Breaking the Static: Dynamic Text Conditioning for Diverse Image Generation

Meng Yu[1,2,3*] Ruidong Chen[3*] Qingfeng Shi[3] Yingmao Miao[3]
Yancheng Bai[3] Lei Sun[3] Xiangxiang Chu[3] Hesheng Wang[1†] Xiaodan Liang[2,4†]

[1]Shanghai Jiao Tong University [2]Zhongguancun Academy
[3]AMAP Alibaba Group [4]Shenzhen Campus of Sun Yat-sen University

Abstract

High-performance text-to-image (T2I) diffusion models suffer from diversity degradation. In this study, we attribute this challenge to a fundamental temporal mismatch: while visual features evolve dynamically in a coarse-to-fine manner in the denoising process, the text condition remains entirely static. Through empirical analysis, we demonstrate that decoupling the text prompt to prioritize low-frequency components in the early denoising stage better aligns with the visual denoising process, while simultaneously enhancing generation diversity. Driven by this insight, we introduce Dynamic Semantic Interpolation (DSI), a training-free strategy that utilizes low-frequency semantics during the initial stages to foster the sampling space and progressively recovers the full semantics, ensuring text-to-image alignment. We also provide a rigorous theoretical justification for DSI from the perspective of conditional entropy, explaining its ability to maintain a broader early sampling space. Extensive experiments across diverse state-of-the-art models demonstrate that DSI significantly boosts generation diversity while preserving text alignment and aesthetic quality.

1 Introduction

Text-to-image (T2I) diffusion models [43, 24, 36, 29, 64] have achieved unprecedented visual fidelity and precise text alignment. However, driven by the memorization of high-probability patterns [42, 63, 45] and reward hacking [55, 14], which stem from enhanced textual alignment [15, 5] and post-training techniques [30, 65, 32], models tend to repeatedly generate similar results. Consequently, restoring the inherent diversity of generated samples while maintaining strict semantic consistency has emerged as a critical challenge in modern T2I models.

In this work, we argue that this diversity deficiency stems from a temporal mismatch in denoising: while visual features evolve dynamically in a coarse-to-fine manner, the textual condition remains entirely static. Considering the dynamics of diffusion models [1, 41], the structural modes are decisively established with an early critical exploration window [13]. As shown in Fig. 1(a), we hypothesize that a full, static prompt imposes undeserved semantic constraints by prematurely injecting high-frequency textual details, which restricts the exploration of diverse structural initializations. Thus, our key insight emerges: **a dynamic text conditioning method that aligns with the low-frequency-structured visual features in the initial stage, thereby reducing unnecessary semantic constraints to enlarge the sampling space.**

To substantiate our hypothesis, we conduct an empirical analysis (in Sec. 4) of prompt semantic granularity. We decouple low-frequency semantic components from the prompt, which capture the subject identity and coarse structural layout, and apply them exclusively in the early denoising stage,

*Equal contribution. †Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

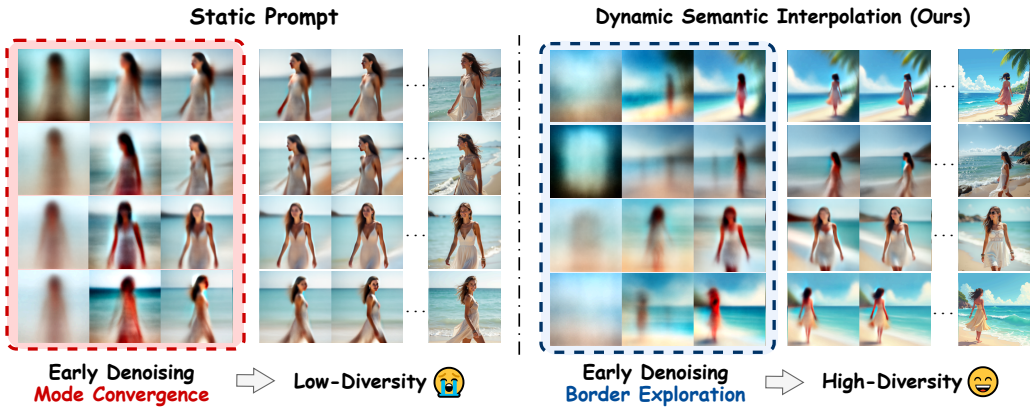


Figure 1: a) Diffusion dynamics [13, 1, 41] reveal a critical early window for structure determination, where static prompts prematurely inject high-frequency details and force visual modes to converge toward narrow, high-probability modes [42, 63, 45]. b) We propose a dynamic semantic interpolation strategy that aligns text conditioning with visual evolution by prioritizing low-frequency semantics during the initial stage, encouraging broader structural exploration to enhance generation diversity.

while retaining the full prompt for the remaining stages. This yields a key observation: compared to static prompt conditioning, injecting only low-frequency semantics in the early stage to better align with visual feature evolution significantly enhances sample diversity, as illustrated in Fig. 1(b).

Building on the observation, we propose a simple yet effective Dynamic Semantic Interpolation (DSI) framework, aimed at aligning visual evolution with dynamic text conditioning to enhance diversity. Our method centers on two core strategies: first, we automate the extraction of low-frequency semantic representations by quantitatively estimating token-wise importance through CLIP-based semantic similarity [39]; Second, we introduce a time-dependent interpolation strategy that prioritizes the core, low-frequency semantics during the early denoising window [13] to encourage structural exploration (as shown in Fig. 2), while smoothly re-injecting high-frequency details in later stages to ensure complete textual alignment. We also provide a rigorous theoretical justification from the perspective of conditional entropy, demonstrating that our strategy inherently maintains a broader early sampling space. We summarize our contribution as follows:

- We reveal the diversity deficiency from a novel perspective (temporal mismatch between dynamic visual features and static semantics), identifying that reducing early-stage semantics constraints effectively enhances diversity.
- We propose the Dynamic Semantic Interpolation (DSI) strategy to achieve dynamically evolving text conditioning, which aligns with diffusion dynamics to enhance diversity while ensuring aesthetic quality and textual alignment.
- We verify the effectiveness of the proposed method through extensive experiments, including on general T2I models [56, 24, 10], Reinforce Learning (RL) models [30, 65], and step-distillation models [27, 48]. Specifically, as an orthogonal perspective, DSI can also effectively integrate with existing diversity-enhancement methods [23, 33, 19]. These experiments confirm the effectiveness, universality, and practical value of our method.

2 Related Work

2.1 Text-to-Image Diffusion Models

T2I models [2, 43, 24] have advanced rapidly, with diffusion models [16, 9, 47] emerging as a dominant generative paradigm. Flow Matching [29] provides a simulation-free framework for learning continuous generative flows. Meanwhile, Diffusion Transformers (DiT) [34] replace U-Net [44] backbones with transformers [54], offering improved architectural scalability. Recent advances in training methods [60, 7, 25] and sampling strategies [61, 38, 62] have significantly improved the generation quality of diffusion models. Beyond visual quality, distillation techniques [31, 59] have

been introduced into T2I diffusion models to accelerate the generation process, while reinforcement learning paradigms [30, 14, 55] are incorporated to further enhance the model's text alignment and human preference alignment capabilities.

However, higher generation quality often comes at the cost of diversity. Pre-training and distillation techniques can strengthen the memorization of high-probability samples [42], while post-training alignment is prone to reward hacking [55, 14], causing models to repeatedly favor similar high-preference images. As a result, these methods may improve image quality but also push generation toward narrow, high-probability modes [1, 13], reducing diversity and practical utility.

2.2 Diverse Image Generation

In recent years, substantial research has investigated the limited diversity of T2I diffusion models. These core methodologies generally include: Condition intervention [45], perturb conditional embeddings with noise and gradually anneal it across timesteps. Gradient-guided strategies [53, 51] enhance diversity by leveraging a small number of scores. Latent intervention [8, 20], injects repulsive forces into latent space to force the separation of generation trajectories. However, both interventions suffer from uncontrollable perturbations, often leading to degradation in image quality. Group-based selection method [33] attempts to filter a diverse and high-quality subset from a vast pool of candidate samples, but incurs substantial computational overhead during inference. Dynamic classifier-free guidance (CFG) methods [23, 19] vary guidance scales across timesteps to improve diversity. Additionally, works such as MinorityPrompt [52], D^2-Align [4], and DAVE [22] have improved generation diversity through different strategies. However, global scaling cannot decouple early structural exploration from later detail constraints, resulting in a suboptimal trade-off.

Distinct from prior work, our method is rooted in the inherent coarse-to-fine dynamics of diffusion models. By interpolating semantics across different "semantic frequencies" throughout generation, early structural exploration is encouraged, improving diversity without sacrificing prompt alignment.

3 Preliminary

To introduce the proposed method, this section reviews the relevant preliminaries.

Denoising Trajectory. Mainstream diffusion models operate in latent space [43] for efficiency, progressively transforming Gaussian noise $\boldsymbol{z}_T$ into the clean latent $\boldsymbol{z}_0$ over timesteps $t \in [0, T]$. Taking the flow-based model [29, 24, 10] as an example, the generation process is governed by a network that predicts a velocity field $\boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c})$, where $\boldsymbol{c}$ is the textural condition. During inference, sampling uses an explicit Euler solver to integrate the ODE, yielding the denoising trajectory:

$$\boldsymbol{z}_{t-1} = \boldsymbol{z}_t \; + \; (\sigma_{t-1} - \sigma_y)\, \boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c}). \tag{1}$$

During this process, to obtain the estimated intermediate image, we can use Tweedie's Formula [16], which is a one-step Euler projection in the flow-based model, formulated as:

$$\boldsymbol{z}_0^t = \boldsymbol{z}_t - \sigma_t \cdot \boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c}). \tag{2}$$

Then, $\hat{\boldsymbol{z}}_0^t$ is mapped back to the pixel space via a pre-trained decoder to produce the image $\hat{\boldsymbol{x}}_0^t$.

Classifier-free Guidance [15] (CFG) is employed in mainstream models for enhancing textual alignment, by linearly combining conditional and unconditional (null-conditioned) predictions:

$$\hat{\boldsymbol{v}}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c}) = \boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \emptyset) + \eta \cdot \big(\boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c}) - \boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \emptyset)\big), \tag{3}$$

where $\emptyset$ represents the null condition and η is the guidance scale. Notably, our methods do not intervene in the model's internal generation process, but modulates semantic injection only through interpolation of the text condition $\boldsymbol{c}$. Thus, it is naturally compatible with various guidance strategies.

Visual Evolution in Diffusion Models. During the denoising process of diffusion models, visual features follow a coarse-to-fine evolution pattern [46, 58, 37]: early timesteps recover low-frequency structures such as layout and shape, while later steps add high-frequency details such as texture and attributes. Meanwhile, diffusion dynamics studies [13, 1, 41] suggest that the early trajectory contains a critical window in which structural modes and concepts become determined.

Motivated by this frequency-aligned dynamics, DSI matches text conditioning to the visual evolution by prioritizing coarse (low-frequency) semantics in the early window and progressively re-introducing

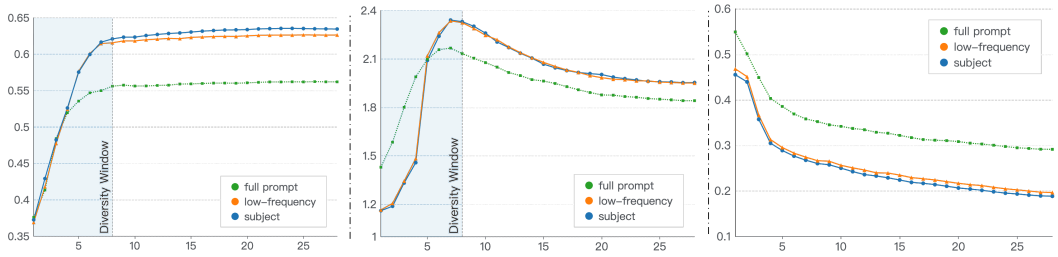


Figure 2: We track the diversity of estimated $\hat{\boldsymbol{x}}_0$ across denoising timesteps, using three metrics: (a) DINO Diversity ↑, (b) Vendi Score ↑, and (c) Mean Pairwise Similarity ↓. We find that generation diversity is largely determined within an early diversity window, where using low-frequency ($\boldsymbol{c}_{low}$) or subject-only ($\boldsymbol{c}_{sub}$) prompts consistently yields much higher diversity than the full static prompt $\boldsymbol{c}$.

the full prompt, establishing a correspondence between the "semantic frequency" decomposition and the low-to-high frequency evolution of denoising generation.

4 The Impact of Semantic Granularity

The mismatch between dynamic visual evolution and static text conditioning motivates our core hypothesis: injecting high-frequency semantics from the full static prompt too early imposes rigid semantic constraints that limit early structural exploration. To verify this, we empirically study how prompt granularity affects generation diversity across denoising timesteps.

We use a randomly sampled subset of MS-COCO captions [28] as our test prompts, denoted by $\boldsymbol{c}$. To separate semantic information, we use an LLM to extract two prompt types from $\boldsymbol{c}$: $\boldsymbol{c}_{low}$ for low-frequency semantics, including basic content, overall structure, and necessary modifiers, and $\boldsymbol{c}_{sub}$ for the main subject only, excluding adjectives and verbs. We adopted the diffusion dynamic analysis from [13], using $\boldsymbol{c}$, $\boldsymbol{c}_{low}$, and $\boldsymbol{c}_{sub}$ for generation in the first 20% of the steps, and then reverting to $\boldsymbol{c}$ in later timesteps to estimate the impact of $\boldsymbol{c}$, $\boldsymbol{c}_{low}$, and $\boldsymbol{c}_{sub}$ on diversity. We use three diversity metrics: DINO Diversity Score [3] to evaluate structural and semantic variance; Vendi Score [11] to measure distinct visual modes across samples; Mean Pairwise Similarity, to evaluate diversity based on image features. During denoising, we estimate the fully denoised image $\hat{\boldsymbol{x}}_0$ (obtained via Eq. 2) to track these metrics. Experiments are conducted on SD3.5-Medium [10], averaged over 100 cases with 4 samples per prompt. More details and the analysis for FLUX [2] are provided in Appendix A.

We identify a clear diversity hierarchy, as shown in Fig 2. Across all timesteps, global structural variance (DINO Diversity Score), the effective number of visual modes (Vendi Score), and visual feature similarity (Mean Pairwise Similarity) consistently follow $\boldsymbol{c}_{sub} \approx \boldsymbol{c}_{low} > \boldsymbol{c}$. This supports the claim that the static prompt ($\boldsymbol{c}$) introduces high-frequency constraints too early, limiting diversity. Contrastly, early guidance with coarse-grained semantics ($\boldsymbol{c}_{low}$ or $\boldsymbol{c}_{sub}$) promotes diversity.

We also observe clear temporal determinism driven by an early diversity window. Diversity rises sharply and is largely determined in the initial denoising phase (e.g., before step 8), after which the DINO Score flattens and the Vendi curves decline smoothly as the model shifts to high-frequency detail refinement. These findings motivate our core idea: **to improve diversity, the semantic granularity of text conditioning should be dynamically scheduled to match the model's evolving focus during this critical temporal window.**

5 Dynamic Semantic Interpolation

Motivated by the diversity patterns across semantic granularities, we propose **Dynamic Semantic Interpolation** (DSI) to better align text conditioning with visual evolution in diffusion models, as shown in Fig. 3. DSI can be expressed in a unified form:

$$\hat{\boldsymbol{e}}_j = (1 - \alpha_j)E(S_k(\boldsymbol{c})) + \alpha_j E(\boldsymbol{c}), \tag{4}$$

where $\hat{\boldsymbol{e}}_j$ denotes the text embedding at denoising step j, $S_k(\boldsymbol{c})$ selects k dominant words from the full prompt $\boldsymbol{c}$ based on CLIP [39], and $E(\cdot)$ denotes the diffusion model's text encoder. The coefficient α_j

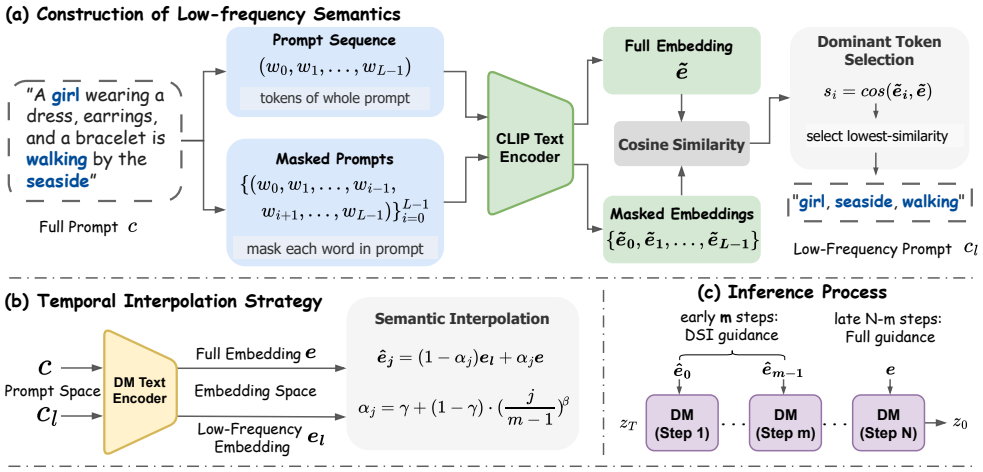


Figure 3: The overall framework of the proposed method. (a) **Construction of low-frequency semantics**, we apply token-level masking to the entire prompt and measure dominant tokens, which serve as the low-frequency prompt c_l for generation. (b) **Dynamic semantic interpolation** dynamically blends the low-frequency $\boldsymbol{e}_l$ and full $\boldsymbol{e}$ condition via an power-law schedule, which recovers the full condition progressively, ensuring the detailed textual alignment.

controls the transition from the selected semantics to the full-prompt condition. Semantic selection determines the starting point of this conditioning path, while temporal interpolation progressively restores full semantic guidance. We describe DSI in three parts: Construction of Low-Frequency Semantics (Sec. 5.1), Temporal Interpolation Strategy (Sec. 5.2), and Theoretical Analysis (Sec. 5.3).

5.1 Construction of Low-frequency Semantics

The "low-frequency" semantics of a textual condition can be interpreted as the dominant elements governing low-frequency visual content (e.g., primary subjects, global background). These semantics can serve as crucial early-stage guidance to initially establish the visual content while ensuring broader structural exploration. In this part, instead of relying on the LLM, we propose an automated, word-level selection for **C**onstructing **L**ow-frequency **S**emantics (CLS) to ensure practicality.

Specifically, we identify tokens that contribute most to the overall semantics as the key low-frequency components, utilizing a pre-trained CLIP text encoder to quantitatively measure the contribution. Let the full prompt $\boldsymbol{c}$ be a sequence of L words: $\boldsymbol{c} = (\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{L-1})$. For each position $i \in \{0, 1, \ldots, L-1\}$, we form a masked prompt $\boldsymbol{c}_i$ by remove $\boldsymbol{w}_i$ in the token sequence:

$$\boldsymbol{c}_i = (\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}, \boldsymbol{w}_{i+1}, \ldots, \boldsymbol{w}_{L-1}), \tag{5}$$

which yields the set $\mathcal{C} = \{\boldsymbol{c}, \boldsymbol{c}_0, \boldsymbol{c}_1, \cdots, \boldsymbol{c}_{L-1}\}$. Then, we extract the embeddings of $\mathcal{C}$ using the CLIP text encoder, denoted as $\{\tilde{\boldsymbol{e}}, \tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1, \ldots, \tilde{\boldsymbol{e}}_{L-1}\}$. The masked token contribution is then measured by the cosine similarity $\boldsymbol{s}_i$ between its masked embedding $\tilde{\boldsymbol{e}}_i$ and the full CLIP embedding $\tilde{\boldsymbol{e}}$:

$$\boldsymbol{s}_i = \cos(\tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{e}}) = \frac{\tilde{\boldsymbol{e}}_i \cdot \tilde{\boldsymbol{e}}}{\|\tilde{\boldsymbol{e}}_i\| \cdot \|\tilde{\boldsymbol{e}}\|}. \tag{6}$$

A lower similarity score indicates that the masked word has a larger effect on the overall semantics and better captures the dominant semantics of the full prompt $\boldsymbol{c}$. Given an integer $k \in \{0, \ldots, L\}$, we select the k words with the lowest similarity scores to construct the low-frequency semantic prompt $\boldsymbol{c}_l = S_k(\boldsymbol{c})$. When k is small, $\boldsymbol{c}_l$ reduces to the core subject; when $k = 0$, $\boldsymbol{c}_l$ becomes the null condition $\emptyset$, an extreme case of DSI discussed in Sec. 6.5.

5.2 Temporal Interpolation Strategy

Building upon the constructed $\boldsymbol{c}_l$, we propose the **T**emporal **I**nterpolation **S**trategy (**TIS**) to dynamically inject coarse-to-fine textual semantics during the early denoising, aligning with the intrinsic modes of the visual evolution process, which will significantly enhance diversity.

Formally, the text encoder $E(\cdot)$ maps the full prompt $\boldsymbol{c}$ and its low-frequency counterpart $\boldsymbol{c}_l$ to text embeddings $\boldsymbol{e} = E(\boldsymbol{c})$ and $\boldsymbol{e}_l = E(\boldsymbol{c}_l)$, respectively. During the first m denoising steps, the conditioning embedding $\hat{\boldsymbol{e}}_j$ at step j is computed via linear interpolation between $\boldsymbol{e}_l$ and $\boldsymbol{e}$:

$$\hat{\boldsymbol{e}}_j = (1 - \alpha_j)\, \boldsymbol{e}_l + \alpha_j \boldsymbol{e}, \tag{7}$$

where α_j is the interpolation coefficient at denoising timestep j, following a power-law schedule:

$$\alpha_j = \gamma + (1 - \gamma) \cdot \left(\frac{j}{m - 1} \right)^{\beta}. \tag{8}$$

Specifically, $\gamma \in [0, 1)$ is the starting coefficient, which sets the proportion of the full-prompt embedding $\boldsymbol{e}$ used at the first denoising step, and $\beta > 0$ is the scheduling exponent controlling the interpolation rate. As shown in Eq. (8), β determines the schedule: $\beta = 1$ gives a linear schedule; $\beta < 1$ increases the coefficients rapidly to introduce detailed semantics early; and $\beta > 1$ increases them slowly to initially prioritize low-frequency structural semantics from $\boldsymbol{e}_l$.

For denoising steps $j \geqslant m$, the full-prompt embedding $\boldsymbol{e}$ is applied directly without modification. Notably, we leverage the intrinsic property of special tokens (e.g., the [PAD] tokens in the T5 encoder [40] and the [EoT] token in the CLIP encoder [39]), which aggregate the global semantics of the entire prompt [50, 26, 6, 17]. When performing semantic interpolation between $\boldsymbol{e}_l$ and $\boldsymbol{e}$, these special tokens act as high-dimensional anchors that ensure the transition is smoothly transformed across the global semantics of both prompts. The interpolation scheme is applied to all text embeddings encoded by the T2I framework, including the pooled prompt embedding, ensuring a consistent transition of global semantic representations for all textual conditions. More discussion is in Appendix B.

5.3 Theoretical Analysis

To further validate DSI, we provide a theoretical analysis from the conditional entropy perspective.

We framework our analysis around the token-level decomposition established in Sec. 5.1. Specifically, let the full prompt $\mathbf{c}$ be defined as the union: $\boldsymbol{c} = \boldsymbol{c}_l \cup \boldsymbol{c}_o$, where $\mathbf{c}_o$ denotes all elements of $\boldsymbol{c}$ excluding those in $\boldsymbol{c}_l$. Subsequently, we let $\boldsymbol{x}$ denote the generated image and analyze how the textual conditioning limits the theoretical uncertainty of $\boldsymbol{x}$ from the perspective of conditional entropy.

Theorem 1. The conditional entropy of $\boldsymbol{x}$ given $\boldsymbol{c}$ satisfies:

$$H(\boldsymbol{x} \mid \boldsymbol{c}) = H(\boldsymbol{x} \mid \boldsymbol{c}_l) - I(\boldsymbol{x};\, \boldsymbol{c}_o \mid \boldsymbol{c}_l), \tag{9}$$

where $I(\boldsymbol{x};\, \boldsymbol{c}_o \mid \boldsymbol{c}_l)$ is the conditional mutual information:

$$I(\boldsymbol{x};\, \boldsymbol{c}_o \mid \boldsymbol{c}_l) = H(\boldsymbol{c}_o \mid \boldsymbol{c}_l) - H(\boldsymbol{c}_o \mid \boldsymbol{x}, \boldsymbol{c}_l). \tag{10}$$

Corollary 1. By the non-negativity of mutual information, $I(\boldsymbol{x};\, \boldsymbol{c}_o \mid \boldsymbol{c}_l) \geq 0$, which yields:

$$H(\boldsymbol{x} \mid \boldsymbol{c}) \leqslant H(\boldsymbol{x} \mid \boldsymbol{c}_l), \tag{11}$$

where equality holds if and only if $\boldsymbol{x} \perp \boldsymbol{c}_o \mid \boldsymbol{c}_l$, i.e., given $\boldsymbol{c}_l$, the generated image $\boldsymbol{x}$ is completely independent of $\boldsymbol{c}_o$. In practice, however, $\boldsymbol{x}$ is continuously guided by $\boldsymbol{c}_o$ throughout the denoising process, thus this condition is never satisfied.

Conclusion. The generated image $\boldsymbol{x}$ has higher conditional entropy under the low-frequency prompt $\mathbf{c}_l$ than under the full prompt $\mathbf{c}$, implying greater output uncertainty and structural diversity. By starting denoising with $\mathbf{c}_l$ and gradually transitioning to $\mathbf{c}$, DSI enables broader structural exploration before applying stricter semantic constraints from $\mathbf{c}_o$.

6 Experiments

To comprehensively evaluate the effectiveness, robustness, and generalizability of our methods, we detail the evaluated models, benchmark dataset, and evaluation metrics as follows.

Evaluated Models. Including general T2I foundation models (SD3-M, SD3.5-M [10] and FLUX-1.dev [2]), reinforcement learning (RL) post-trained models (Flow-GRPO [30] and Diffusion-NFT [65]) and efficient distillation models (SDXL-Light [27] and FLUX-Turbo [48]). As DSI

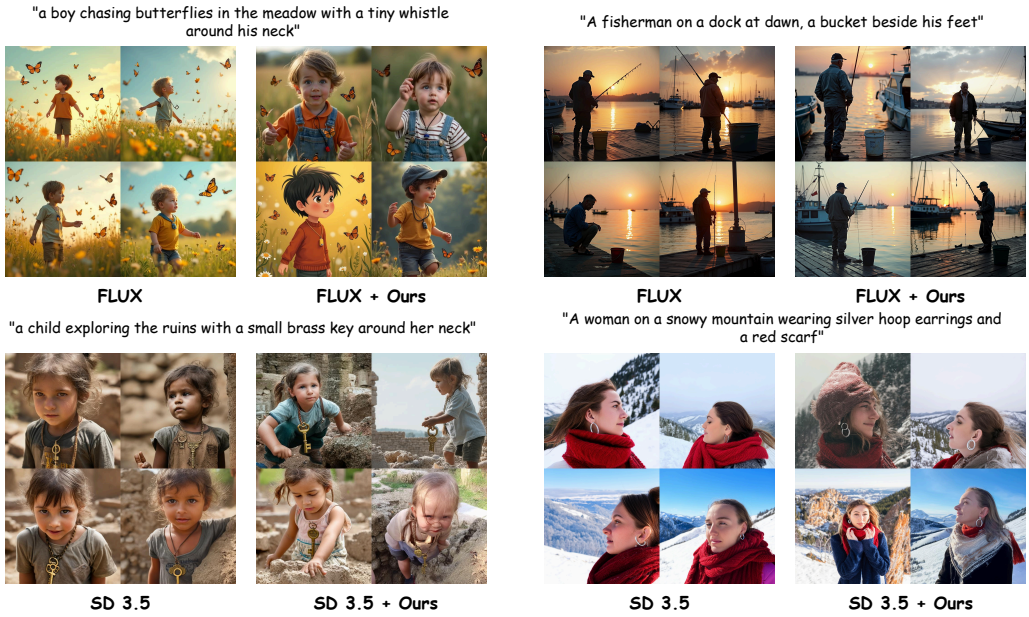


Figure 4: Visualizations of integrating our proposed DSI to various T2I models.

boosts diversity orthogonally to prior work, we validate its compatibility by integrating it with SOTA methods to enhance diversity: CFG-TV [19], Scale-Group [33], and CFG-In [23].

Dataset. All quantitative evaluations are conducted on 1000 randomly sampled prompts from the widely adopted MS-COCO captions [28], and we generate 4 images per prompt for evaluation. In addition, we introduce the more complex GenEval [12] and DPG-Bench [18] for evaluation.

Table 1: The quantitative evaluation on General T2I Models.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
SD3-M	**0.292**	**0.351**	**0.227**	**5.587**	0.574	1.806	0.255
+Ours	0.290	0.348	0.226	5.558	**0.637**	**1.916**	**0.160**
SD3.5-M	**0.280**	**0.353**	**0.225**	4.752	0.576	1.856	0.277
+Ours	0.277	**0.353**	**0.225**	**4.777**	**0.625**	**1.921**	**0.220**
FLUX.1-dev	**0.308**	**0.337**	**0.230**	5.925	0.563	1.804	0.351
+Ours	0.305	0.330	**0.230**	**5.995**	**0.603**	**1.901**	**0.295**

Evaluation Metrics. We evaluate the generated images across three dimensions: **Textual Alignment & Preference**: HPSv2 [57], CLIPScore [39], and PickScore are used to assess prompt adherence and human preference. **Image Aesthetics**: The Aesthetic Score is employed to quantify overall visual quality. **Diversity**: generation Diversity is measured via DINO Diversity [3] and Vendi Score [11]. Additionally, we measure Mean Pairwise Similarity (MSS), where lower values indicate reduced feature similarity and fewer mode collapses. We empirically define semantic preservation as changes in semantic alignment scores within 0.008. Diversity improvements are considered meaningful only when this constraint is satisfied. For further details, refer to Appendix C.

6.1 Qualitative Comparison

To intuitively demonstrate DSI's effect on generation diversity, identical random seeds and sampling steps are used for both baseline and improved models during inference, keeping denoising trajectories as consistent as possible. Fig. 4 visualizes DSI applied to FLUX and SD3.5. Baseline models

Table 2: The results on DPG-bench [18].

Method	DPGS	Entit	CLIPS	Aesth	Vendi	Dino
FLUX	**87.32**	92.80	0.349	5.16	1.74	0.53
+Ours	87.27	**92.96**	**0.350**	**5.18**	**1.76**	**0.55**

Table 3: The results on GenEval [12].

Method	Over	Color	HPSv2	PickS	Vendi	Dino
FLUX	**0.69**	0.80	**0.307**	0.24	1.87	0.55
+Ours	0.67	**0.89**	0.306	**0.24**	**1.92**	**0.56**

Table 4: The results on CFG-In, CFG-TV, and Scale-Group, which are diversity-enhancing methods.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
CFG-In [23]	**0.272**	**0.343**	**0.224**	**5.905**	0.641	2.114	0.142
+Ours	0.267	0.338	0.223	5.902	**0.662**	**2.180**	**0.118**
CFG-TV [19]	**0.292**	**0.351**	**0.228**	**5.958**	0.619	1.982	0.195
+Ours	0.286	0.346	0.226	5.943	**0.649**	**2.070**	**0.151**
Scale-Group [33]	**0.301**	**0.338**	**0.231**	5.994	0.629	1.866	0.274
+Ours	**0.301**	**0.338**	0.230	**6.121**	**0.651**	**1.931**	**0.245**

produce consistent, high-quality but low-diversity results with high inter-sample similarity. With DSI, significantly greater variation is achieved while preserving semantic alignment and aesthetics. Additional visualizations are provided in Appendix D.

Additionally, we conduct a human evaluation using 100 prompts, generating four samples per prompt for each method. Evaluators independently assign binary semantic-alignment scores: 1 if a sample matches the prompt and 0 otherwise. The baseline and our method achieve alignment rates of 97.00% and 95.75%, respectively. For diversity, evaluators select the more diverse sample group in pairwise comparisons. Our method is preferred in 78 out of 100 comparisons, compared with 22 for the baseline. The human evaluation shows that our method substantially improves generation diversity while maintaining semantic alignment close to the baseline.

6.2 Results of General T2I diffusion Models.

We validate DSI on three general-purpose T2I foundation models: SD3-M, SD3.5-M, and FLUX.1-dev. As shown in Tab. 1, DSI consistently improves diversity across all models, increasing DivDino and Vendi Score while reducing MSS, which indicates lower inter-sample feature similarity and effective mitigation of mode collapse. Meanwhile, text-alignment metrics, including HPSv2, CLIPScore, and PickScore, as well as aesthetic scores, remain largely unchanged. Notably, DSI even improves aesthetic scores on SD3.5-M and FLUX.1-dev. These results demonstrate that DSI is a plug-and-play module that enhances diversity across different architectures without sacrificing semantic consistency.

Then, we selected GenEval and DPG-Bench to test DSI on complex datasets. GenEval is an object-focused framework for evaluating compositional image properties (e.g., object co-occurrence, position, count, color). DPG-Bench contains highly detailed, dense, and compositionally complex long prompts. Tab. 3 and 2 show that DSI remains effective on complex benchmarks. On DPG-Bench, DSI outperforms the baseline across Entit, CLIPScore, Aesthetic, and DivDino. It also performs well in object binding, color attributes, and human preference, while significantly improving diversity.

6.3 Integration with Diversity-enhanced Methods

To examine whether DSI further improves diversity-oriented baselines, we apply it to CFG-TV, CFG-In, and Scale-Group. As shown in Tab. 4, DSI consistently improves diversity across these baselines while preserving text consistency and aesthetics. For example, on CFG-In, it increases DivDino from 0.641 to 0.662 and Vendi Score from 2.114 to 2.180, while reducing MSS from 0.142 to 0.118. These results show that DSI is orthogonal and complementary to existing diversity-enhancement methods. Additional comparisons on Scale-Group are provided in Appendix E.

Table 5: Quantitative Experiments on RL-Finetuned and Distilled Models.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
Flow-GRPO [30]	**0.310**	**0.354**	**0.236**	**5.539**	0.528	1.705	0.380
+Ours	0.309	**0.354**	0.235	5.531	**0.562**	**1.744**	**0.345**
SDXL-Light [27]	**0.290**	**0.346**	**0.228**	6.370	0.613	1.918	0.247
+Ours	0.285	0.338	0.227	**6.383**	**0.632**	**1.955**	**0.219**

Table 6: Analysis on core design of our method.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
SD3.5	0.283	0.354	0.226	4.820	0.562	1.842	0.291
$\hat{\boldsymbol{e}}_j = \boldsymbol{e}_l$ (w/o TIS)	0.270	0.348	0.223	4.813	0.641	2.010	0.166
$\boldsymbol{c}_l = \emptyset$ (w/o CLS)	0.272	0.349	0.224	4.783	0.669	2.017	0.159
DSI (Ours)	0.280	0.353	0.225	4.866	0.617	1.910	0.218

6.4 Results on RL-finetuned and Distilled Models.

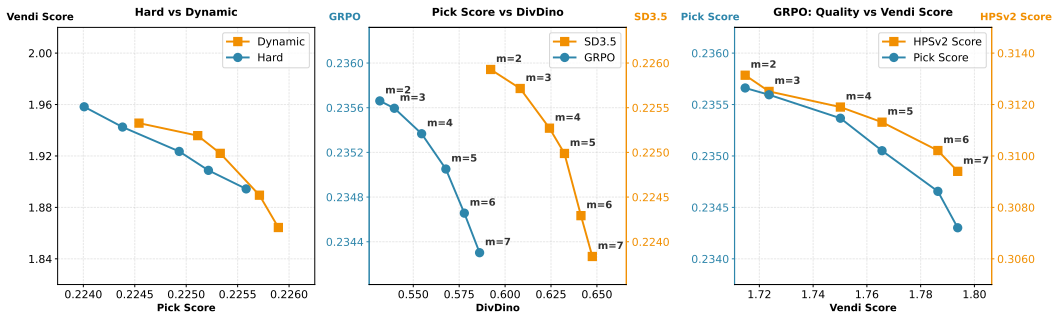
We apply DSI to RL-finetuned and distilled models, including Flow-GRPO (SD3.5 Medium-Pick), DiffusionNFT, SDXL-Light, and FLUX-Turbo, to verify its generality. As shown in Tab. 5, DSI improves diversity on Flow-GRPO, increasing DivDino from 0.528 to 0.562 and Vendi Score from 1.705 to 1.744. On SDXL-Light, DSI also enhances the diversity and aesthetics under 8-step sampling. These results show that DSI remains effective across both fine-tuned and distilled models. Results on DiffusionNFT and FLUX-Turbo are provided in Appendix E.

6.5 Module Analysis

Effectiveness of Core Designs. To examine the effectiveness and necessity of DSI's core components, we consider two extreme configurations derived from Sec.5.1, which can also be regarded as variants of DSI. Specifically, $\hat{\boldsymbol{e}}_j = \boldsymbol{e}_l$ uses $\boldsymbol{c}_l$ during the first m steps without Temporal Interpolation Strategy (TIS) in Sec. 5.2, while $\boldsymbol{c}_l = \emptyset$ replaces $\boldsymbol{c}_l$ with empty text, omitting Construction of Low-Frequency Semantics (CLS) in Sec. 5.1. Tab. 6 shows that both extreme variants improve diversity metrics but substantially compromise semantic alignment. For example, the HPSv2 scores of $\hat{\boldsymbol{e}}_j = \boldsymbol{e}_l$ and $\boldsymbol{c}_l = \emptyset$ decrease by 0.013 and 0.011, respectively, both exceeding the empirical semantic-preservation threshold of 0.008. We emphasize that, in text-to-image generation, diversity is meaningful only when the generated images remain faithful to the input text. In contrast, DSI's dominant token selection and dynamic interpolation are highly complementary, maximizing diversity under the semantic-alignment constraint and achieving an optimal trade-off between the two.

Semantic Switching Strategies. To evaluate the effectiveness of dynamic switching, we sweep the switching step m for hard switching ($\hat{\boldsymbol{e}}_j = \boldsymbol{e}_l$) and retain non-dominated configurations to construct its Pareto frontier. We similarly obtain the Pareto frontier for dynamic switching. As shown in Fig. 5(a), within the overlapping metric ranges, the dynamic frontier generally lies closer to the upper-right region, indicating a better balance between image quality and diversity.

Semantic Extraction Models. We adopt CLIP [39] for key semantic extraction based on task suitability, computational efficiency, and empirical performance. Its image–text alignment capability is better suited to identifying semantics associated with low-frequency visual features. With only 0.15B parameters, CLIP requires less time for model loading and semantic extraction (see Tab. 7) and avoids hallucination risks associated with LLM text generation. Experiments show that CLIP performs comparably to Qwen3-8B and Qwen3-0.6B [49] on simple prompts while achieving a better balance between semantic alignment and diversity on complex prompts, as shown in Appendix F.



(a) Switching strategy (b) SD3.5 vs GRPO (c) Only GRPO

Figure 5: Pareto frontiers illustrating the balance between image quality and diversity. (a) Hard versus dynamic switching. (b) PickScore versus DivDino for SD3.5 and Flow-GRPO. (c) PickScore and HPSv2 versus Vendi Score for Flow-GRPO.

Table 7: Semantic extraction time.

Model	Parameters (↓)	Avg. Time (↓)
Qwen3	8.00 B	6.1164
Qwen3	0.60 B	1.6940
CLIP	**0.15 B**	**0.9713**

Table 8: Inference time for a batch.

Model	Resolution	Time (↓)	DSI Time (↓)	Overhead (↓)
SD3-M	512	5.05	5.30	5.05%
FLUX.1-dev	1024	15.97	16.45	2.97%
Flow-GRPO	512	5.00	5.30	5.92%

Sensitivity of Hyperparameter. DSI is robust to hyperparameter. We analyze the sensitivity of the key parameter m in Eq. (8) by plotting the Pareto frontier between consistency and diversity scores. As shown in Figs. 5(b) and (c), sweeping m produces smooth Pareto curves on both SD3.5-M and Flow-GRPO without abrupt fluctuations. As m increases, DivDino improves while consistency-related scores gradually decrease, showing that m controls the diversity–quality trade-off. Furthermore, the effect of m can be compensated for by adjusting β in Eq. (8). When m is large, a smaller β makes the interpolation coefficient rapidly approach 1, introducing full semantics earlier. When m is small, a larger β makes the coefficient approach 1 more slowly, maintaining stronger guidance from $\boldsymbol{c}_l$ within the early interpolation window. Adjusting β yields comparable results across different values of m, indicating that their effects are compensable and allowing two-parameter tuning to be reduced in practice to primarily tuning β. We provide a more detailed analysis of parameter robustness and search difficulty in Appendix G.

Computational overhead. We evaluate the impact of the method on computational overhead. To ensure fair comparison, we maintain consistent settings for the random seed, timesteps, and captions across all trials, with SD3-M, FLUX.1-dev and Flow-GRPO selected as the baseline models. Without loss of generality, we report the average runtime for generating a single sample from a single prompt. Tab. 8 shows that our method adds only a small inference-time overhead across baseline models. Notably, the DSI time already includes the prompt processing for low-frequency semantic extraction and the dynamic semantic interpolation, yet the overhead stays within $2.97\% \sim 5.92\%$. The above evidence demonstrates that the efficiency and practicality of our approach.

7 Conclusion

In this work, we study diversity degradation in text-to-image diffusion models from the temporal mismatch between dynamic visual evolution and static text conditioning. We propose Dynamic Semantic Interpolation (DSI), a training-free strategy that adapts semantic conditioning to the coarse-to-fine denoising process. By using low-frequency semantics in early steps and progressively restoring the full prompt, DSI enables broader structural exploration while maintaining fine-grained text alignment. We also provide a theoretical explanation from the perspective of conditional entropy. Experiments on general, RL-aligned, and distilled models show that DSI consistently improves generation diversity while preserving text fidelity and aesthetic quality.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203100), the National Natural Science Foundation of China (Grant Nos. 62225309, U24A20278, 62361166632, and 62476293), the Scientific Research Innovation Capability Support Project for Young Faculty (Grant No. ZYGXQNJSKYCXNLZCXM-I28), the Explorers Program of Shanghai (Basic Research Funding, Grant No. 25TS1415100), the Shenzhen Science and Technology Program (Grant No. QNXMA20250701093544048), the Zhongguancun Academy (Grant Nos. C20250204 and C20250602), and the General Embodied AI Center of Sun Yat-sen University.

Competing Interests. The authors declare no competing interests.

References

[1] Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

[2] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, et al. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In *CVPR*, 2026.

[5] Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *ICLR*, 2026.

[6] Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18927–18936, 2025.

[7] Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. In *ICCV*, 2025.

[8] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[11] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=g97OHbQyk1`.

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

[13] Ada Gorgun, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. *arXiv preprint arXiv:2512.08486*, 2025.

[14] Haoran He, Yuxiao Ye, Jie Liu, Jiajun Liang, Zhiyong Wang, Ziyang Yuan, Xintao Wang, Hangyu Mao, Pengfei Wan, and Ling Pan. Gardo: Reinforcing diffusion models without reward hacking. *arXiv preprint arXiv:2512.24138*, 2025.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[17] Taihang Hu, Linxuan Li, Joost Van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37: 137646–137672, 2024.

[18] 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.

[19] Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[20] Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[21] 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 *NeurIPS*, 2023.

[22] Dahee Kwon, Haeun Lee, and Jaesik Choi. Breaking the lock-in: Diversifying text-to-image generation via representation modulation. In *ICML*, 2026.

[23] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

[24] Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025.

[25] Jiachen Lei, Keli Liu, Julius Berner, Y HoiM, Hongkai Zheng, Jiahong Wu, and Xiangxiang Chu. There is no VAE: End-to-end pixel-space generative modeling via self-supervised pre-training. In *ICLR*, 2026.

[26] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. *arXiv preprint arXiv:2402.05375*, 2024.

[27] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[29] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2024.

[30] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[31] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

[32] Yingmao Miao, Pengfei Zhang, Xiaochen Lv, Meng Yu, Lei Sun, Xiangxiang Chu, Chao Shen, and Chenhao Lin. Evaluation-verification reward for consistent multi-reference image editing. *arXiv preprint arXiv:2607.29025*, 2026.

[33] Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Jackson Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *ICLR*, 2026.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023.

[35] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, 2022.

[36] 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.

[37] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *CVPR*, 2024.

[38] Xiangyan Qu, Zhenlong Yuan, Jing Tang, Rui Chen, Datao Tang, Meng Yu, Lei Sun, Yancheng Bai, Xiangxiang Chu, Gaopeng Gou, Gang Xiong, and Yujun Cai. From scale to speed: Adaptive test-time scaling for image editing. In *CVPR*, 2026.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[41] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36:66377–66389, 2023.

[42] Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024.

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

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023.

[46] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024.

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[48] Alimama Creative Team. Flux.1-turbo-alpha. `https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha`, 2024. Accessed: 2026-03-04.

[49] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[50] 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 *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7618–7632, 2025.

[51] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *ECCV*, pages 414–430, 2024.

[52] Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *CVPR*, 2025.

[53] Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *ICLR*, 2024.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[55] Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025.

[56] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

[57] 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 preprint arXiv:2306.09341*, 2023.

[58] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. In *NeurIPS*, 2024.

[59] 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.

[60] Meng Yu and Kun Zhan. Bias mitigation in graph diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=CSj72Rr2PB`.

[61] Meng Yu and Kun Zhan. Frequency regulation for exposure bias mitigation in diffusion models. *arXiv preprint arXiv:2507.10072*, 2025.

[62] Meng Yu, Lei Sun, Jianhao Zeng, Xiangxiang Chu, and Kun Zhan. Elucidating the snr-t bias of diffusion probabilistic models. In *CVPR*, 2026.

[63] Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, and Ling Pan. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23625–23635, 2025.

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[65] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *ICLR*, 2026.

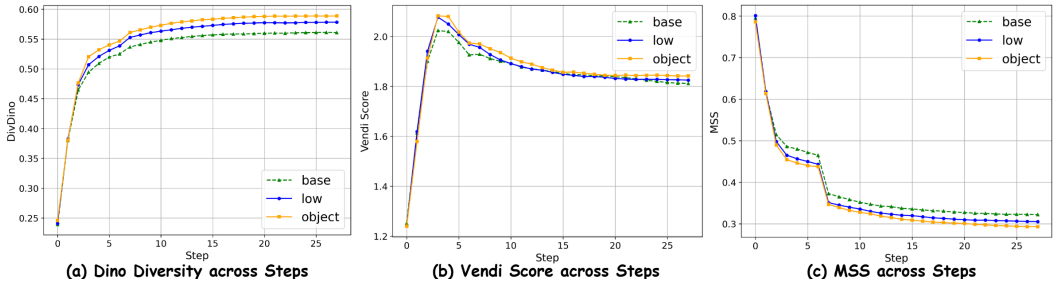


Figure 6: We track the diversity of estimated $\hat{\boldsymbol{x}}_0$, which predicts $\boldsymbol{x}_0$ directly given $\boldsymbol{x}_t$, across denoising timesteps on FLUX.1-dev, using three metrics: (a) DINO Diversity ↑, (b) Vendi Score ↑, and (c) Mean Pairwise Similarity ↓. We found that using low-frequency ($\boldsymbol{c}_{low}$) and subject-only ($\boldsymbol{c}_{sub}$) prompts at the early stages consistently yield significantly higher diversity than the complete static prompt ($\boldsymbol{c}$).

A Semantic Granularity and Diversity

A1. Experimental Setup

We randomly sample 100 captions from MS-COCO [28] and denote each original caption as c. We use SD3.5-Medium as the baseline model and Claude 4.5 as the LLM to extract two sub-prompts from each c: $\boldsymbol{c}_{low}$ and $\boldsymbol{c}_{sub}$.

$\boldsymbol{c}_{low}$ preserves low-frequency semantics such as the overall scene structure and basic content. We extract $\boldsymbol{c}_{low}$ using the following prompt: "For a given batch of prompts, I want to extract sub-prompts that represent low-frequency information. Similar to frequency components in image signals, low-frequency represents overall contours while high-frequency represents texture details. Specifically, articles and quantifiers modifying subject objects must be retained, and the order cannot be changed, nor can new words be added. Below are my 100 prompts; please help me extract the corresponding low-frequency prompts."

$\boldsymbol{c}_{sub}$ contains only the main subject entities. We extract $\boldsymbol{c}_{sub}$ using the following prompt: "For a given batch of prompts, I want to extract sub-prompts that represent only the subject objects. Specifically, I do not care about adjectives, verbs, or high-frequency detail words describing these subjects. These words also include entities representing minor details; for example, for a person, earrings can be considered high-frequency details. Articles and quantifiers modifying the subject objects must be retained. Below are my 100 prompts; please help me extract the corresponding subject object words."

For example, given "Two people flying a kite on a field with clouds in the sky", we extract $\boldsymbol{c}_{low}$ as "Two people flying a kite on a field" and $\boldsymbol{c}_{sub}$ as "Two people, a kite, a field".

A2. The results of FLUX

We perform the same set of tests on FLUX 1-dev [2] to demonstrate that our findings are robust, with the number of sampling steps set to 28.

Fig. 6 presents the experimental results on the FLUX.1-dev model. The results show a consistent hierarchy of diversity across all time steps: regardless of global structural differences, number of visual patterns, or feature similarity, the trend consistently follows $\boldsymbol{c}_{sub} > \boldsymbol{c}_{low} > \boldsymbol{c}$. This phenomenon validates our hypothesis: fixing the full text condition $\boldsymbol{c}$ introduces high-frequency constraints too early, thereby limiting generation diversity. In contrast, using only coarse-grained semantic information (such as $\boldsymbol{c}_{low}$ or $\boldsymbol{c}_{sub}$) in the early stages allows for more exploration of structural variations.

B Further Discussion on the Role of Special Tokens

In Sec 5.2, we highlighted the critical role of special tokens—specifically the `[EoT]` in CLIP and `[PAD]` in T5—as high-dimensional anchors during the Temporal Interpolation Strategy. This section further elucidates their mechanism in aggregating global semantics [50, 26, 6, 17].

First, these special tokens possess an innate capacity for global semantic aggregation rooted in their architectures. In CLIP, the [EoT] token is explicitly optimized via contrastive learning to align with global visual representations, encapsulating the macroscopic semantics of the sequence. Similarly, [PAD] tokens in T5 encoders (used in SD3.5 and FLUX) leverage bidirectional self-attention over all valid tokens. Notably, because T5 typically pads input sequences to a substantial fixed length, [PAD] tokens often dominate the conditioning embeddings in sheer quantity. This significant volume collectively amplifies their impact, allowing them to effectively function as a massive and robust "global semantic pool" for the entire prompt.

Second, this global aggregation provides a robust foundation for our Temporal Interpolation Strategy (Eq. (7)). When dynamically interpolating between the low-frequency embedding $\boldsymbol{e}_l$ and the full-prompt embedding $\boldsymbol{e}$, these tokens act as high-dimensional anchors bridging the hierarchical gap between core structures ($\boldsymbol{c}_l$) and fine-grained details ($\boldsymbol{c}$). By smoothly interpolating these global anchors, the model establishes a unified and continuous semantic reference frame. This ensures that the trajectory from $\boldsymbol{e}_l$ to $\boldsymbol{e}$ is not merely a discrete feature blend, but a cohesive evolution from macroscopic structures to localized details, significantly enhancing the structural consistency of the coarse-to-fine generation process.

C Experimental Setup

In this section, we provide a detailed description of the experimental setup.

C1. Evaluation Metrics

Textual Alignment. To evaluate the alignment between generated images and prompts, we adopt three widely recognized metrics. (1) CLIPScore [39]: it is used to estimate the semantic match between an image and its prompt. It does so by extracting text and image embeddings with a pre-trained CLIP model and then measuring their cosine similarity. (2) PickScore [21]: to evaluate both prompt relevance and perceptual appeal, we use PickScore, which is a reward model learned from large-scale human preference annotations. (3) Hpsv2 [57]: this is a human-preference-oriented evaluation metric built by fine-tuning CLIP on the Human Preference Dataset v2. Since it is trained on large-scale human preference data, HPSv2 better captures human judgments regarding image quality and prompt alignment. It has also been shown to exhibit stronger generalization across diverse image distributions than earlier metrics.

Image Aesthetics. To evaluate the visual aesthetics of generated images, we use Aesthetic Predictor V2.5, a SigLIP-based model that predicts an aesthetic quality score on a 1–10 scale and generalizes well across diverse image domains, including illustrations.

Diversity. We assess the diversity of images generated under the same prompt from three complementary perspectives, aiming to capture both the overall spread of the sample set and the average similarity among individual images. (1) Vendi Score [11]. We use the Vendi Score to comprehensively assess diversity in our generated samples. This metric analyzes the eigenvalue spectrum of the similarity matrix computed from a set of images, capturing both the range of variations present and how uniformly these variations are represented across the collection. (2) Mean Pairwise Similarity (MSS). To complement this analysis with a more intuitive measure of sample redundancy, we also calculate the MSS. We first encode each image using the SSCD model [35] , a self-supervised feature extractor designed for image copy detection. We then compute cosine similarity between all pairs of image embeddings and average these values (excluding self-comparisons along the diagonal). Lower MSS values indicate higher diversity, meaning the generated images are more distinct from one another on average.

C2. Other Setting

For the main experiments, we use 28 sampling steps for FLUX.1-dev, 50 steps for the SD series, and 8 steps for distilled models. For the sampling batch, all our experiments generate four diverse images per prompt by using different random seeds, which are then used to compute diversity metrics. Regarding the number of prompts, we use 1,000 captions from the MS-COCO dataset in the main experiments, while for ablation studies, we select 100 captions due to computational constraints. Specifically, since Scale-Group [33] is an inference-time scalable framework that demands additional



Figure 7: Visualizations of integrating our proposed DSI on FLUX.1-dev, which illustrate that DSI can improve diversity without sacrificing image quality.

computational resources, we also select 100 captions for our evaluation. For all baseline models, we use the official open-source code provided in the original papers for reproduction and comparison. Specifically, since the CFG-TV [19] code repository also includes a branch for CFG-In [23], we use this repository to conduct the comparison with CFG-In.

D Qualitative Comparison

To evaluate the effectiveness of our proposed DSI in enhancing generation diversity, we conducted qualitative comparisons against two state-of-the-art baselines: FLUX.1-dev (Fig. 7) and SD3.5-M [10] (Fig. 8). As observed, given specific prompts, the baseline models exhibit noticeable homogenization. They frequently generate images with repetitive compositions, camera angles, and character poses. For instance, the baseline model in Fig. 8 produces nearly identical silhouettes for the "a girl walking through the forest holding a small lantern" prompt.

In contrast, integrating our DSI method significantly improves visual diversity. For the exact same prompts, DSI successfully yields varied camera perspectives, character attributes, lighting conditions, and even desirable stylistic variations. A clear example of this is seen in Fig. 7, where DSI introduces distinct styles ranging from 3D animation to photorealism. Crucially, this enhanced diversity effectively mitigates mode collapse without sacrificing text alignment or the high-fidelity image quality of the original models.

E More Comparative Experiments

Scale-Group [33] proposes a scalable group reasoning approach that simultaneously enhances both the diversity and quality of a batch of samples. At the beginning of the denoising process, Scale-Group initializes K random noise instances as a candidate subset. It then performs denoising inference on each noise sample and applies a reward-based pruning strategy during denoising, ultimately producing M samples. We emphasize that our method, which aligns textual and visual frequency characteristics throughout the denoising process, is orthogonal to this group reasoning framework and thus can be seamlessly integrated into Group to further enhance its performance.

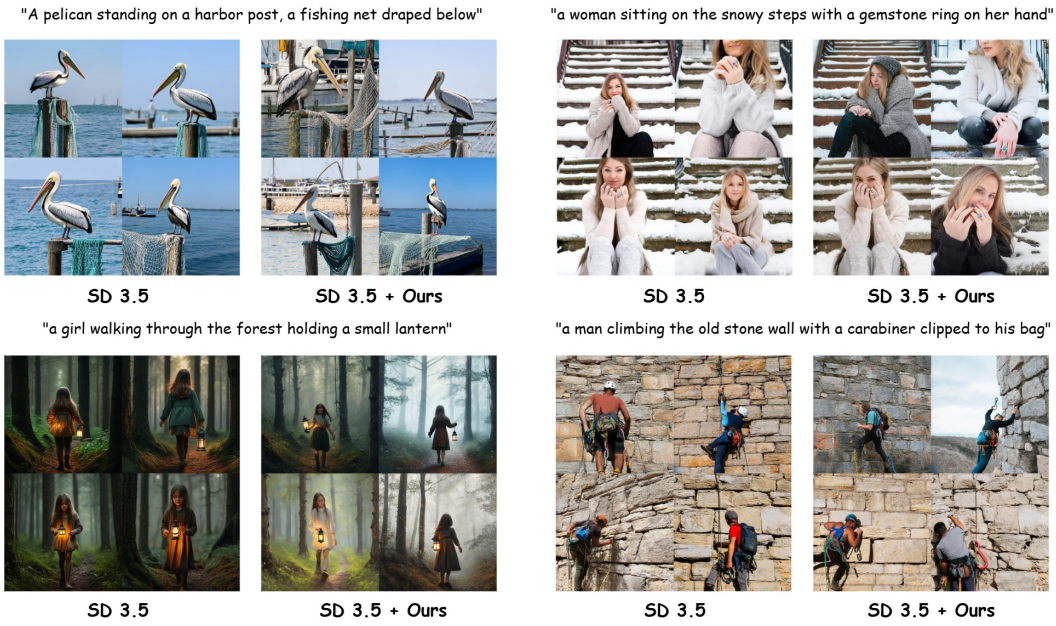


Figure 8: Visualizations of integrating our proposed DSI on SD3.5-M, which illustrate that DSI can improve diversity without sacrificing image quality.

Table 9: Quantitative experiments on Scale-Group (ICLR 2026).

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
K=8	**0.306**	**0.337**	**0.231**	6.041	0.571	1.782	0.316
+Ours	0.304	0.333	0.230	**6.076**	**0.581**	**1.787**	**0.302**
K=16	**0.305**	**0.336**	**0.231**	6.048	0.583	1.829	0.306
+Ours	0.302	0.331	0.230	**6.084**	**0.608**	**1.849**	**0.277**
K=32	**0.305**	**0.340**	**0.232**	6.039	0.610	1.843	0.288
+Ours	0.302	0.335	0.230	**6.083**	**0.621**	**1.897**	**0.278**
K=64	**0.301**	**0.338**	**0.231**	5.994	0.629	1.866	0.274
+Ours	**0.301**	**0.338**	0.230	**6.121**	**0.651**	**1.931**	**0.245**

Tab. 9 shows the quantitative comparison between Scale-Group and Scale-Group augmented with our method under different scaling factors, where a larger K indicates a larger initial candidate noise set. Overall, our method consistently improves the performance of Scale-Group across different values of K, showing that it is highly complementary to the group reasoning paradigm. In particular, although Scale-Group itself benefits from increasing K, integrating our frequency-alignment strategy further enhances generation quality and diversity simultaneously. Across all settings, our method consistently improves Aesthetic, DivDino, Vendi, and MSS, indicating better visual quality, stronger diversity, and lower semantic redundancy. Notably, the gains become more evident at larger scales: for example, at $K = 64$, our method improves Aesthetic from 5.994 to 6.121, DivDino from 0.629 to 0.651, Vendi from 1.866 to 1.931, and reduces MSS from 0.274 to 0.245. These results show that our method remains effective even when combined with strong large-scale sampling and pruning, validating that textual-visual frequency alignment is orthogonal to Scale-Group and can serve as a plug-and-play enhancement for scalable group generation.

DiffusionNFT [65] is an online reinforcement learning framework for diffusion models that directly optimizes generation on the forward flow-matching process, avoiding the likelihood estimation and sampling-trajectory constraints in prior RL-based diffusion tuning methods. Instead of discretizing

Table 10: Quantitative experiments on DiffusionNFT (ICLR 26) and FLUX-Turbo.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
DiffusionNFT	**0.321**	**0.350**	**0.238**	5.702	0.467	1.608	0.497
+Ours	0.317	0.340	0.237	**5.713**	**0.500**	**1.655**	**0.471**
FLUX-Turbo [48]	**0.310**	**0.339**	**0.232**	**6.216**	0.561	1.781	0.327
+Ours	0.295	0.317	0.227	6.121	**0.623**	**1.928**	**0.259**

the reverse sampling process, it contrasts positive and negative generations to define an implicit policy improvement direction, allowing reinforcement signals to be naturally incorporated into the standard supervised training objective. This design supports arbitrary black-box reward models, simplifies integration with diffusion training, and enables more efficient optimization.

As presented in Tab. 10, applying our method to DiffusionNFT yields a substantial enhancement in generation diversity. Specifically, all three diversity metrics exhibit marked improvements: DivDino increases from 0.467 to 0.500, Vendi score rises from 1.608 to 1.655, and MSS drops from 0.497 to 0.471, indicating a much broader exploration of the generation space.Importantly, this significant boost in diversity is achieved with virtually no sacrifice to semantic alignment or visual quality. The semantic consistency metrics, including HPSv2, CLIPScore, and PickScore, experience only negligible fluctuations and remain highly competitive with the baseline. Furthermore, the Aesthetic score sees a marginal improvement from 5.702 to 5.713. These quantitative results confirm that our method effectively mitigates the homogenization issue even in RL-tuned diffusion models, unlocking highly diverse outputs while strictly preserving the original model's text fidelity and aesthetic appeal.

F Comparison of Semantic Extractors

To evaluate the choice of semantic extractor, we compare CLIP-0.15B with Qwen3-8B [49] under hard switching by varying the switching step m. As shown in Tab. 11, the two extractors exhibit comparable Pareto frontiers on simple MS-COCO prompts, with identical PickScores at the reported precision and Vendi differences of at most 0.019. On the longer and more complex DPG-Bench prompts, CLIP achieves a superior empirical Pareto frontier: every evaluated Qwen3-8B configuration is dominated by at least one CLIP configuration. For example, CLIP at $m = 5$ achieves a PickScore of 0.2180 and a Vendi score of 1.8746, exceeding Qwen3-8B at $m = 6$ on both metrics (0.2177 and 1.8669, respectively). At the same PickScore of 0.2183, CLIP also improves Vendi from 1.8304 to 1.8358. These results show that CLIP performs comparably to Qwen3-8B on simple prompts while providing a better Pareto trade-off between semantic alignment and diversity on complex prompts, supporting its use as an effective and lightweight semantic extractor.

G Sensitivity of Hyperparameter

In this section, we further investigate how sensitive the proposed method is to various parameters. We emphasize that, due to computational resource constraints, we used 100 MS-COCO captions as prompts in our ablation studies.

G1. Number of low-frequency words k

We first fix the parameters m and β to investigate the sensitivity of our method to the number of low-frequency semantic tokens k. Specifically, k denotes the top-k low-frequency words selected from the complete prompt as shown in Sec 5.1, which controls the degree of "relaxation" in the low-frequency semantic condition. As shown in Tab. 12, as k increases, $\boldsymbol{c}_{low}$ progressively incorporates more semantic information from the full prompt, thereby narrowing the search space: this improves text-image alignment while reducing the diversity of generated outputs.

In summary, our method exhibits a clear and consistent trend as k varies: diversity gradually decreases while text alignment steadily improves with increasing k. This demonstrates that our method achieves

Table 11: Comparison of semantic extractors on MS-COCO and DPG-Bench prompts.

		Prompt: MS-COCO			Prompt: DPG-Bench		
Semantic Extractor	Size	m	PickScore ↑	Vendi ↑	m	PickScore ↑	Vendi ↑
Qwen3	8B	7	**0.224**	1.939	6	**0.2177**	1.8669
CLIP	0.15B	7	**0.224**	**1.958**	6	0.2175	**1.9031**
Qwen3	8B	5	**0.225**	**1.932**	5	**0.2180**	1.8506
CLIP	0.15B	5	**0.225**	1.924	5	**0.2180**	**1.8746**
Qwen3	8B	4	**0.225**	1.903	4	0.2180	1.8463
CLIP	0.15B	4	**0.225**	**1.909**	4	**0.2181**	**1.8657**
Qwen3	8B	3	**0.226**	1.876	3	**0.2183**	1.8304
CLIP	0.15B	3	**0.226**	**1.894**	3	**0.2183**	**1.8358**

Table 12: Sensitivity experiments on the hyperparameter k.

Method	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
$k = 1$	0.288	0.327	0.227	5.335	0.618	1.934	0.269
$k = 3$	0.294	0.334	0.229	5.420	0.590	1.862	0.293
$k = 5$	0.297	0.332	0.229	5.398	0.580	1.851	0.303
$k = 7$	0.296	0.334	0.229	5.403	0.573	1.840	0.312
$k = 9$	0.298	0.336	0.230	5.414	0.567	1.837	0.317

a favorable trade-off between diversity and text consistency, and remains robust to the choice of hyperparameter k, reflecting strong adaptability and balance.

G2. Power-law scheduling exponent β

The parameter β controls the temporal evolution of the interpolation coefficient in Eq. (8). We conduct a sensitivity analysis over $\beta \in [0.06, 5.00]$ to examine its influence on semantic consistency and generation diversity. As shown in Tab. 13, small values within $\beta \in [0.06, 0.14]$ yield relatively stable performance, while the broader sweep reveals a clear trade-off between prompt alignment and output diversity.

Semantic consistency generally decreases as β increases. Specifically, HPSv2 drops from 0.298 at $\beta = 0.06$ to 0.227 at $\beta = 5.00$, accompanied by similar declines in CLIPScore and PickScore. Within the narrower range of 0.06–0.14, these alignment metrics remain relatively stable: HPSv2 varies from 0.298 to 0.296, CLIPScore from 0.338 to 0.336, and PickScore from 0.230 to 0.229. This behavior is consistent with the interpolation schedule: smaller β values produce a faster initial increase in α_j, introducing full-prompt semantics earlier in denoising. Larger values instead delay this transition, allowing the low-frequency embedding $\mathbf{e}_l$ to exert greater influence during the early stages, but potentially weakening prompt alignment.

Diversity exhibits a non-monotonic response to β. DivDino reaches its maximum of 0.633 at $\beta = 0.50$, whereas Vendi reaches its maximum of 1.982 and MSS its minimum of 0.246 at $\beta = 1.00$. Increasing β further to 5.00 reduces DivDino and Vendi to 0.613 and 1.958, respectively, while increasing MSS to 0.252. Thus, delaying the introduction of full-prompt semantics can improve diversity, but an excessively slow initial transition does not necessarily provide additional gains.

Table 13: Sensitivity experiments on the hyperparameter β.

β	HPSv2 ↑	CLIPScore ↑	PickScore ↑	Aesthetic ↑	DivDino ↑	Vendi ↑	MSS ↓
0.06	0.298	0.338	0.230	5.394	0.594	1.849	0.288
0.08	0.297	0.337	0.230	5.388	0.595	1.859	0.284
0.10	0.297	0.337	0.230	5.389	0.599	1.876	0.282
0.12	0.296	0.336	0.229	5.400	0.599	1.866	0.280
0.14	0.296	0.336	0.229	5.397	0.601	1.870	0.279
0.50	0.276	0.307	0.223	5.331	0.633	1.968	0.249
1.00	0.251	0.273	0.215	5.210	0.628	1.982	0.246
5.00	0.227	0.238	0.208	5.145	0.613	1.958	0.252

These results show that β provides an interpretable means of adjusting the balance between semantic consistency and generation diversity. In the evaluated setting, $\beta \in [0.06, 0.14]$ maintains strong prompt alignment with relatively stable performance, while $\beta \in [0.50, 1.00]$ favors greater diversity at the cost of alignment. We expect the appropriate operating range to depend on the task and model, rather than being universally fixed, and recommend selecting β empirically according to the desired balance between prompt fidelity and output diversity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the temporal-mismatch motivation, the proposed DSI method, and the diversity-improvement claims, all of which are supported by main manuscript.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the diversity-consistency trade-off in experiment section, and additional limitations are discussed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theorem are stated with assumptions in Sec. 5.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiment and appendix provides all backbones, datasets, metrics and hyperparameters needed to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include anonymized code in the supplemental material; the code will be released under a permissive license upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experiment and appendix provides all backbones, datasets, metrics and hyperparameters needed to reproduce the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined, or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: All metrics are averaged over 4000 generated images per setting (1000 prompts $\times$ 4 samples), which provides empirical stability; we omit error bars due to the substantial cost of repeated full-scale evaluations across many backbones.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiment section reports per-image inference time on a single NVIDIA A6000 GPU for our methods.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly available pre-trained models and datasets, involves no human subjects, and preserves anonymity, in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: DSI mitigates mode collapse and stereotyped outputs (positive impact); as a training-free inference-time modifier, it does not introduce misuse risks beyond those of the underlying base models.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release no new pre-trained model or dataset; DSI is a training-free strategy applied on top of existing public models, which retain their original safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used pre-trained models, the CLIP text encoder, and the MS-COCO dataset are properly cited and used under their original licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The DSI codebase is provided anonymously in the supplemental material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work involves no crowdsourcing or human subjects; HPSv2 and PickScore are computed using publicly available pre-trained reward models.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve research with human subjects, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An LLM is used in Sec. 4 to extract motivational analysis. The proposed DSI method itself relies on an automated CLIP-based token-importance estimation and does not require an LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.