Breaking Semantic Boundaries: Distribution-Guided Semantic Exploration for Creative Generation

Fu Feng[1,2] Yucheng Xie[1,2] Ruixiao Shi[1,2] Xu Yang[1,2] Jing Wang[1,2*] Xin Geng[1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{fufeng, xieyc, eric_xiao, xuyang_palm, wangjing91, xgeng}@seu.edu.cn

Figure 1. We introduce **Distribution-Conditional Generation**, a novel paradigm that breaks semantic boundaries through fine-grained, controllable concept fusion, and propose DisTok, an encoder–decoder framework **unifying conditional and unconditional semantic exploration**, enabling efficient and flexible discovery of novel visual concepts.

Abstract

Text-to-image (T2I) diffusion models effectively produce semantically aligned images, but their reliance on training distributions constrains their capacity for synthesizing truly novel, out-of-distribution concepts. Existing methods attempt to enhance creativity through semantic exploration, such as fusing known concept pairs, but the resulting images remain linguistically describable and confined to familiar semantic spaces. Inspired by the soft probabilistic outputs of classifiers on novel or out-of-distribution inputs, we propose Distribution-Conditional Generation, a paradigm that models novel concepts as image synthesis conditioned on class distributions, enabling controllable yet semantically unconstrained creative generation. Building on this, we propose DisTok, an encoder–decoder framework that unifies conditional and unconditional creative generation by decoding latent representations—either randomly sampled or mapped from conditions (e.g., class distributions)—into tokens representing novel concepts. DisTok is trained by iteratively sampling and fusing concept pairs from a dynamic pool to model progressively complex distributions, while enforcing semantic consistency through a vision-language model that aligns the class distributions of generated images with the input distributions. Extensive experiments demonstrate that DisTok enables efficient and flexible semantic exploration for token-level creative synthesis, achieving state-of-the-art text–image alignment and human preference.

1. Introduction

Recent advances in text-to-image (T2I) models such as Stable Diffusion 3 [7] and Midjourney demonstrate remarkable ability to generate *semantically aligned* images from natural language prompts [4, 34, 35], reflecting the maturity of current generative models in reproducing established visual concepts. However, this reproducibility primarily stems from their proficiency in modeling complex training data distributions [3, 33], which tends to reinforce familiar semantics and limits their capacity for out-of-distribution creativity

*Corresponding authors.

involving unknown or ambiguous concepts [5, 43].

Consequently, recent efforts have sought to foster creativity in generative models [9, 17, 30] through semantic exploration aimed at synthesizing novel concepts. BASS [17] explores complex semantics by combining pairs of known concepts, enabling semantically controlled out-of-distribution synthesis while remaining constrained within established conceptual boundaries. In contrast, ConceptLab [30] identifies tokens beyond established regions to explore novel semantic spaces, yet this unconditional approach excludes human-guided priors, making alignment with human intent challenging. Hence, a more practical approach to semantic exploration is needed to break semantic boundaries while maintaining controllable creative concept generation.

A key challenge in semantic exploration is ***how to semantically characterize unknown concepts***. Classification models offer a useful perspective: when confronted with ambiguous or out-of-distribution inputs, they typically produce ***soft probability distributions over known classes***, capturing partial similarities without assigning a definitive label. This suggests that novel concepts, though semantically unknown, often follow recognizable patterns of known categories. Building on this insight, we invert the classification process by generating images conditioned on class distributions—a paradigm we term **Distribution-Conditional Generation**—which enables fine-grained, controllable concept fusion and provides an expanded interpretable space for unconditional semantic exploration.

However, existing methods are limited to combinatorial creativity between two known concepts, typically achieved by combining reference images [17] or tokens [9], or via diffusion-process interpolation [18, 36, 40], which restricts their scalability to fine-grained control over class distributions involving multiple concepts. Thus, we propose **DisTok**, an encoder–decoder framework with a learnable **Distribution Encoder** that maps arbitrary class distributions into a latent space and a **Creative Decoder** that transforms any latent vectors—either encoded or randomly sampled—into creative concept tokens (Fig. 1). In this way, DisTok inherently unifies conditional and unconditional exploration, enabling controllable generation and efficient discovery of novel concepts, while its generated tokens can be seamlessly embedded into natural language prompts to preserve semantic consistency across diverse scenes and styles.

DisTok maintains a Concept Pool $\mathcal{P}$ initialized with tokens of known concepts and jointly optimizes two complementary objectives to generate ***increasingly complex*** and ***semantically consistent*** creativity. To facilitate an increasingly complex semantic composition, DisTok continuously encode concept pairs (c_1, c_2) sampled from $\mathcal{P}$ and decodes them into a single concept token t_{crt}. This process is indirectly supervised by aligning an adaptive prompt ("a photo of a $\langle t_{\text{crt}} \rangle$") with restrictive prompts ("a c_1 c_2" and "a photo of a cute pet"). To enhance semantic consistency, DisTok periodically samples latent vectors from a Gaussian prior, decodes them into novel tokens, and synthesizes corresponding images. A pretrained vision-language model is then queried to infer class distributions (e.g., via "What animal is in the photo?"). The predicted distribution is used to supervise the alignment between the input class distribution and the visual semantics of the output token. Resulting tokens are added to $\mathcal{P}$, enabling subsequent semantic composition toward increasingly complex class distributions.

DisTok is trained on known classes from *CangJie* [9] and, once trained, can ***directly generate tokens*** representing creative concepts without additional optimization, achieving up to 13× and 40× speedups over BASS and ConceptLab, respectively. The generated tokens maintain strong semantic consistency across prompts, enabling high-quality image synthesis with enhanced human preference and text–image alignment. Comprehensive evaluations with GPT-4o and user studies further validate DisTok's superiority in concept integration and originality.

Our key contributions are as follows: (1) We introduce Distribution-Conditional Generation, a novel paradigm that leverages class distributions to more effectively break semantic boundaries for creative generation. (2) We propose DisTok, an encoder–decoder framework that unifies conditional and unconditional exploration for fine-grained, controllable concept composition and efficient discovery of novel visual concepts. (3) Extensive experiments validate DisTok's state-of-the-art performance in creative generation, consistently outperforming both standard T2I models and creativity-focused methods in efficiency and diversity.

2. Related Work

Creative Generation. Text-to-image (T2I) models have demonstrated impressive capabilities in synthesizing semantically coherent images [13, 23, 39], yet extending these models toward creative generation remains a central challenge for machine intelligence [21, 22]. Some diffusion-based methods [11, 20] directly explore the latent space for creative image generation, but the resulting outputs often lack semantic flexibility and are difficult to reuse with natural language prompts for diverse scenes.

To preserve semantic flexibility, recent methods explore the creativity of T2I models by discovering tokens for previously unseen concepts. For instance, BASS [17] and AGSwap [41] employ sampling and swapping strategies to combine pairs of distinct concepts, CreTok [9] introduces a dedicated token explicitly encoding combinational semantics, and ConceptLab [30] promotes novelty by progressively decreasing similarity to existing concepts. However, these techniques remain constrained by discrete pairwise combinations, restricting exploration of unknown semantic space and preventing the full breaking of semantic boundaries. Thus,

we propose Distribution-Conditional Generation, which conditions semantic exploration on distributions over known classes, enabling fine-grained, controllable exploration in a semantically unconstrained space.

Personalized Visual Content Generation. Personalization aims to generate coherent representations of target concepts from a few exemplar images [27, 32]. Early methods such as Textual Inversion [10, 31] embed new identities into unique tokens via optimization on reference images. Recent work emphasizes modularity and compositionality, with PartCraft [24] and Chirpy3D [25] modeling visual entities at the part level to enable flexible recombination. Beyond token-level adaptation, some approaches adapt specific network components [12, 15] or introduce external adapters [2, 6, 38] to improve flexibility. Other methods, such as MagicMix [18], DiffMorpher [40], and ATIH [36], promote innovative visual representations by interpolating semantic representations during the diffusion process [42]. In contrast, DisTok directly generates conceptual mixtures from latent vectors—either derived from specified class distributions or randomly sampled—producing tokens that capture abstract concepts and enable zero-shot generation of unseen concepts without further training or adaptation.

3. Methods

3.1. Overview of DisTok

Unlike existing semantic exploration approaches for creative generation [9, 17, 30], DisTok unifies conditional and unconditional exploration within an encoder-decoder framework. The **Distribution Encoder** $\mathcal{E}_{\text{dis}}$ projects class distributions into a latent space, enabling the modeling of complex, hard-to-describe visual concepts, while the **Creative Decoder** $\mathcal{D}_{\text{tok}}$ maps these latent representations—either derived from specified class distributions or randomly sampled—into tokens that can be seamlessly integrated into prompts (Fig. 2).

Formally, given a class distribution $p_c \in \Delta^K$, $\mathcal{E}_{\text{dis}}$ projects p_c into a latent vector $z = \mathcal{E}_{\text{dis}}(p_c) \in \mathbb{R}^\delta$, which is then decoded into a creative concept token $t_{\text{crt}} = \mathcal{D}_{\text{tok}}(z) \in \mathbb{R}^d$ ($\delta \ll d$). To progressively model more complex and semantically coherent concepts, DisTok maintains a **Concept Pool** $\mathcal{P}$ initialized with tokens of known concepts and employs three key components to jointly train $\mathcal{E}_{\text{dis}}$ and $\mathcal{D}_{\text{tok}}$:

- **Continuous Concept Fusion** (Sec. 3.2). DisTok continuously samples class pairs from $\mathcal{P}$ and trains to fuse them into unified concepts, progressively enabling $\mathcal{D}_{\text{tok}}$ to decode tokens corresponding to increasingly complex class distributions from latent representations.
- **Class Distribution Estimation** (Sec. 3.3). DisTok generates novel concepts by decoding randomly sampled latent vectors and predicting their class distributions via VLMs, with the resulting tokens and distributions added to $\mathcal{P}$ as novel concepts for further training.
- **Distribution Consistency Enforcement** (Sec. 3.4). DisTok enforces consistency between the input class distribution and the one predicted from generated images, providing direct supervision for both $\mathcal{E}_{\text{dis}}$ and $\mathcal{D}_{\text{tok}}$ using novel concepts from $\mathcal{P}$.

3.2. Continuous Concept Fusion based on Class Pairs

To train the Distribution Encoder $\mathcal{E}_{\text{dis}}$ and Creative Decoder $\mathcal{D}_{\text{tok}}$ for effective class distribution encoding and creative concept generation, it is first necessary to synthesize distribution-describable images, since no direct training data are available for this purpose.

BASS [17] and CreTok [9] demonstrate that basic combinatorial creativity can be achieved through text-pair fusion, while ConceptLab [30] shows that newly generated tokens can be recursively combined. This suggests that continuously composing existing or novel concepts enables the generation of images aligned with specific class distributions. For example, given classes $c_1, c_2, c_3 \in \mathcal{C}$, a target class distribution $p_c = (0.25, 0.25, 0.5)$ over (c_1, c_2, c_3) can be approximately synthesized by first combining (c_1, c_2) into an intermediate concept c_{inter}, then merging c_{inter} with c_3.

Thus, DisTok iteratively samples class pairs from the current Concept Pool $\mathcal{P}$ [8] and performs concept fusion to progressively enhance the Creative Decoder's ability to synthesize images corresponding to increasingly complex class distributions, as illustrated in Fig. 2a. Specifically, given two concepts (i.e. classes) $c_1, c_2 \in \mathcal{P}$, their corresponding tokens t_1, t_2 are combined and encoded into a latent vector $z = \mathcal{E}_{\text{dis}}(t_1 + t_2)$, which is subsequently decoded into a creative token $t_{\text{crt}} = \mathcal{D}_{\text{tok}}(z)$.

To facilitate the fusion of c_1 and c_2, we construct an *adaptive prompt* $q_a =$ "a photo of a $\langle t_{\text{crt}} \rangle$" and a *restrictive prompt* $q_r(c_1, c_2) =$ "a c_1 c_2", following the fusion strategy in CreTok [9]. To further align generation with human-preferred semantics, we introduce an auxiliary restrictive prompt $q_s =$ "a photo of a cute pet". Semantic fusion is encouraged by minimizing:

$$\tilde{\mathcal{L}}_{\text{mix}} = (1-\cos(E(q_r), E(q_a))) + (1-\cos(E(q_s), E(q_a))). \tag{1}$$

where $\cos(a, b) = \frac{a \cdot b}{||a|| ||b||}$ denotes cosine similarity, and $E(\cdot)$ is the text encoder (i.e., CLIP-L/14 [28] in Kandinsky 2.1). To prevent overfitting to a dominant concept, which can artificially inflate similarity by disproportionately emphasizing one concept while neglecting others, we introduce thresholding to regulate concept fusion. The final loss $\mathcal{L}_{\text{mix}}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{mix}} =& (1 - \min[\cos(E(q_r), E(q_a)), \theta_1]) \\ &+ (1 - \min[\cos(E(q_s), E(q_a)), \theta_2]). \end{aligned} \tag{2}$$

where θ_1 and θ_2 are predefined thresholds controlling the maximum allowed similarity.

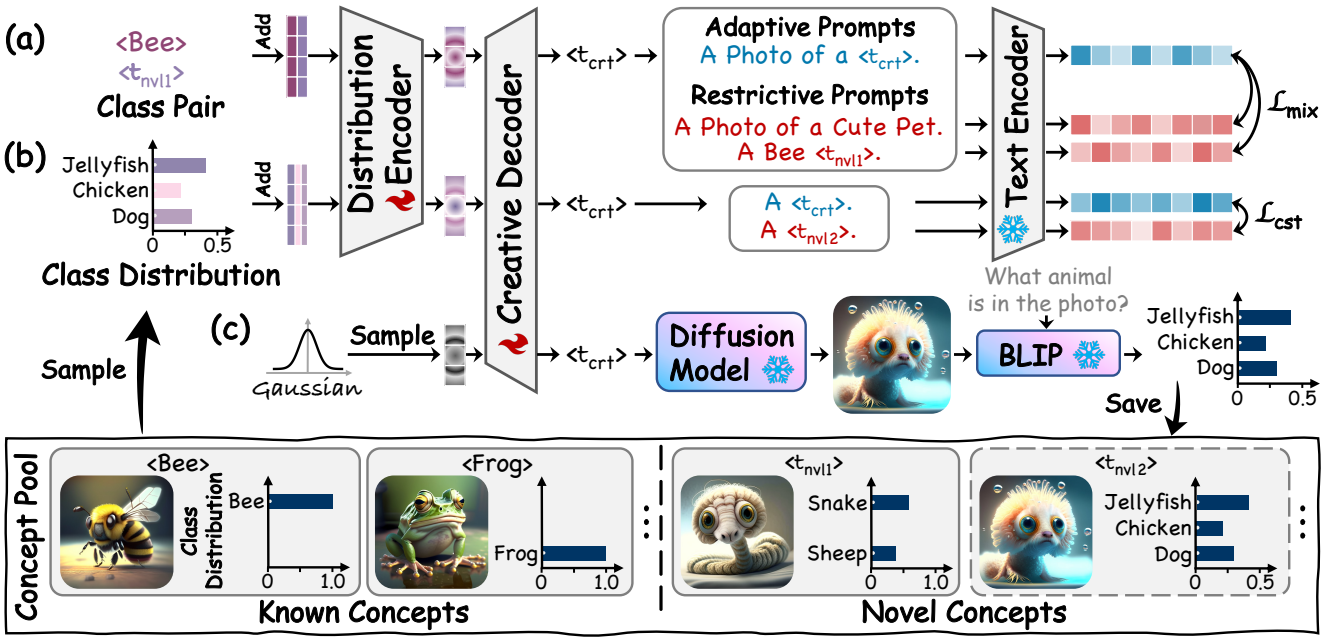

Figure 2. Overview of DisTok. At each training step, DisTok performs either **(a) Continuous Concept Fusion**, where a class pair is sampled to train the model to generate tokens reflecting increasingly complex class distributions; or **(b) Distribution Consistency Enforcement**, where a class distribution is sampled to align the encoder and decoder with the visual semantics of generated tokens. **(c) Class Distribution Estimation** is periodically conducted by randomly sampling latent vectors and decoding them into tokens, whose class distributions are subsequently predicted by a VLM. Resulting tokens and distributions are saved in *Concept Pool* for subsequent fusion and supervision.

3.3. Class Distribution Estimation via Vision-Language Models (VLMs)

Building on the continuous concept fusion strategy in Sec. 3.2, DisTok progressively learns to generate tokens whose corresponding images x_{crt}, produced by diffusion models via $x_{\text{crt}} = \mathcal{G}_{\text{diff}}(t_{\text{crt}})$, reflect increasingly complex class distributions. However, the fusion objective in Eq. (2) provides only indirect supervision, encouraging the generated images to exhibit overall semantic fusion of (c_1, c_2) at the prompt level, without explicitly constraining the contribution of each class within t_{crt}, thereby limiting fine-grained control over concept composition.

To enhance supervision, we extract fine-grained class distributions from the visual domain using a pretrained VLM, such as BLIP [16]. Given a generated image x_{crt}, the VLM is queried with a prompt like "What animal is in the photo?" and its token-level output logits $p_{\text{vlm}}(c|x_{\text{crt}})$ are obtained over the known concept set $\mathcal{C}_{\text{knw}} \subset \mathcal{P}$ (Fig. 2c). The resulting class distribution $p_{\text{crt}}(c) = \text{softmax}(p_{\text{vlm}}(c|x_{\text{crt}}))$ reflects the visual semantic distribution, and can thus be used to encourage consistency between the input class distribution and the visual semantics of the generated token t_{crt}.

Images x_{crt} are generated through unconditional exploration by randomly sampling latent vectors $z \sim \mathcal{N}(0, I)$ from a standard Gaussian distribution, decoding them into creative tokens $t_{\text{crt}} = \mathcal{D}_{\text{tok}}(z)$, and synthesizing the corresponding images $x_{\text{crt}} = \mathcal{G}_{\text{diff}}(t_{\text{crt}})$. To ensure these random samples represent meaningful concepts rather than semantic noise, we regularize the latent space [25] by minimizing:

$$\mathcal{L}_{\text{reg}} = \frac{1}{\sigma(z)^2}\mathbb{E}_z\left[||\mu(z)||_2^2\right], \tag{3}$$

where $\mu(z)$ and $\sigma(z)$ denote the mean and standard deviation of the latent vectors, respectively.

This regularization encourages a near-zero mean and sufficient variance in the latent space, promoting diversity and preventing mode collapse. Consequently, new tokens and images can be generated by sampling z from any distribution $\mathcal{Q}$ satisfying $\mathbb{E}_{z\sim\mathcal{Q}}[z] = 0$, enabling open-ended and unconditional exploration (Sec. 5.2).

The sampling process of latent vectors for class distribution prediction is periodically performed during training. The resulting tokens t_{crt} and their associated distributions p_{crt} are added to $\mathcal{P}$ as novel concepts $(t_{\text{nvl}}, p_{\text{nvl}})$ if they satisfy $\max_c p_{\text{crt}}(c) < \tau$, where τ is a predefined threshold controlling concept novelty. These novel concepts enable higher-order semantic fusion (Sec. 3.2) and provide more precise, fine-grained supervision for enforcing distribution consistency (Sec. 3.4).

3.4. Distribution Consistency Enforcement between Language and Visual Semantics

Building upon the expanded $\mathcal{P}$, which includes tokens of novel concepts t_{nvl} with VLM-predicted class distributions

p_{nvl} from the visual domain, we further supervise $\mathcal{E}_{\text{dis}}$ to improve alignment between input distributions and the visual semantics of decoder outputs, as shown in Fig. 2b.

Specifically, during training, we sample a novel token t_{nvl} along with its associated class distribution p_{nvl}. We compute the weighted combination of relevant known concepts as $\Sigma_i p_{\text{nvl}}(i) t_i$, where t_i denotes the token vector associated with class $c_i \in \mathcal{C}_{\text{knw}}$. This combined token is then encoded into a latent vector $z = \mathcal{E}_{\text{dis}}(\Sigma_i p_{\text{nvl}}(i) t_i)$, and subsequently decoded into a creative token t_{crt}.

To enforce distribution consistency, we minimize the cosine distance between the text embeddings of the generated token t_{crt} and the novel token t_{nvl} sampled from $\mathcal{P}$.

$$\mathcal{L}_{\text{cst}} = 1 - \cos(E(t_{\text{crt}}), E(t_{\text{nvl}})), \tag{4}$$

where $E(\cdot)$ denotes the text encoder and $\cos(\cdot, \cdot)$ denotes the cosine similarity, both defined in Eq. (1).

Unlike the indirect supervision provided by concept fusion (Eq. (2)), VLM-based consistency explicitly grounds creative tokens in visual semantics. By directly aligning generated tokens with their intended class distributions, it encourages $\mathcal{E}_{\text{dis}}$ to capture distributional semantics more faithfully, enabling DisTok to produce tokens of creative concepts that are both visually novel and semantically coherent with multi-concept compositions.

3.5. Iterative Training of DisTok

Each training iteration consists of n sampling steps, where at each step we randomly perform either concept fusion (Sec.3.2) or distribution consistency enforcement (Sec.3.4) to jointly optimize $\mathcal{E}_{\text{dis}}$ and $\mathcal{D}_{\text{tok}}$. The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \frac{1}{n} \sum_{i=1}^{n} \left(\alpha \mathbb{I}_{\text{mix}}^{(i)} \mathcal{L}_{\text{mix}}^{(i)} + \beta \mathbb{I}_{\text{cst}}^{(i)} \mathcal{L}_{\text{cst}}^{(i)} + \gamma \mathcal{L}_{\text{reg}}^{(i)} \right), \tag{5}$$

where $\mathbb{I}_{\text{mix}}^{(i)}, \mathbb{I}_{\text{cst}}^{(i)} \in \{0, 1\}$ with $\mathbb{I}_{\text{mix}}^{(i)} + \mathbb{I}_{\text{cst}}^{(i)} = 1$, and α, β, γ are weighting coefficient.

4. Experiments

Datasets. We use the *CangJie* [9] dataset, which contains 60 common concepts. Originally designed for class-pair conditioned semantic exploration, it includes 30 text pairs for evaluating two-concept combinatorial creativity. To enable distribution-conditional generation, we extend *CangJie* by creating 30 class distributions through random combinations and weighting of the original concepts.

Experimental Setup. Our implementation is based on Kandinsky 2.1 [29], which employs CLIP-L/14 [28] as the text encoder. The distribution encoder $\mathcal{E}_{\text{dis}}$ and creative decoder $\mathcal{D}_{\text{tok}}$ are two-layer MLPs with a hidden dimension of 768 and a latent space dimension of 20. DisTok is trained for 20K steps on an NVIDIA 4090 GPU with a batch size of 1 and gradient accumulation over $n = 8$ steps (∼30 min). We use a cosine learning rate schedule with an initial rate of 0.0005, and set $\alpha = 1$, $\beta = 1$, $\gamma = 0.001$ and $\tau = 0.85$. Similarity thresholds θ_1 and θ_2 are set to 0.85 and 0.80.

Evaluation Metrics. We evaluate model performance using VQAScore [19], PickScore [14], and ImageReward [37], which assess alignment with text prompts, aesthetic quality, and human preferences. In addition, we employ GPT-4o and conduct a user study to evaluate creativity in terms of conceptual integration, originality, and aesthetics.

5. Results

5.1. Performance on Conditional Exploration

5.1.1. Exploration Guided by Class Distributions

Distribution-Conditional Generation poses a novel challenge in creative synthesis by requiring the generation of visual concepts conditioned on fine-grained semantic distributions over multiple known classes. We evaluate DisTok against state-of-the-art T2I diffusion models, including Stable Diffusion 3 and FLUX.

As shown in Fig. 3, while state-of-the-art diffusion models are effective at generating visually appealing images, they struggle with creative synthesis when conditioned on class distributions. In the absence of explicit mechanisms for modeling such distributions, these models often fail to integrate three or more concepts coherently—leading to missing semantic elements, disjoint object representations, or visually fragmented compositions. Moreover, they exhibit low sensitivity to the specified class proportions (e.g., "55% pig"), resulting in semantically unbalanced outputs and limited control over fine-grained concept composition.

In contrast, DisTok employs an encoder-decoder architecture that directly transforms class distributions into tokens of creative concepts (i.e., $\langle t_{\text{crt}} \rangle$), which are subsequently generated by diffusion models. This framework enables the generation of semantically coherent, visually integrated concepts that faithfully reflect the specified class distributions, supporting fine-grained and controllable synthesis beyond discrete prompt-based generation.

5.1.2. Exploration Guided by Class Pairs

Semantic exploration guided by class pairs, typically formulated as the Text Pair-to-Object (TP2O) task [17], fuses two known concepts into an integral representation and can be viewed as a special case of distribution-conditional generation with a uniform distribution over two classes. DisTok naturally generalizes to this setting and outperforms TP2O-specific methods, including BASS [17] and CreTok [9].

As shown in Fig. 4, DisTok generates visually coherent and semantically integrated hybrids that better reflect both

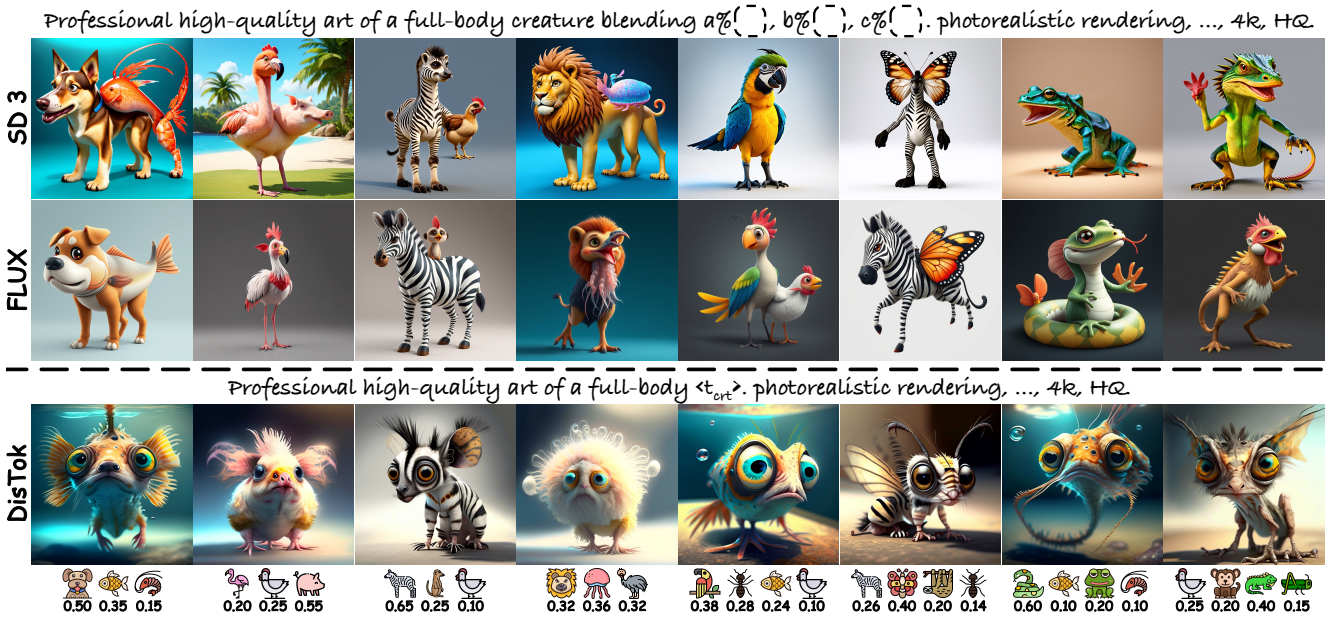

Figure 3. Performance of DisTok on Semantic Exploration Guided by Class Distributions.

Figure 4. Performance of DisTok on Semantic Exploration Guided by Class Pairs.

input classes. Compared to BASS, which requires costly iterative sampling and candidate filtering (∼40s per concept), DisTok directly encodes a concept pair and decodes a creative token in a single pass, incurring negligible additional overhead over standard diffusion inference (∼3s).

While CreTok successfully captures the semantics of combinatorial creativity through a shared token, enabling efficiency and flexibility, its fixed fusion mechanism constrains output diversity. As shown by its near-identical outputs for `(Lion, Snake)` and `(Lion, Shrimp)`, CreTok struggles to distinguish between similar concept pairs, reflecting limitations inherent to prompt-level fusion. In contrast, DisTok generates distinct and semantically coherent hybrids for each pair, demonstrating stronger compositional sensitivity and greater expressiveness. Thus, by modeling combinatorial creativity in latent space and aligning generated concepts with human-preferred aesthetics, DisTok enables efficient, diverse, and coherent creative generation.

5.2. Performance on Unconditional Exploration

DisTok provides a distinctive advantage over prior methods such as CreTok [9] and BASS [17] by enabling reference-free creative generation. As shown in Fig. 5a, CreTok relies on explicit concept pairs and fails to produce diverse outputs without such guidance. ConceptLab [30], enables unconstrained concept generation through iterative token optimization to deviate from known classes, but suffers from high computational cost (∼120s per concept) due to its gradient-based search process.

In contrast, DisTok employs an encoder-decoder frame-

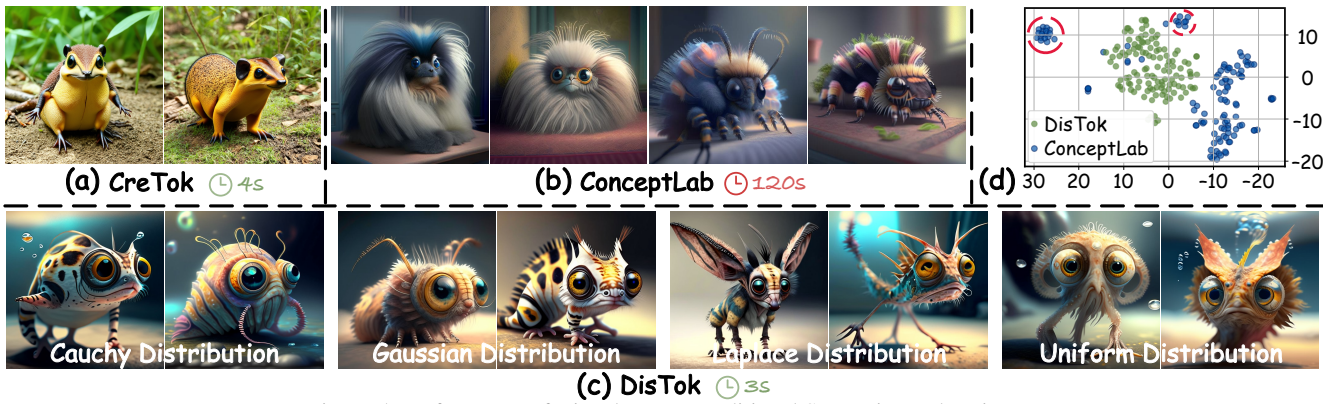

Figure 5. Performance of DisTok on Unconditional Semantic Exploration.

Figure 6. Tokens generated by DisTok can be seamlessly combined with natural language prompts to support diverse styles while preserving conceptual consistency. See Appendix A.4 for more styles.

Table 1. Quantitative Comparisons for Image-Text Alignment and Human Preference Ratings.

	Class Pair			Class Distri.
	BASS	CreTok	DisTok	DisTok
VQAScore↑ [19]	0.667	0.695	**0.840**	**0.734**
PickScore↑ [14]	21.67	21.97	**22.33**	**21.23**
ImageReward↑ [37]	0.387	1.018	**1.168**	**0.661**

work that enables direct sampling of latent vectors and decoding into creative tokens without reliance on external prompts. This design enables efficient exploration of novel and diverse concepts beyond simple recombinations of existing ones, while eliminating iterative search and mitigating the limitations of prompt-dependent methods.

More importantly, ConceptLab [30] remains semantically constrained, as each optimization step merely pushes tokens away from limited known classes, resulting in sparse, repetitive clusters (red circle in Fig. 5d) and restricting diversity and novelty, particularly in large-scale generation. In contrast, DisTok leverages a structured latent space regularized during training (Sec. 3.3), enabling ***direct sampling*** from any zero-mean, unit-variance distribution (e.g., Gaussian, Laplace, Cauchy) without iterative optimization. As shown in Fig. 5c, intra-distribution sampling yields substantial diversity, while inter-distribution sampling further enhances variability, demonstrating DisTok's effectiveness as a unified framework for open-ended semantic exploration.

5.3. Universal Style Adaptability of DisTok

DisTok demonstrates strong universality in style-aware creative generation by directly producing creative tokens that serve as reusable semantic anchors, enabling seamless style transfer via natural language prompts without retraining. As shown in Fig. 6, DisTok consistently preserves conceptual identity across diverse styles. In contrast, BASS [17] and other diffusion-based exploration methods [11, 20], which sample and filter images without token representations, lack the flexibility to adapt concepts via textual conditioning.

Table 2. Creativity evaluated by GPT-4o.

		Inte.	Align.	Orig.	Aesth.	*Compr.*
Distribution	SD 3	6.3±1.8	5.7±3.4	5.8±4.8	8.0±0.5	*6.5±2.0*
	SD 3.5	6.7±1.3	6.1±2.3	6.4±2.3	8.2±0.4	*6.9±1.2*
	Kandinsky	7.7±1.5	7.4±2.6	7.3±2.2	8.6±0.4	*7.8±1.4*
	FLUX	8.0±0.9	7.7±1.5	8.0±1.4	8.8±0.2	*8.1±0.9*
	DisTok	**9.2±0.2**	**9.2±0.2**	**9.8±0.1**	**9.9±0.1**	***9.5±0.1***
Pair	BASS [17]	5.7±2.0	5.5±4.6	5.6±4.5	8.2±0.3	*6.3±2.1*
	CreTok [9]	7.3±2.3	7.7±5.3	7.1±3.0	8.8±0.2	*7.7±1.9*
	DisTok	**9.3±0.3**	**9.9±0.1**	**9.2±0.2**	**9.1±0.3**	***9.4±0.1***

5.4. Evaluation for Creativity

Quantitative Comparisons. We evaluate DisTok on image-text alignment and on human preference under both text-pair and class-distribution conditions (Table 1). DisTok consistently outperforms BASS [17] and CreTok [9] across all metrics on TP2O tasks, demonstrating its ability to synthesize semantically faithful and aesthetically preferred creative images. Although existing metrics may underestimate outputs with strong out-of-distribution characteristics, VQAScore still highlights DisTok's strength in capturing fine-grained conditions (e.g., class distributions).

Evaluation via GPT-4o. Given the limitations of existing metrics in capturing out-of-distribution creativity, we adopt GPT-4o for assessments (prompts in Appendix B.1). Table 2 shows that DisTok outperforms all SOTA diffusion models on distribution-conditional generation task, with notable

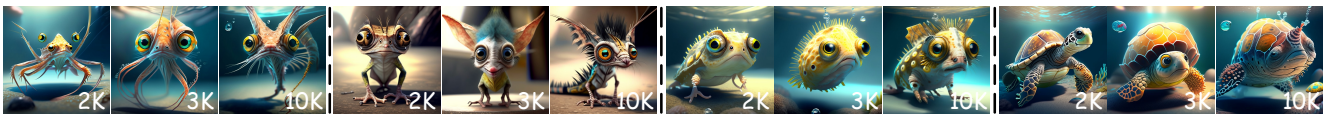

Figure 7. Progressively complex creative concepts sampled at 2K, 3K, and 10K training steps.

Figure 8. Comparison with Prompt Engineering. For each DisTok-generated concept, a corresponding textual prompt for Kandinsky [29] is derived using GPT-4o. All detailed prompts are provided in Fig. 14.

Table 3. Results of the user study.

	Class Distribution				Class Pair		Uncondition
	SD 3	SD 3.5	Kandinsky	FLUX	BASS	CreTok	ConceptLab
DisTok	**312**:188	**320**:180	**355**:145	**278**:222	**396**:104	**330**:170	**308**:192

gains in originality and aesthetics, highlighting its capacity to synthesize novel and creative concepts.

User Study. We conduct a user study with 100 participants from diverse creative domains and higher education backgrounds (see Appendix B.2 for details). Each participant evaluates 50 image pairs generated by DisTok and baseline methods across three tasks, selecting the image they find more creative. As shown in Table 3, DisTok consistently receives the majority of votes, outperforming both SOTA T2I models and creativity-focused baselines. These results highlight DisTok's ability to synthesize semantically integrated, visually coherent, and creatively compelling concepts.

5.5. Ablation and Analysis

Increasingly Complex Semantic through Concept Fusion. DisTok decode creative concepts with increasingly complex class distributions by repeatedly sampling concept pairs from the Concept Pool and applying semantic fusion (Sec. 3.2). To illustrate this progression, we visualize tokens decoded at different training stages. As shown in Fig. 7, the resulting images exhibit a clear trajectory from simple compositions to semantically rich and entangled concepts, demonstrating DisTok's capacity to model fine-grained compositional variation within its latent space.

Distribution Consistency between Language and Visual Semantics. DisTok promotes alignment between input class distributions and the semantic content of generated images through distribution consistency enforcement (Sec. 3.4). We evaluate its impact via an ablation study by removing this supervision and computing the KL divergence between the input token distribution and the visual distribution predicted by BLIP. As shown in Table 4, removing this objective significantly increases divergence, highlighting the importance of distribution consistency enforcement in maintaining fine-grained semantic consistency during generation.

Table 4. Ablation of Distribution Consistency Supervision.

	w/o Cons.	DisTok
KL↓	0.0732	**0.0602**

Creativity Beyond Linguistic Expression. Prompt engineering guides diffusion models toward creative outputs via handcrafted natural language prompts [1, 26]. However, it presupposes that users have already conceptualized the desired idea, shifting the burden of creativity to humans and limiting machine intelligence. Nevertheless, we still compare DisTok with prompt engineering to illustrate DisTok's generative capacity beyond linguistic expression.

For each image generated by DisTok, we extract a detailed description using GPT-4o and reuse it as a prompt in standard diffusion models (Fig. 8). Despite their linguistic richness, these prompts often yield images with compositional artifacts, or diminished semantic coherence—underscoring representational limits of natural language. In contrast, DisTok directly generates semantically rich, linguistically inexpressible concepts without human intervention.

6. Conclusion

Inspired by classifiers' tendency to produce soft probability distributions over known classes for out-of-distribution inputs, we propose Distribution-Conditional Generation, a paradigm that models creativity as image synthesis conditioned on class distributions, enabling controllable yet semantically unconstrained exploration for creative generation. Building on this, we present DisTok, an encoder–decoder framework that maps class distributions into tokens of creative concepts, unifying conditional and unconditional generation for both controllable and open-ended exploration. Extensive experiments show that DisTok achieves state-of-the-art performance in creative generation with enhanced aesthetics, originality and semantic coherence.

Acknowledgement

We sincerely appreciate Freepik for contributing to the figure design. This research was supported by the Jiangsu Science Foundation (BG2024036, BK20243012), the National Natural Science Foundation of China (625B2045, 62125602, U24A20324, 92464301, 62306073), the New Cornerstone Science Foundation through the XPLORER PRIZE, the Fundamental Research Funds for the Central Universities (2242025K30024), and SEU Innovation Capability Enhancement Plan for Doctoral Students (CXJH_SEU 26023).

References

[1] Shm Garanganao Almeda, JD Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. Prompting for discovery: Flexible sense-making for ai art-making with dreamsheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24)*, pages 1–17, 2024. 8

[2] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 8619–8628, 2024. 3

[3] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the International Conference on Machine Learning (ICML'23)*, pages 4672–4712, 2023. 1

[4] Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 7727–7736, 2024. 1

[5] Payel Das and Lav R Varshney. Explaining artificial intelligence generation and creativity: Human interpretability for novel ideas and artifacts. *IEEE Signal Processing Magazine*, 39(4):85–95, 2022. 2

[6] Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH'25)*, pages 1–11, 2025. 3

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML'24)*, pages 12606–12633, 2024. 1

[8] Fu Feng, Jing Wang, Xu Yang, and Xin Geng. Learngene: Inheritable "genes" in intelligent agents. *Artificial Intelligence*, page 104421, 2025. 3

[9] Fu Feng, Yucheng Xie, Xu Yang, Jing Wang, and Xin Geng. Redefining <creative> in dictionary: Towards an enhanced semantic understanding of creative generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25)*, pages 18444–18454, 2025. 2, 3, 5, 6, 7

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the International Conference on Learning Representations (ICLR'23)*, 2023. 3

[11] Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, and Jaesik Choi. Enhancing creative generation on stable diffusion-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25)*, pages 28609–28618, 2025. 2, 7

[12] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 7323–7334, 2023. 3

[13] Minghui Hu, Jianbin Zheng, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. One more step: A versatile plug-and-play module for rectifying diffusion schedule flaws and enhancing low-frequency controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 7331–7340, 2024. 2

[14] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS'23)*, pages 36652–36663, 2023. 5, 7

[15] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 1931–1941, 2023. 3

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML'22)*, pages 12888–12900, 2022. 4

[17] Jun Li, Zedong Zhang, and Jian Yang. Tp2o: Creative text pair-to-object generation using balance swap-sampling. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, pages 92–111, 2024. 2, 3, 5, 6, 7, 1

[18] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 2, 3

[19] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, pages 366–384, 2024. 5, 7

[20] Jack Lu, Ryan Teehan, and Mengye Ren. Procreate, don't reproduce! propulsive energy diffusion for creative generation. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, pages 397–414, 2024. 2, 7

[21] Deborah Mateja and Armin Heinzl. Towards machine learning as an enabler of computational creativity. *IEEE Transactions on Artificial Intelligence*, 2(6):460–475, 2021. 2

[22] Marian Mazzone and Ahmed Elgammal. Art, creativity, and the potential of artificial intelligence. *Arts*, 8(1):26, 2019. 2

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, pages 4296–4304, 2024. 2

[24] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, pages 420–437, 2024. 3

[25] Kam Woh Ng, Jing Yang, Jia Wei Sii, Jiankang Deng, Chee Seng Chan, Yi-Zhe Song, Tao Xiang, and Xiatian Zhu. Chirpy3d: Creative fine-grained 3d object fabrication via part sampling. *arXiv preprint arXiv:2501.04144*, 2025. 3, 4

[26] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. Prompting ai art: An investigation into the creative skill of prompt engineering. *International Journal of Human–Computer Interaction*, 41(16):10207–10229, 2025. 8

[27] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 27080–27090, 2024. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML'21)*, pages 8748–8763, 2021. 3, 5

[29] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'23)*, pages 286–295, 2023. 5, 8

[30] Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3):1–14, 2024. 2, 3, 6, 7

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 22500–22510, 2023. 3

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 6527–6536, 2024. 3

[33] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang "Atlas" Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS'23)*, pages 72137–72154, 2023. 1

[34] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Multimodal large language models make text-to-image generative models align better. In *Advances in Neural Information Processing Systems (NeurIPS'24)*, pages 81287–81323, 2024. 1

[35] Yucheng Xie, Fu Feng, Ruixiao Shi, Jing Wang, Yong Rui, and Xin Geng. Kind: Knowledge integration and diversion for training decomposable models. In *Proceedings of the International Conference on Machine Learning (ICML'25)*, pages 68626–68645, 2025. 1

[36] Zeren Xiong, Zedong Zhang, Zikun Chen, Shuo Chen, Xiang Li, Gan Sun, Jian Yang, and Jun Li. Novel object synthesis via adaptive text-image harmony. In *Advances in Neural Information Processing Systems*, pages 139085–139113, 2024. 2, 3

[37] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS'23)*, pages 15903–15935, 2023. 5, 7

[38] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[39] Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, and Qianru Sun. Few-shot learner parameterization by diffusion time-steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 23263–23272, 2024. 2

[40] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 7912–7921, 2024. 2, 3

[41] Zedong Zhang, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. Agswap: Overcoming category boundaries in object fusion via adaptive group swapping. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SIGGRAPH Asia'25)*, pages 1–12, 2025. 2

[42] Ming Zhong, yelong shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-loRA composition for image generation. *Transactions on Machine Learning Research*, 2024. 3

[43] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 13246–13257, 2024. 2