Redefining `<Creative>` in Dictionary: Towards an Enhanced Semantic Understanding of Creative Generation

Fu Feng[1,2] Yucheng Xie[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, xuyang_palm, wangjing91, xgeng}@seu.edu.cn

Abstract

"Creative" remains an inherently abstract concept for both humans and diffusion models. While text-to-image (T2I) diffusion models can easily generate out-of-distribution concepts like "a blue banana", they struggle with generating combinatorial objects such as "a creative mixture that resembles a lettuce and a mantis", due to difficulties in understanding the semantic depth of "creative". Current methods rely heavily on synthesizing reference prompts or images to achieve a creative effect, typically requiring retraining for each unique creative output—a process that is computationally intensive and limits practical applications. To address this, we introduce CreTok, which brings meta-creativity to diffusion models by redefining "creative" as a new token, `<CreTok>`, thus enhancing models' semantic understanding for combinatorial creativity. CreTok achieves such redefinition by iteratively sampling diverse text pairs from our proposed CangJie dataset to form adaptive prompts and restrictive prompts, and then optimizing the similarity between their respective text embeddings. Extensive experiments demonstrate that `<CreTok>` enables the universal and direct generation of combinatorial creativity across diverse concepts without additional training, achieving state-of-the-art performance with improved text-image alignment and higher human preference ratings. Code will be made available at https://github.com/fu-feng/CreTok.

1. Introduction

"Creativity is the power to connect the seemingly unconnected."
— William Plomer

Recent advancements have witnessed the impressive capabilities of diffusion models, such as DALL-E 3 [42], Stable Diffusion 3 [10], and Midjourney [31], which can now generate images comparable to those created by human artists [2, 38, 58]. The high-quality image generation results from models' strength of capturing complex data distributions [5, 24, 51]. However, such strength also leads diffusion models to replicate patterns from their training data, limiting their potential for genuine creativity [8, 61].

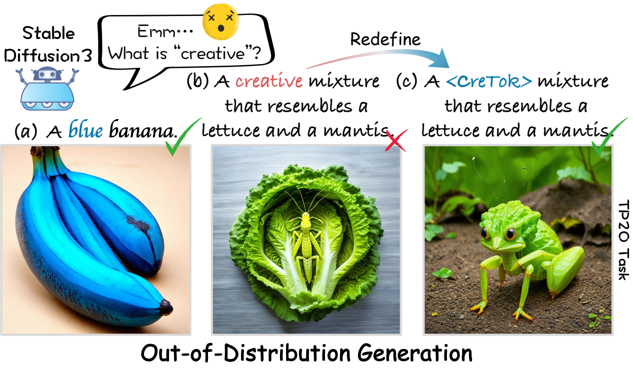
Stable Diffusion3
Emm···
What is "creative"?
Redefine
(a) A blue banana.
(b) A creative mixture that resembles a lettuce and a mantis.
(c) A <CreTok> mixture that resembles a lettuce and a mantis.
TP2O Task
Out-of-Distribution Generation

Figure 1. (a) In out-of-distribution generation, diffusion models can ***directly*** generate "a blue banana" without additional training, benefiting from the clear and concrete semantics of "blue". (b) However, they lack an intrinsic understanding of the abstract and ambiguous semantics of "creative". (c) Leveraging the TP2O (i.e., Creative Text Pair to Object) task, we redefine the token associated with "creative" as `<CreTok>` to bring models meta-creativity, allowing them to ***directly*** generate combinatorial creativity by enhancing their semantic understanding of "creative".

Creativity, an inherently abstract concept even for humans [28, 48], presents a significant challenge for diffusion models. Recent efforts have sought to integrate creativity into diffusion models in more concrete ways. For instance, ConceptLab [44] approaches creativity as the generation of novel, indescribable concepts beyond conventional language. BASS [23] extends this interpretation, defining creativity as the ***combinatorial generation*** of unique objects from text pairs (e.g., `(Lettuce, Mantis)`), exemplified in Figure 1c. Both approaches suggest that creative ability involves generating out-of-distribution images.

[*] Corresponding authors

Diffusion models can directly generate out-of-distribution images by understanding concrete prompts like “a blue banana”, but often struggle with the abstract semantics of “creative”, as illustrated in Figure 1a,b. Given the challenge that diffusion models face in directly generating creativity, existing methods typically rely on synthesizing reference prompts or images to achieve creative effects. For instance, to combine “Lettuce” and “Mantis” creatively, ConceptLab [44] merges tokens representing these concepts into a new composite token, while BASS [23] uses predefined sampling rules to search for creative outcomes from a large pool of candidate images. Similarly, personalization-based methods like MagicMix [26], DiffMorpher [57], and ATIH [52] use semantic mixing or interpolation during the diffusion process to generate novel visual representations.

However, these methods rely heavily on reference prompts and images, demanding a new training process for each generation, which leads to high computational costs and limited practicality for online applications. In contrast, “a blue banana” can be generated directly without additional training, due to its clear and concrete semantics, especially by the adjective “blue”. Inspired by this, we may ask: *Can we awaken the creativity of diffusion models by enhancing their semantic understanding of “creative”?* To achieve this, we propose **CreTok**, which redefines “creative” as a new specialized token, `<CreTok>`, allowing it to function similarly to “blue” in “a blue banana”. This redefinition enhances the model’s semantic understanding for combinatorial creativity, as shown in Figure 1c.

Unlike traditional token-based personalization methods, such as textual inversion [11, 46, 50] and ConceptLab [44], which assign a unique token to each static novel concept, CreTok introduces `<CreTok>` as a ***universal*** “adjective” applicable across all creative concept generation. Specifically, CreTok builds on the definition of “creativity” from the TP2O task [23] for combinatorial object generation, and refines this concept for meta-creativity in an image-free manner on our proposed dataset of text pairs, termed *CangJie*[1] for the learning of `<CreTok>`. In each training step, a text pair (t_1, t_2) is randomly sampled to generate creative outputs by optimizing the similarity between the text embedding of a restrictive prompt (e.g., *“A t_1 t_2”*) and an adaptive prompt (e.g., *“A photo of a `<CreTok>` mixture*”). This process enhances the semantic understanding of `<CreTok>` for concept-combinatorial creativity beyond the literal meanings of t_1 and t_2.

Through this approach, CreTok establishes `<CreTok>` as a universal token that brings ***meta-creativity*** to diffusion models, transforming creativity from static concept synthesis [23, 44, 52] to a more adaptable and flexible creative capability. This meta-creativity enables the model to generate novel combinatorial concepts, even when the corresponding text pairs have not been encountered during training. For instance, the combination of `(Lettuce, Mantis)` in Figure 1c, though unseen during training, can be creatively generated using `<CreTok>`. Furthermore, this meta-creativity enables direct concept combinations without requiring additional training, much like generating “a blue banana”. This significantly reduces both time and computational complexity compared to state-of-the-art (SOTA) creative generation methods, such as ConceptLab [44] (4s vs. 120s per image, **30× speedup**) and BASS [23] (4s vs. 40s per image, **10× speedup**), while maintaining linguistic flexibility for diverse applications and styles.

Notably, images generated by CreTok achieve higher human preference ratings (↑0.009 in PickScore [20] and ↑0.169 in ImageReward [53]) and better text-image alignment (↑0.03 in VQAScore [52]) compared to SOTA diffusion models, such as Stable Diffusion 3.5 [36]. Further evaluations using GPT-4o [1] and user studies indicate superior performance of CreTok in terms of integration, originality, and aesthetics, underscoring its effectiveness in fostering combinatorial creativity.

Our contributions are as follows: (1) We propose CreTok, a method designed to enhance models’ meta-ability by enabling an enhanced understanding of abstract and ambiguous adjectives (e.g., “creative” or “beautiful”) through their redefinition as new tokens. (2) Leveraging CreTok, we redefine the abstract term “creative” within our proposed *CangJie* dataset for the TP2O task, and introduce `<CreTok>`, a universal token that imparts meta-creativity to diffusion models, enabling direct application to the creative generation of diverse combinatorial concepts. (3) Experimental results demonstrate the effectiveness of CreTok in generating combinatorial creativity, outperforming SOTA text-to-image (T2I) models and creative generation methods in terms of computational complexity, human preference ratings, text-image alignment, and other key metrics.

2. Related Work

2.1. Creative Generation

Advancing machine intelligence necessitates models with human-like creativity, a critical yet underexplored aspect of AI research [29, 30]. Early approaches to creativity involves heuristic search methods [7, 54]. With the rise of image generation, interest in exploring creativity expands [9, 16, 34], particularly within Generative Adversarial Networks [12] and Variational Autoencoders [19].

More recently, text-to-image (T2I) models [15, 25] have incorporated tasks specifically targeting creativity. ConceptLab [44] introduces Creative Text to Image Generation (CT2I) task, which focuses on generating novel visual concepts beyond conventional language description. In con-

[1] CangJie comes from 仓颉, the creator of Chinese characters.

trast, BASS [23] proposes the Creative Text Pair to Object (TP2O) task, which combines attributes of existing concepts into new compositions. Compared to the open-ended nature of CT2I, TP2O offers more controlled, user-aligned creativity. Parallel advancements in creative text generation [49, 61] further emphasize the significance of creativity across modalities.

In this work, we enhance the creativity of diffusion models—particularly in the TP2O task—by refining their semantic understanding of "creative" through a redefined token, `<CreTok>`. `<CreTok>` brings meta-creativity to diffusion models, transforming it into a universal token for expressing "creative" and enabling the combination of diverse concepts without additional training.

2.2. Personalized Visual Content Generation

Personalization aims to generate diverse images of specific concepts from limited reference images [39, 47]. Foundational approaches, such as Textual Inversion [11] and DreamBooth [46], optimize text embeddings to capture unique visual concepts as new tokens. Building on these, recent methods employ compositional techniques for creative recombination of visual elements. For instance, Concept Decomposition [50] breaks personalized concepts into distinct visual aspects captured by specific tokens, while PartCraft [33] deconstructs images into modular, fine-grained components for selective reassembly.

Beyond text embedding optimization, advanced methods like SVDiff [14] and others [4, 22, 59] enhance model adaptability through targeted network tuning. Techniques such as MagicMix [26], DiffMorpher [57], and ATIH [52] use semantic mixing or interpolation [60] during the diffusion process to create innovative visual representations.

In this context, `<CreTok>` functions not as a representation of a specific concept but as a descriptor of *how to generate creativity*, imparting meta-creativity to models and establishing `<CreTok>` as a universally adaptable token.

3. Methods

CreTok arouses the creativity of diffusion models by enhancing their semantic understanding of "creative" and redefining it as a new token, `<CreTok>`. This section first presents the basic principles of T2I models, followed by a detailed method for combining text pairs into novel concepts. Finally, we describe the iterative process to continually refine `<CreTok>` for enhanced creative expression.

3.1. Preliminary

Latent Diffusion Models (LDMs) [45] have been widely adopted in T2I generation [3, 13, 17, 18, 32, 37, 55, 56]. LDMs shift the diffusion process from pixel space to a compact latent space with an encoder $\mathcal{E}$, which maps images x into spatial latent codes $z = \mathcal{E}(x)$. The diffusion model is then trained to generate these latent codes through denoising, minimizing the following objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}[||\varepsilon - \varepsilon_\theta(z_t, t, c_\theta(y))||_2^2] \quad (1)$$

where ε_θ represents the noise prediction network, trained to estimate the noise ε added to the latent variable z_t at timestep t, conditioned on $c_\theta(y)$, a vector derived from the input y (e.g., text prompt) through a mapping function c_θ.

In our study, we utilize Stable Diffusion 3 (SD 3) [10] as the base model, which employs a transformer-based architecture to facilitate bidirectional information flow between image and text tokens. Our work focuses on the text encoder in c_θ and condition input y, while keeping the parameters of other components frozen (See Appendix A for details).

3.2. Creative Generation from a Single Text Pair

To redefine "creative" as a universally applicable token, `<CreTok>`, for the combinatorial generation of various text pairs, we begin by performing token-based concept fusion with a single text pair. Building on ConceptLab [44], we achieve such fusion by increasing the semantic similarity of two distinct prompts in the embedding space.

As shown in Figure 2, given a text pair (t_1, t_2) (e.g., `(Lettuce, Mantis)`), we generate a restrictive prompt $\mathcal{P}_r(t_1, t_2)$ by combining the pair into a phrase like "a t_1 t_2" (e.g., "a lettuce mantis."). The **trainable** token `<CreTok>`, which redefines "creative", is then used to form an adaptive prompt $\mathcal{P}_a$ representing the combinatorial results (e.g., "a photo of a `<CreTok>` mixture."). To optimize the semantic alignment of $\mathcal{P}_r$ and $\mathcal{P}_a$, we increase the similarity between the embeddings of $\mathcal{P}_r$ and $\mathcal{P}_a$ using the following objective:

$$\mathcal{L}_{\text{mix}} = 1 - \cos(E(\mathcal{P}_r(t_1, t_2)), E(\mathcal{P}_a)) \quad (2)$$

where $\cos(a, b) = \frac{a\cdot b}{||a||||b||}$ denotes cosine similarity, and $E(\cdot)$ is the text encoder (e.g., CLIP L/14 [40] in SD 3) that maps prompts to corresponding text embeddings.

As noted in ConceptLab [44], overfitting can artificially inflate similarity by disproportionately reinforcing one concept while neglecting others. To address this issue, we introduce a loss threshold θ to regulate concept integration. Moreover, to ensure the coherent fusion of two concepts, rather than their independent generation (e.g., a lettuce and a mantis), θ must remain moderate to avoid low similarity between $\mathcal{P}_r$ and $\mathcal{P}_a$ (see Section 6.2 for details).

$$\tilde{\mathcal{L}}_{\text{mix}} = 1 - \min[\cos(E(\mathcal{P}_r(t_1, t_2)), E(\mathcal{P}_a)), \theta] \quad (3)$$

Additionally, we observe that the order of t_1 and t_2 in $\mathcal{P}_r$ can bias the model's subject focus. For instance, "a lettuce mantis" may prioritize mantis features with lettuce-like elements. To mitigate this bias, we alternate the positions of the two texts in each pair (i.e., (t_2, t_1)), and compute the loss for both $\mathcal{P}_r(t_1, t_2)$ and $\mathcal{P}_r(t_2, t_1)$ during training, encouraging a balanced fusion of both concepts.

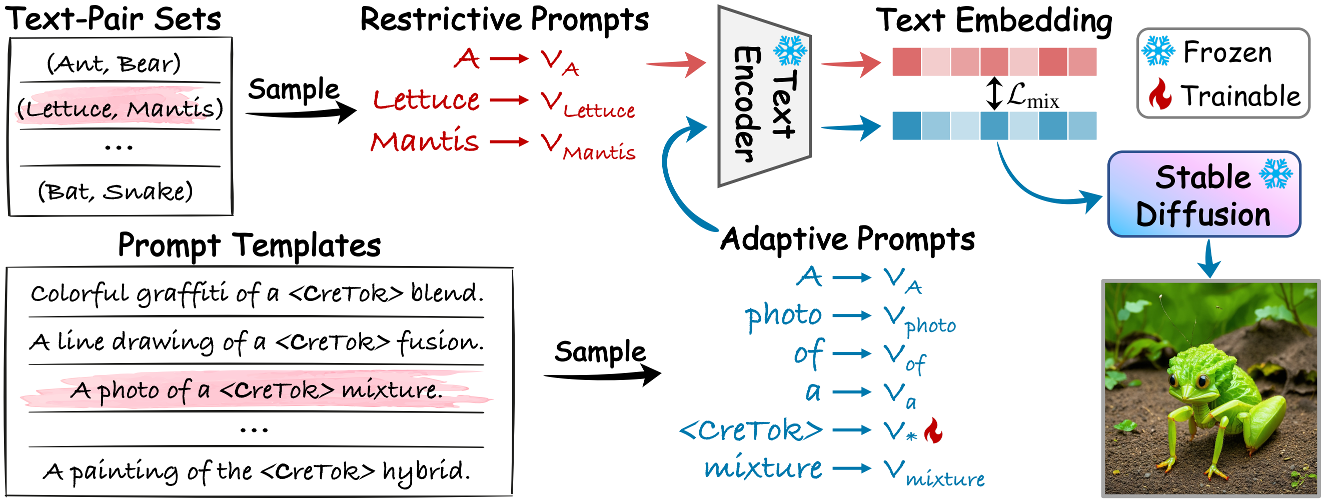

Figure 2. In each training iteration, a text pair and a prompt template are sampled to create a restrictive prompt and an adaptive prompt. The trainable <CreTok> token is then optimized to minimize the cosine similarity between the text embeddings of the adaptive and restrictive prompt. Then the refined adaptive prompt is input into a diffusion model (e.g., Stable Diffusion 3 [10]) for creative image generation.

3.3. Refining <CreTok> in a Continuous Process

Our ultimate objective is not to create a token representing a specific new concept, as done in ConceptLab [44] and other token-based personalization methods [11, 46]. Instead, we aim to enhance the diffusion model's semantic understanding of "creative" through <CreTok>, thus guiding the model on "how to generate creativity". This meta-creativity cannot be achieved through direct optimization on a single text pair, which risks embedding the specific semantics of t_1 and t_2 into <CreTok>.

To achieve such meta-creativity, we construct a dataset specifically for the TP2O task, termed *CangJie*, comprising diverse text pairs (see Appendix D for details). Then we iteratively refine <CreTok> on *CangJie* through a continuous training process, gradually embedding generalized semantics of "creative" into <CreTok>. In each training iteration, a set of n text pairs is randomly sampled, and the cumulative loss is calculated as:

$$\mathcal{L}_{\text{iter}} = \frac{\sum_{i=1}^{n} \tilde{\mathcal{L}}^{i}_{\text{mix}}}{n} \qquad (4)$$

where $\mathcal{L}_{\text{mix}}$ represents the cosine similarity loss between $\mathcal{P}_r$ and $\mathcal{P}_a$, as defined in Eq. (2). After each update, new sets of n text pairs are sampled, ensuring that <CreTok> remains generalizable across a wide range of concepts.

4. Experiments

4.1. Datasets

To comprehensively evaluate creativity, we develop *CangJie*, the first dataset specifically designed for the TP2O task. *CangJie* combines concepts from categories like animals and plants, forming text pairs through diverse combinations. The dataset includes 200 text pairs for training <CreTok>, and 27 text pairs from the original BASS [23] results for unified comparison. Detailed specifications are provided in Appendix D.

4.2. Experimental Setup

Our implementation is based on the official Stable Diffusion 3 [10], which integrates three text encoders: CLIP L/14 [40], OpenCLIP bigG/14 [6], and T5-v1.1-XXL [41]. In our experiments, only the two CLIP models are used as text encoders $E(\cdot)$ without significant performance loss, owing to the simplicity of prompts. The training of <CreTok> runs for 10K steps on a single NVIDIA 4090 GPU, using an initial learning rate of 0.01 with a cosine scheduler, a batch size of 1, and gradient accumulation over $n = 16$ steps. The training process can be completed within approximately 30 minutes. Notably, there is **NO** additional computational overhead after the training of <CreTok>.

4.3. Evaluation Metrics

To evaluate the creativity generated by CreTok and related methods, we first apply VQAScore [27] to measure alignment between the generated image and the text prompt, particularly for combinatorial generation. We also employ PickScore [20] and ImageReward [53] to evaluate alignment with aesthetic standards and human preferences. Additionally, we use GPT-4o [35] and conduct a user study to comprehensively evaluate creativity in terms of conceptual integration, originality, and aesthetic quality. Generation time per image is recorded to highlight deployment considerations, demonstrating CreTok's zero-shot efficiency.

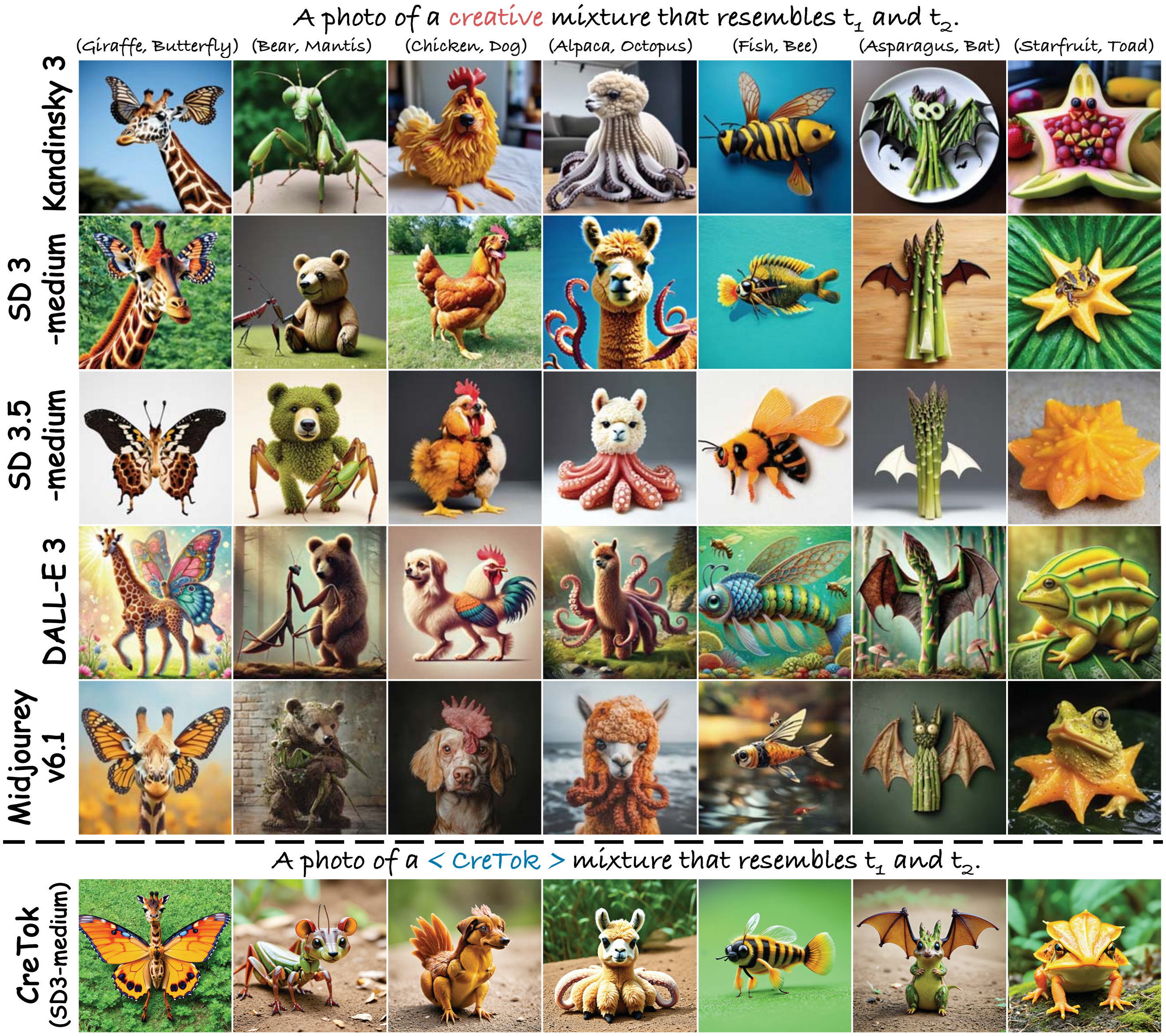

Figure 3. `<CreTok>` enhances diffusion models' semantic understanding of combinatorial creativity. We compare CreTok with SOTA T2I diffusion models including Stable Diffusion 3 [42], Kandinsky 3 [43], Stable Diffusion 3.5 [36], DALL-E 3 [42] and Midjourney v6.1 [31] with identical prompts. CreTok, built on Stable Diffusion 3, replaces "creative" in prompts with the redefined `<CreTok>`.

5. Results

5.1. Performance of Redefined `<CreTok>`

5.1.1. Comparison with State-of-the-Art T2I Models

We evaluate CreTok against state-of-the-art (SOTA) T2I models, including Stable Diffusion 3.5 [36], DALL-E 3 [42] and Midjourney v6.1 [31] under identical prompts, as shown in Figure 3. Despite extensive training on large-scale datasets, SOTA models still struggle to capture the abstract concept of "creative" and struggle to generalize beyond their training distributions, often rendering two objects as separate entities rather than as a cohesive, integrated concept, such as `(Bear, Mantis)`.

Models like DALL-E 3 and Midjourney demonstrate some improvements in combinatorial generation over Stable Diffusion 3, benefiting from advanced architectures and extensive training. However, their outputs often favor an artistic style with vivid colors and intricate details, which contrasts with the realism expected in "photo" prompts, making realistic combinatorial generation a more challenging out-of-distribution (OOD) task.

In contrast, CreTok significantly enhances the model's semantic understanding of "creative" through `<CreTok>`,

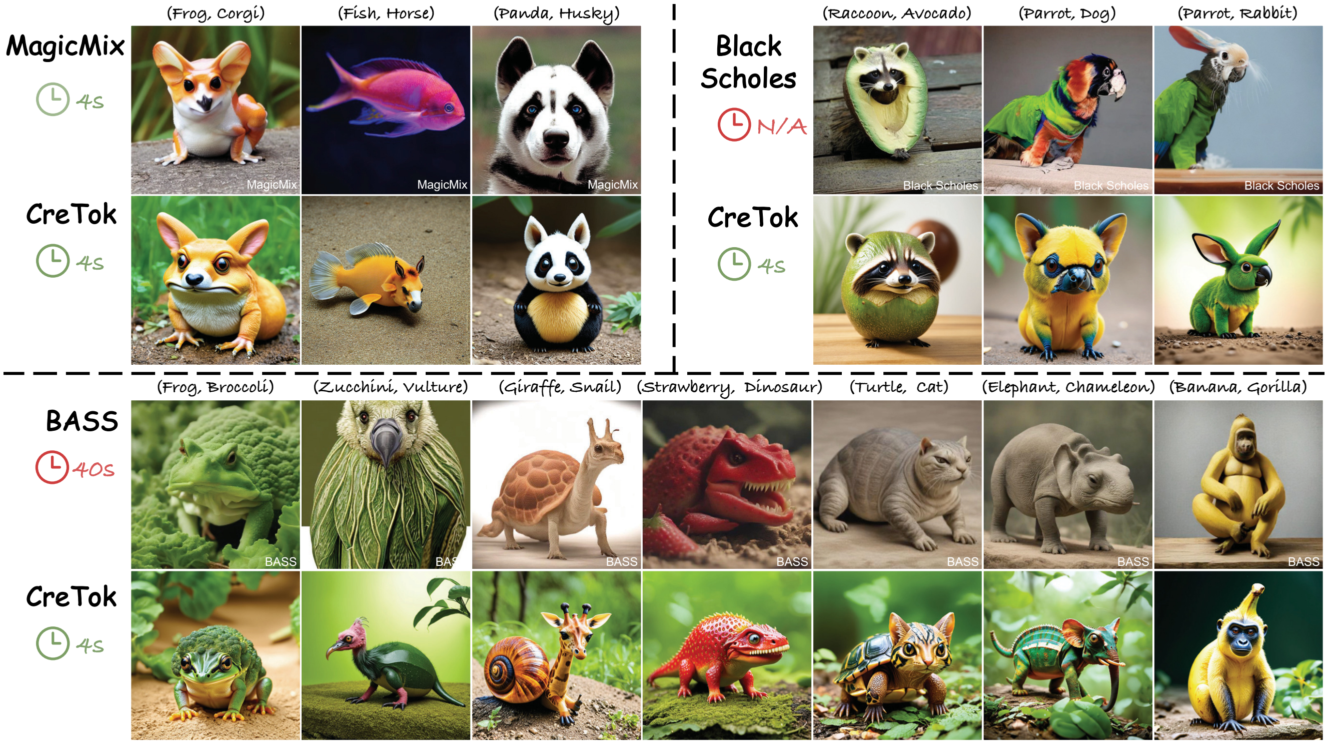

Figure 4. Visual comparisons of combinatorial creativity. We compare CreTok with BASS [23], and other methods achieving similar combinatorial effects, including MagicMix [26] and Black-Scholes [21], to highlight CreTok's superior performance. For fair comparison, most images from these methods are sourced directly from the original papers, with a white watermark added in the bottom right corner. Additionally, generation time per image is recorded to emphasize CreTok's meta-creativity and zero-shot capability.

enabling a more cohesive and realistic fusion while preserving the interpretability of each component. Additional images generated by CreTok are available in Appendix B.2.

5.1.2. Comparison with Creative Generation Methods

Beyond comparisons with SOTA T2I models, we evaluate CreTok against methods specifically designed for creativity and personalization to further highlight its advantages. Creative generation methods like BASS [23] achieve creative outputs through rule-based searches across large-scale candidate generations, while personalization methods, such as MagicMix [26] and Black Scholes [21], can generate combination effects via interpolation between noise, prompts.

As shown in Figure 4, interpolation techniques used in personalization can produce similar visual effects, but they are heavily dependent on reference images, limiting their further adaptability. When significant visual disparities exist between images, the resulting fusion often lacks coherence (see Appendix B.1 for additional comparisons). While BASS is capable of producing high-quality creative images without reference images, it demands substantial computational resources (40 seconds vs. CreTok's 4 seconds) for generation and filtering.

CreTok addresses limitations by introducing a universal token, <CreTok>, specifically for combinatorial creativity, which directly redefines "creative" for meta-creativity rather than merely synthesizing reference images. This allows <CreTok> to seamlessly integrate with other tokens, facilitating novel concept generation without additional training.

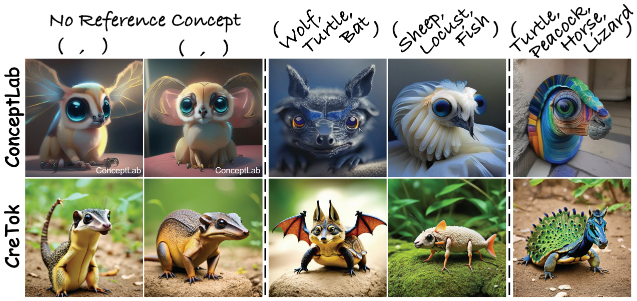

Figure 5. Combinatorial creativity with no concepts or two more concepts. Images with white watermarks are directly sourced from the original paper of the comparison method.

5.2. Creativity beyond Text Pair

Beyond text-pair-based TP2O tasks, we extend our exploration to the Creative Text-to-Image (CT2I) task, as defined in ConceptLab [44], which allows for the fusion of three or more concepts or the creation of entirely new concepts without referencing existing ones.

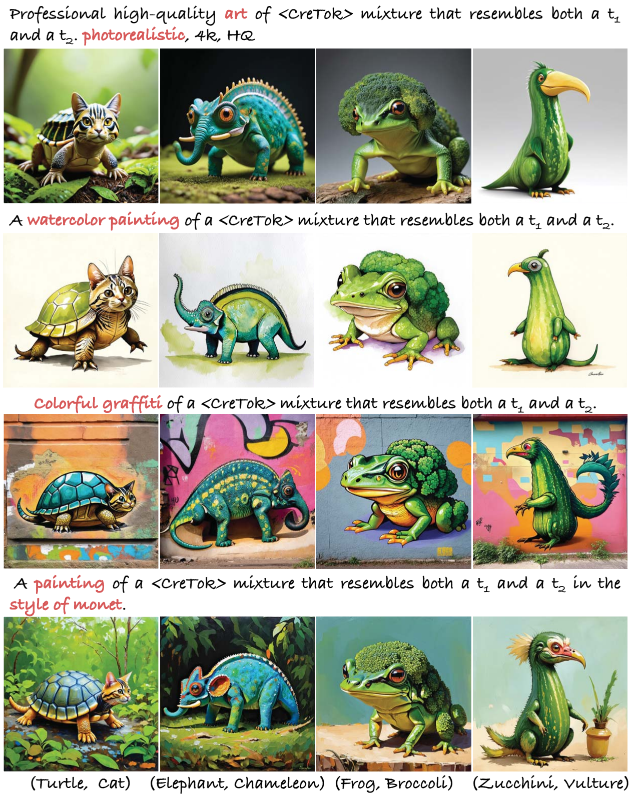
Professional high-quality art of <CreTok> mixture that resembles both a t_1 and a t_2. photorealistic, 4k, HQ
A watercolor painting of a <CreTok> mixture that resembles both a t_1 and a t_2.
Colorful graffiti of a <CreTok> mixture that resembles both a t_1 and a t_2.
A painting of a <CreTok> mixture that resembles both a t_1 and a t_2 in the style of monet.
(Turtle, Cat)
(Elephant, Chameleon)
(Frog, Broccoli)
(Zucchini, Vulture)

Figure 6. Redefined `<CreTok>` can be combined with natural language to showcase combinatorial creativity in various styles. Additional styles are illustrated in Appendix B.3.

Figure 5 compares CreTok and ConceptLab on the CT2I task, showcasing creative images generated from multiple or undefined concepts. Although `<CreTok>` is primarily redefined for combinatorial object generation from text pairs, it extends seamlessly to multi-concept fusion, enabling novel creative outputs without reference text. For instance, CreTok can generate new concepts using prompts like "A photo of a `<CreTok>` mixture." without any predefined concepts. Moreover, when combining multiple concepts (e.g., `(Turtle, Peacock, Horse, Lizard)`), ConceptLab struggles to preserve individual concept features, while CreTok effectively maintains the distinct characteristics of each concept.

While ConceptLab also supports multi-concept generation through token updates, each token is tailored to a specific new concept (e.g., "A photo of `<concept>`"), requiring repeated training for each new creative instance. Moreover, CreTok operates directly in CLIP semantic space [40], without relying on diffusion priors [42], offering a more streamlined framework for creative generation.

5.3. Universality of `<CreTok>` Among Styles

A key limitation of existing methods is their inability to transfer generated creativity across various styles. As previously discussed, `<CreTok>` serves as a universal "adjective", functioning similarly to "blue", allowing it to be seamlessly combined with other prompts for various styles, such as "painting" or "art". Figure 6 presents our results across diverse image styles, highlighting CreTok's unique adaptability—a capability that cannot be achieved by methods like MagicMix [26] and BASS [23].

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

	SD 3	SD 3.5	Kand 3	BASS	CreTok
VQAScore↑	0.793	0.805	0.771	0.710	**0.835**
PickScore↑	21.716	21.766	21.637	20.799	**21.775**
ImageReward↑	0.896	0.881	0.634	0.481	**1.065**

Table 2. Creativity evaluated by GPT-4o.

	Integ.	Align.	Orig.	Aesth.	*Compr.*
SD 3 [10]	8.1±4.1	8.7±4.0	8.2±4.1	9.0±1.3	*8.5±3.1*
Kand 3 [43]	8.9±0.8	9.7±0.3	9.0±0.4	9.2±0.2	*9.2±0.3*
SD 3.5	9.1±0.7	9.9±0.2	9.1±0.6	9.4±0.4	*9.4±0.3*
BASS [23]	8.9±1.3	9.3±1.4	8.7±1.2	8.3±0.7	*8.8±0.9*
CreTok	**9.5±0.4**	**9.9±0.1**	**9.3±0.4**	**9.6±0.3**	***9.6±0.3***

Unlike ConceptLab [44], where each token is tied to a specific concept, `<CreTok>` does not correspond directly to any single concept, yet it consistently maintains adaptability across a wide range of prompts.

5.4. Evaluation for Creativity

5.4.1. Quantitative Comparisons

We conduct quantitative comparisons to evaluate the alignment between images and prompts using VQAScore [27], along with human preference ratings via PickScore [20] and ImageReward [53]. Table 1 presents comparisons between CreTok and SOTA open-source T2I models.

Despite being built upon SD 3, CreTok outperforms both SD 3.5 and Kandinsky 3 in terms of human preference ratings and image-text alignment, even though these models use advanced architectures and extensive training data tailored to human aesthetic preferences.

5.4.2. Evaluation via GPT-4o

Since existing metrics are insufficient for assessing such abstract "creativity", we employ GPT-4o to objectively assess image creativity through quantitative analysis across four dimensions: Integration, Alignment, Originality, and Aesthetics. Detailed prompts are available in Appendix C.1.

Table 2 presents GPT-4o's assessments of creativity for images generated by CreTok compared to other methods. The results indicate that CreTok-generated images demonstrate significant advantages across all evaluated dimensions, especially in the concept integration and originality.

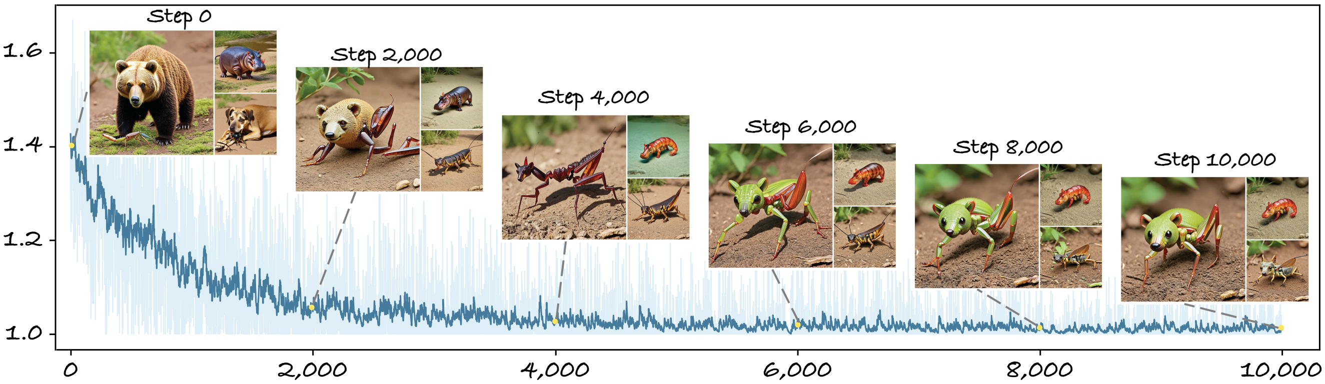

Figure 7. Convergence rate of <CreTok> during the continuous redefinition process, showing training curves and corresponding images.

Table 3. Results of the user study.

	SD 3	SD 3.5	Kand 3	BASS	CreTok
Avg. Rank↓	3.4±1.5	3.1±1.1	3.3±1.4	3.1±1.3	**1.9±1.1**

5.4.3. User Study

To comprehensively evaluate creativity, we conduct a user study involving 50 highly educated participants. Each participant ranks the creativity of images generated by CreTok in comparison with other methods. The average ranks, summarized in Table 3, reveal that CreTok significantly outperforms the current T2I diffusion models lacking specialized design for creativity and receives a higher ranking than BASS [23], achieving an average ranking of 1.9. Further details are provided in Appendix C.2.

6. Ablation and Analysis

6.1. Process of Continual Redefinition

The continual refinement of “creative” within CreTok is illustrated in Figure 7, which captures the convergence process of <CreTok> over time. Additionally, visualizations of randomly selected text pairs, captured every 2,000 training steps, demonstrate the evolving representation.

In early stages, <CreTok> primarily absorbs semantic content from individual concepts, as observed at steps 2,000 and 4,000, where the generated creative output closely resembles one of the concepts (e.g., “Bear” and “Mantis”). However, as training progresses, <CreTok> transitions toward encapsulating a generalized creative representation, independent of specific concepts. This transition is evidenced by increasingly aligned text-image relationships and enhanced image quality, culminating in final convergence.

6.2. Effect of Loss Threshold on Creativity

When optimizing semantic similarity between text embeddings of restrictive and adaptive prompts, improper thresholds can hinder the combinatorial generation of two concepts. To evaluate this, we analyze different thresholds during the refinement of <CreTok>, as shown in Figure 8.

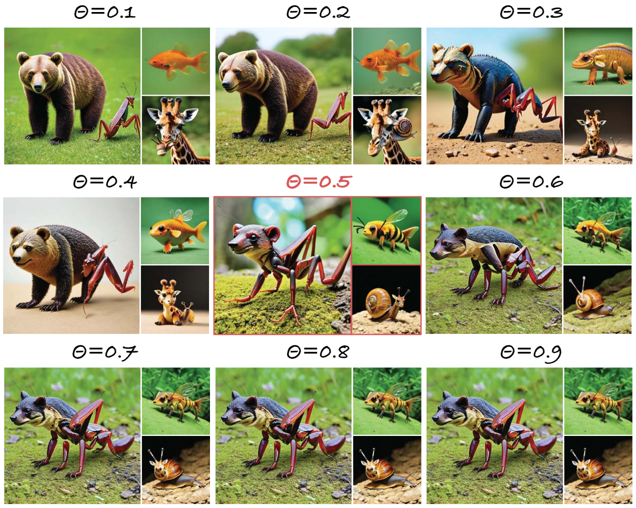

Figure 8. Combinatorial creativity with different threshold θ.

Low semantic similarity results in the two concepts being generated independently, rather than merging into a cohesive combination, as demonstrated by current T2I model in Figure 3. On the other hand, high similarity increases the likelihood of overfitting to one of the concepts. Therefore, we identify an optimal threshold of $\theta = 0.5$, which strikes a balance between capturing semantic representations and promoting combinatorial object generalization.

7. Conclusion

We propose CreTok, a novel approach that imparts meta-creativity to T2I diffusion models by enhancing their semantic understanding of “creative”. CreTok achieves this by redefining “creative” as a universal token, <CreTok>, enabling the model to achieve combinatorial creativity in a zero-shot, image-free manner. Moreover, <CreTok> integrates seamlessly with natural language, facilitating concept combinations across various styles without additional training. Extensive experiments demonstrate that CreTok significantly enhances model creativity, outperforming SOTA T2I models and creative generation methods.

Acknowledgement

We sincerely thank Wenqian Li, Jianlu Shen, and Ruixiao Shi for their insightful discussions on this work. We also appreciate Freepik for contributing to the figure design. This research is supported by the Science and Technology Major Project of Jiangsu Province under Grant BG20240305, Key Program of Jiangsu Science Foundation under Grant BK20243012, National Natural Science Foundation of China under Grants U24A20324, 62125602, and 62306073, Natural Science Foundation of Jiangsu Province under Grant BK20230832, and the Xplorer Prize.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1

[3] Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 821–830, 2024. 3

[4] 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

[5] 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 International Conference on Machine Learning (ICML'23)*, pages 4672–4712, 2023. 1

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 2818–2829, 2023. 4, 12

[7] Daniel Cohen-Or and Hao Zhang. From inspired modeling to creative modeling. *The Visual Computer*, 32:7–14, 2016. 2

[8] 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. 1

[9] Ahmed Elgammal. Can: Creative adversarial networks, generating “art” by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 6:2017, 2017. 2

[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 *Proceedings of International Conference on Machine Learning (ICML'24)*, pages 1–13, 2024. 1, 3, 4, 7

[11] 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. 2, 3, 4

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 7548–7558, 2024. 3

[14] 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

[15] Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*, 2024. 2

[16] Amin Heyrani Nobari, Muhammad Fathy Rashad, and Faez Ahmed. Creativegan: Editing generative adversarial networks for creative design synthesis. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, page V03AT03A002, 2021. 2

[17] Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality control for text-conditional image generation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, pages 1–13, 2023. 3

[18] 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. 3

[19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] 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 *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, pages 36652–36663, 2023. 2, 4, 7

[21] Divya Kothandaraman, Ming Lin, and Dinesh Manocha. Prompt mixing in diffusion models using the black scholes algorithm. *arXiv preprint arXiv:2405.13685*, 2024. 6

[22] 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

[23] 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 1–19, 2024. 1, 2, 3, 4, 6, 7, 8

[24] Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, pages 2097–2127, 2023. 1

[25] Zongrui Li, Minghui Hu, Qian Zheng, and Xudong Jiang. Connecting consistency distillation to score distillation for text-to-3d generation. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, pages 274–291, 2024. 2

[26] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 2, 3, 6, 7, 12, 13

[27] 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. 4, 7

[28] Lars Lindström. Creativity: What is it? can you assess it? can it be taught? *International Journal of Art & Design Education*, 25(1):53–66, 2006. 1

[29] 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

[30] Marian Mazzone and Ahmed Elgammal. Art, creativity, and the potential of artificial intelligence. In *Arts*, page 26, 2019. 2

[31] Midjourney. Midjourney.com. https://www.midjourney.com, 2022. Accessed: 2024-11-14. 1, 5

[32] 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. 3

[33] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. *arXiv preprint arXiv:2407.04604*, 2024. 3

[34] Amin Heyrani Nobari, Wei Chen, and Faez Ahmed. Range-constrained generative adversarial network: Design synthesis under constraints using conditional generative adversarial networks. *Journal of Mechanical Design*, 144(2):021708, 2022. 2

[35] OpenAI. Gpt-4: Openai language model. https://openai.com/research/gpt-4, 2023. Accessed: 2024-11-14. 4

[36] OpenAI. Stable diffusion 3.5. https://github.com/Stability-AI/sd3.5, 2024. Accessed: 2024-11-14. 2, 5

[37] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *arXiv preprint arXiv:2402.05195*, 2024. 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 4195–4205, 2023. 1

[39] 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

[40] 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 *Proceedings of International Conference on Machine Learning (ICML'21)*, pages 8748–8763, 2021. 3, 4, 7, 12

[41] 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. 4, 12

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 5, 7

[43] 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*, pages 286–295, 2023. 5, 7

[44] 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. 1, 2, 3, 4, 6, 7

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pages 10684–10695, 2022. 3

[46] 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. 2, 3, 4

[47] 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

[48] Robert Keith Sawyer and Danah Henriksen. *Explaining creativity: The science of human innovation*. Oxford university press, 2024. 1

[49] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*, 30(5):1–57, 2023. 3

[50] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics*, 42(6):1–13, 2023. 2, 3

[51] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, pages 72137–72154, 2023. 1

[52] 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. *arXiv preprint arXiv:2410.20823*, 2024. 2, 3

[53] 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 *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, pages 15903–15935, 2023. 2, 4, 7

[54] Kai Xu, Hao Zhang, Daniel Cohen-Or, and Baoquan Chen. Fit and diverse: Set evolution for inspiring 3d shape galleries. *ACM Transactions on Graphics*, 31(4):1–10, 2012. 2

[55] Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I Chang, Hanwang Zhang, et al. Exploring diffusion time-steps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*, 2024. 3

[56] 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. 3

[57] 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, 12, 13

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 3836–3847, 2023. 1

[59] Xinxi Zhang, Song Wen, Ligong Han, Felix Juefei-Xu, Akash Srivastava, Junzhou Huang, Hao Wang, Molei Tao, and Dimitris N Metaxas. Spectrum-aware parameter efficient fine-tuning for diffusion models. *arXiv preprint arXiv:2405.21050*, 2024. 3

[60] 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. *arXiv preprint arXiv:2402.16843*, 2024. 3

[61] 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. 1, 3