity gains and noticeably degrade image quality. These results support the effectiveness of the proposed coarse-to-fine scheduling strategy.

Parameter Initialization. Initialization affects optimization stability and convergence [39]. We compare different initialization strategies in Tab. IV, showing that $\mathcal{N}(0, 10^{-4})$ yields the best performance.

Inference Efficiency. Our method incurs only a small inference-time overhead of 3.6%–8.9% across Stable Diffusion variants (Table VII).

TABLE VII
INFERENCE TIME ACROSS STABLE DIFFUSION VERSIONS.

Method	Time (sec / sample) ↓ SDv1.5	SDv2.1	SDv3.5
Stable Diffusion (SD) [4, 30]	**0.78**	**1.80**	**7.46**
Dynamic CFG [6]	0.79	1.80	7.46
CADS [6]	0.80	1.80	7.47
Particle Guidance [7]	0.79	1.80	7.46
TPSO	0.85 (+8.9%)	1.93 (+7.2%)	7.73 (+3.6%)

VI. LIMITATIONS AND FUTURE WORK

Limitations. TPSO remains bounded by the representational capacity of the underlying diffusion model, as it does not modify model parameters. It relies on CLIP embeddings for semantic alignment and is therefore limited by CLIP's capacity. Large perturbations may still cause mild quality degradation in rare cases. TPSO also introduces a small inference-time overhead due to per-prompt optimization.

Future Work. Although TPSO operates at inference time, its effectiveness suggests potential extensions to training. Incorporating embedding-space exploration or diversity-aware regularization during training may reduce mode collapse and improve coverage, representing a promising direction for future research.

VII. CONCLUSION

We present TPSO, a training-free and model-agnostic module for enhancing diversity in text-to-image generation without sacrificing fidelity. By optimizing learnable offsets at the token level and regulating semantic deviation in the prompt space, TPSO enables controlled exploration of underrepresented regions of the token embedding space, reducing repeated sampling from strong modes of the distribution. Diversity is further enhanced through an explicit diversity loss. A coarse-to-fine scheduling strategy introduces diversity at early denoising stages while preserving high-quality refinement in later steps. Extensive experiments on SD1.5, SD2.1, and SD3.5 demonstrate consistent improvements in per-prompt and global diversity metrics, while preserving image fidelity with minimal

inference-time overhead. Given its simplicity, generality, and compatibility with modern diffusion and rectified-flow models, TPSO provides a practical solution for diversity-enhanced image generation.

Acknowledgements DM's and GT's work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant No. 10099264) and by the European Union (under EC Horizon Europe grant agreement No. 101135800 (RAIDO)).

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] P. Astolfi, M. Careil, M. Hall, O. Mañas, M. Muckley, J. Verbeek, A. R. Soriano, and M. Drozdzal, "Consistency-diversity-realism pareto fronts of conditional image generative models," *arXiv preprint arXiv:2406.10429*, 2024.

[6] S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R. M. Weber, "Cads: Unleashing the diversity of diffusion models through condition-annealed sampling," *arXiv preprint arXiv:2310.17347*, 2023.

[7] G. Corso, Y. Xu, V. De Bortoli, R. Barzilay, and T. Jaakkola, "Particle guidance: non-iid diverse sampling with diffusion models," *arXiv preprint arXiv:2310.13102*, 2023.

[8] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[10] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[11] Y. Su, H. Zhu, Y. Tian, C. Zhao, Z. Peng, Y. Liu, L. Fan, B. Li, and L. Zhang, "Agentic-sql taxonomy: A survey of autonomous and interactive text-to-sql with llms," Mar. 2026. [Online]. Available: http://dx.doi.org/10.22541/au.177430005.57777158/v1

[12] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[13] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.

[14] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[15] Z. Guo, K. Zhao, and L. Zhang, "Instancersr: Real-world super-resolution via instance-aware representation alignment," 2026. [Online]. Available: https://arxiv.org/abs/2603.24240

[16] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 490–22 499.

[18] Z. Gao, D. Meng, Y. Miao, Z. Zhang, S. Xu, I. Patras, and J. Song, "Diffusion-based makeup transfer with facial region-aware makeup features," *arXiv preprint arXiv:2603.20012*, 2026.

[19] D. Meng, C. Tzelepis, I. Patras, and G. Tzimiropoulos, "Mm2latent: Text-to-facial image generation and editing in gans with multimodal assistance," in *European Conference on Computer Vision*. Springer, 2024, pp. 88–106.

[20] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[21] B. Wu, P. Liu, Y. Hu, L. Zhang, A. Hu, and Z. Xu, "Indexnet: Timestamp and variable-aware modeling for time series forecasting," *arXiv preprint arXiv:2509.23813*, 2025.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[23] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[24] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[25] B. Trang, P. Saremi, A. Q. Wang, F. Huang, Z. TehraniNasab, A. Kumar, T. Arbel, L. Fei-Fei, and E. Adeli, "Discovering latent graphs with gflownets for diverse conditional image generation," *arXiv preprint arXiv:2510.22107*, 2025.

[26] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. [Online]. Available: https://arxiv.org/abs/2208.01618

[27] C. Jin, R. Tanno, A. Saseendran, T. Diethe, and P. A. Teare, "An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning," in *Forty-first International Conference on Machine Learning*, 2024.

[28] S. Um and J. C. Ye, "Minority-focused text-to-image generation via prompt optimization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 926–20 936.

[29] T. Hu, L. Li, J. van de Weijer, H. Gao, F. Shahbaz Khan, J. Yang, M.-M. Cheng, K. Wang, and Y. Wang, "Token merging for training-free semantic binding in text-to-image synthesis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 137 646–137 672, 2024.

[30] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[31] D. Decatur, T. Groueix, W. Yifan, R. Hanocka, V. Kim, and M. Gadelha, "Reusing computation in text-to-image diffusion for efficient generation of image sets," *arXiv preprint arXiv:2508.21032*, 2025.

[32] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[33] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[35] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[36] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *arXiv preprint arXiv:2210.02410*, 2022.

[37] E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, "A self-supervised descriptor for image copy detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 532–14 542.

[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:326772

[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.