FreeBlend: Advancing Concept Blending with Staged Feedback-Driven Interpolation Diffusion

Yufan Zhou[1*] **Haoyu Shen**[2*] **Huan Wang**[3]
[1]Harbin Institute of Technology
[2]University of Science and Technology of China
[3]Westlake University
[*]equal contribution

Figure 1: We introduce **FreeBlend**, a training-free approach that effectively blends two concepts to generate new objects through feedback interpolation and auxiliary inference. This method consistently produces visually coherent and harmonious blends, enabling users to create customized images with diverse combinations of concepts.

Abstract

Have you ever imagined a creature combining the features of a cat and a car? In this work, we tackle the problem known as concept blending, which represents a fascinating yet underexplored area in generative models. While recent approaches, such as embedding mixing and latent modification based on structural sketches, have been proposed, they often suffer from incompatible semantic information and discrepancies in shape and appearance. In this work, we introduce FreeBlend, an effective, training-free framework designed to address these challenges. To mitigate cross-modal loss and enhance feature details, we leverage transferred image embeddings as conditional inputs. We also design a stepwise increasing interpolation strategy between latents to seamlessly integrate auxiliary features. Moreover, we introduce a feedback-driven mechanism that updates the auxiliary latents in reverse order, facilitating global blending and preventing unnatural outputs. Extensive experiments demonstrate that our method significantly improves both the semantic coherence and visual quality of blended images.The project homepage is available at `https://petershen-csworld.github.io/FreeBlend/`.

[*]These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

1 Introduction

With the breakneck development of diffusion models, the field of image generation has witnessed rapid progress [1, 2, 3, 4]. These models exhibit strong instruction-following capabilities and high-fidelity synthesis, achieving remarkable performance across a variety of downstream tasks, including local inpainting [5, 6], personalized image generation [7, 8], and more [9, 10, 11].

Such advances have paved the way for a spectrum of novel and creative applications. Among them, concept blending stands out as an interesting direction previously overlooked by the community. For example, one could create a creature combining the features of an orange and a teddy bear, resulting in a surprising yet coherent visual concept. More concept pair blends are shown in Figure 1.

Formally, concept blending involves blending two distinct *text* concepts to create a new one that retains the defining characteristics of its components [12, 13]. It can be viewed as a way of integrating elements from different domains into a novel and meaningful output and creating new objects, scenes, or alterations that are coherent and creatively synthesized. Although current models can generate realistic combinations of individual concepts [14, 15], more sophisticated concept blending techniques are needed to ensure that the generated images not only blend features but also maintain their semantic consistency and visual appeal.

Early methods of concept blending rely on a simple, "black-box" approach. These methods achieve coarse concept blending by manipulating text embeddings from the input side, which typically involves adding or reassembling the embeddings [16, 17]. However, these approaches often lead to inaccurate representations and a lack of correspondence between visual and semantic features due to cross-modal discrepancies. MagicMix [18] interpolates the original class latent into another latent space corresponding to a text prompt. While it creatively introduces latent interpolation for mixing, it struggles with shape mismatch issues and lacks flexibility in visual transformations. ConceptLab [19] utilizes VLMs and latent space manipulation, but the constraints imposed during training the additional module limit the flexibility of its application. ATIH [20], based on MagicMix, injects trainable parameters into the denoising process and enforces similarity constraints to harmonize the fusing of texts and images. However, the limitations of its model structure, similar to those of MagicMix, combined with its inability to address mismatched shapes or semantically irrelevant features, hinder its overall blending performance.

In this paper, we introduce FreeBlend, a training-free method for concept blending in image generation. As shown in Figure 2, FreeBlend smoothly interpolates between auxiliary latents by incorporating a feedback-driven mechanism into the diffusion model's denoising process. As denoising advances, the influence of auxiliary latents diminishes, allowing the blending latent to take greater control, enhancing the blending performance and leveraging the creative power of diffusion models.

Specifically, FreeBlend consists of three core components: transferred unCLIP [21] image conditions for Stable Diffusion [1], a stepwise increasing interpolation strategy, and a feedback-driven mechanism of the denoising process. Instead of using traditional text-based conditions, we employ images generated from text as conditions to guide the generation process via the unCLIP model. This approach reduces the uncertainty caused by cross-modal differences. In addition to this, we divide the denoising process into three stages: the initialization stage, the blending stage, and the refinement stage. At the initialization stage, the pretrained Stable Diffusion model starts with random noise sampled from a standard normal distribution, which is then denoised under the guidance of unCLIP image condition. At the blending stage, we add noise to the auxiliary latents derived from the condition images to ensure to be in the same period as the blending latent. The blending and auxiliary processes are all denoised simultaneously. Ultimately, in the final refinement stage, only the unCLIP image condition is used to provide additional information, enabling the model to generate images with greater clarity and finer details.

The core contributions of this paper are (i) We propose a feedback-driven latent interpolation approach that leverages diffusion models to address the concept blending problem. Our method is designed to be training-free and computationally efficient. (ii) Our approach incorporates unCLIP to use images as conditions, along with a stepwise increasing interpolation strategy and a feedback-driven denoising process to effectively blend different concepts. (iii) We conduct extensive qualitative and quantitative experiments to validate the effectiveness of our proposed method. Multiple evaluation metrics, including our newly introduced blending score metric, CLIP-BS, confirm that our approach achieves state-of-the-art performance in generating blended concepts.

2 Related Work

Image Editing and Training-free Guidance. Generative models like Stable Diffusion [1], Imagen [4], and DALL-E [22] have advanced text-to-image synthesis [2, 23, 24, 25] and diffusion models have significantly advanced image editing [26, 26, 27, 28, 29, 30], image interpolation [31, 32], style transfer [33, 34], and 3D generation [35]. DiffEdit [28] and InstructPix2Pix [29] primarily focus on semantic image editing. DiffEdit leverages mask guidance to facilitate intuitive modifications, while InstructPix2Pix enables efficient editing through natural language instructions. In training-free guidance, Structure Diffusion [36] introduces structured diffusion guidance to tackle compositional challenges, enhancing the handling of multiple objects. FreeDoM [37] is a training-free method using energy-guided conditional diffusion for dynamic editing with conditions like segmentation maps. FreeControl [38] offers zero-shot spatial control over pretrained models. These methods enhance the flexibility and precision of diffusion models for effective, retraining-free image editing. Our work builds on this intuition and introduces a novel method for achieving a unique blending style of image editing through training-free guidance.

Concept Composition and Blending. The field of concept composition has made significant advancements, with the combination of multiple concepts to generate complex, multi-faceted images [14, 15, 39, 40, 41]. Composable Diffusion [14] introduces a compositional approach that utilizes multiple models to generate distinct components of an image, effectively addressing challenges associated with complex object compositions and their interrelationships. Custom Diffusion [41] explores multi-concept customization, enabling the generation of unified, intricate images that seamlessly blend multiple concepts. In contrast, while concept blending holds substantial promise, its practical applications remain more limited and underexplored when compared to compositional approaches. MagicMix [18] addresses semantic mixing while preserving spatial layout, though its use is more restricted due to shape limitations. Melzi et al. [16] and Olearo et al. [17] investigate concept blending by manipulating the relationships between different prompts, with the former focusing on prompt ratios and the latter experimenting with various mechanisms for blending text embeddings. ConceptLab [19] leverages Diffusion Prior models to generate novel concepts within a category through CLIP-based constraints, enabling the creation of unique hybrid concepts. ATIH [20] further advances the field by introducing adaptive text-image harmony, merging text and image inputs to generate novel objects while maintaining their original layout. Although these studies contribute to the domain of concept blending, their overall impact has been more limited compared to concept composition, which has inspired the completion of our work.

3 Method

3.1 Preliminaries

Our method is based on Stable Diffusion, a prominent application of the latent diffusion model [1]. It operates within a compressed latent space, improving sampling and denoising efficiency compared to DDPM [42] and DDIM [43]. A VAE [44] is used to compress image representations, enhancing computational efficiency. During training, at timestep t, a noisy latent vector $\mathbf{z}_t$ is generated by adding noise to the latent representation $\mathbf{z}_0$ (the original clean image latent). This noisy latent is progressively denoised over several timesteps to recover the original image.

Besides, Stable Diffusion employs classifier-free guidance [45] to control the synthesized image content with $\mathbf{c}$ representing given additional conditions like text prompts, images, or other customized properties. The model ϵ_θ could be trained via the following objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{c},\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{1}$$

where ϵ_θ, typically implemented by U-Net [46] or DiT [47], represents the denoiser.

During inference, a latent $\mathbf{z}_t$ is sampled from the standard normal distribution $\mathcal{N}(0, I)$ and the trained denoiser is used to iteratively remove the noise $\epsilon_t = \hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$ in $\mathbf{z}_t$ to produce $\mathbf{z}_0$. This process is expressed as:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) \approx -\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t|\mathbf{c}), \tag{2}$$

where $\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$ represents the noise removal at timestep t, and the gradient of the log marginal distribution provides the direction for noise reduction. In the end, the latent $\mathbf{z}_0$ is passed to the decoder $\mathcal{D}$ to generate the output image $\tilde{\mathbf{x}}$.

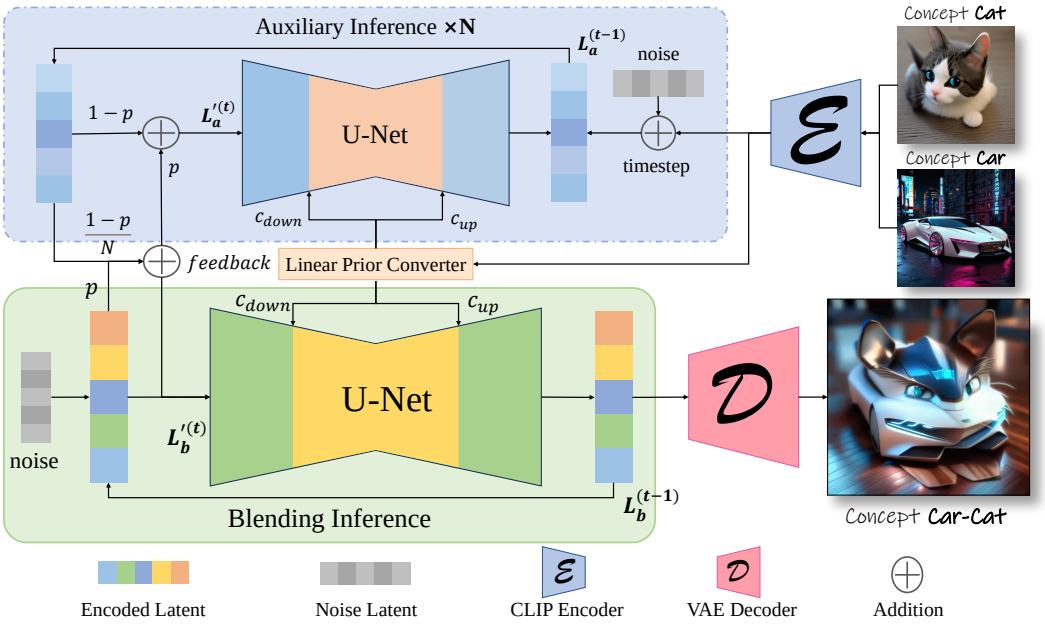

Figure 2: **Overview of our method.** Two input images, generated by Stable Diffusion with respective concepts, are encoded into CLIP embeddings and mapped to a shared text space via the Linear Prior Converter from unCLIP [21]. These embeddings condition the U-Net, one for downsampling and the other for upsampling. The blending latent L_b is initialized with Gaussian noise and processed during initialization. The module within the dashed box is used only in the blending stage. Noise ϵ is added to the image embeddings to generate initial auxiliary latents, which are interpolated into $L_b^{(t)}$ for feedback. The latent $L_a^{(t)}$ is combined with $L_b'^{(t)}$ by proportion p. Updated latents $L_a'^{(t)}$ are refined in auxiliary inference using unCLIP embeddings to preserve original features, while $L_b'^{(t)}$ is denoised in the blending inference. Finally, the blending latent is refined and passed to the VAE decoder to generate the final image.

3.2 Concept Blending with unCLIP Condition

In multimedia production, the innovation of novel entities is crucial, particularly in the development of anthropomorphic characters. A prevalent method for conceptual innovation involves the fusion of distinct ideas. For instance, numerous animated characters are conceived by combining elements from various sources, such as the "Iron Man" suit and a traditional Japanese samurai's armor.

In this section, we present a novel image generation task called concept blending. This task aims to merge two distinct text concepts and generate blended images that retain the original shape, color, texture, and other relevant details. To achieve this task, we first attempt to directly use texts as condition [16, 17], but due to the low quality and visual uncertainty due to the gap of vision and text, we decide to choose images as condition [48, 49, 50].

We adopt the *Linear Prior Converter* from unCLIP [21], which, like DALL-E [22], enables images as conditional inputs. This converter maps image embeddings to the text embedding space, and the resulting embeddings are fed into the U-Net's intermediate layers via cross-attention, guiding the denoising process across all stages. While mapping image embeddings to the text space, the transformed embeddings preserve visual-like information, similar to textual inversion [51], providing more deterministic features and richer details than the original text embeddings.

Our model uses two types of embeddings at different phases: one during the downsampling phase and another during the upsampling phase. Simply averaging the embeddings [16] could lead to the loss of essential details, such as fine textures or spatial positioning. By interacting with the text embedding space through unCLIP, the model captures the semantic essence of the concepts while maintaining precise control over visual attributes like shape and color.

3.3 Staged Denoising Process

To enable more flexible and precise control over the image generation process, and inspired by previous works [18, 52, 53], we propose dividing the generation process into three stages: the initialization stage, the blending stage, and the refinement stage. We introduce two coefficients, t_s and t_e, to represent the start and end timesteps of the blending stage, respectively.

At the initialization stage, the denoising process begins with random noise, $\epsilon \sim \mathcal{N}(0, I)$. In this stage, the blending latent is iteratively updated based on the input unCLIP image embeddings, which define the basic layout of the subject and background. The process starts with $L_b^{(T)} = \epsilon$, and the blending latent is updated according to the following equation:

$$L_b^{(t-1)} = \epsilon_\theta(L_b^{(t)}, t, \varnothing) + w \cdot \left[\epsilon_\theta(L_b^{(t)}, t, \mathbf{c}_{\text{stage}}) - \epsilon_\theta(L_b^{(t)}, t, \varnothing)\right], \tag{3}$$

where w denotes the classifier-free guidance [45] scale, and $\epsilon_\theta(L_b^{(t)}, t, \mathbf{c})$ represents the pre-trained U-Net model with the conditional input $\mathbf{c}$ and timestep t.

In the blending stage, the focus shifts to incorporating original features into the initially formed latent. This process is guided by methods described in Section 3.4 and Section 3.5, which enable the effective blending and filtering of features, ensuring the integration of relevant information.

Finally, at the refinement stage, the general structure of the latent has been established, and the model's focus moves toward enhancing finer details. The objective here is to improve the overall appearance, making the image more natural and cohesive, while addressing any disjointedness or artifacts from earlier stages. This phase continues to refine and improve the final image output, building on the layout established in the initialization stage.

3.4 Stepwise Increasing Interpolation Strategy

In this section, we elaborate on how our method ensures semantic coherence and consistency. We interpolate two image latents with the current denoising latents to encode semantic information from both sources together, which shares similar intuition with MagicMix [18]. The key challenge is to select appropriate blending ratio at each timestep. One issue of using a constant blending ratio is that we may result in blurring and unclear images, dramatically decreasing the generation quality. To address this, for a denoising process with T timesteps, at timestep t, we apply a stepwise declining blending ratio p, which could be expressed as

$$p = 1 - \frac{t}{T}, \tag{4}$$

and the proportion of the k-th auxiliary latent $L_a^{(k)}$ is

$$\lambda = \frac{1-p}{N}, \tag{5}$$

where N is the number of concepts needed to blend and we set it two. Another challenge is the potential for biased or catastrophic forgetting of content. Specifically, the model should account for the varying influence of different concepts. For instance, when subject concepts such as "animals" are blended with background concepts like "plants", the model might give more weight to the subject concepts, resulting in generated images with a higher representation of animals.

To address this, we introduce an additional hyper-parameter, denoted as γ, which regulates the contribution of each image embedding in the latent space. In the case of a blending task involving N images, the interpolation process can be expressed as follows:

$$L_b'^{(t)} = p \cdot L_b^{(t)} + \lambda \cdot \sum_{n=1}^{N} \gamma_n \cdot L_a^{(t,n)}, \tag{6}$$

where γ_n denotes the interpolation strength of n-th image, $L_b^{(t)}$ refers to the blending latent, and $L_a^{(t,n)}$ specifies the n-th auxiliary latent at the t-th timestep.

3.5 Feedback-Driven Mechanism

In terms of synthesis quality, setting the time proportion of the blending stage too high can lead to overlapping issues, particularly when the blending stage is extended too long. This indicates that the interpolation methods used in previous approaches are unable to handle significant changes, resulting in incomplete processing and causing overlapping effects. In contrast, earlier methods [18] terminate the blending stage prematurely, which restricts further blending and fine-tuning. To address this, we propose a feedback-driven mechanism that strikes an optimal balance between these two extremes.

Specifically, $L_a^{(t_s,k)}$ is initialized using the initial auxiliary latent $L_a^{(0,k)}$ as follows:

$$L_a^{(t_s,k)} = \sqrt{\bar{\alpha}_t} L_a^{(0,k)} + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{7}$$

where $L_a^{(0,k)}$ is the embedding encoded from image. Once $L_b^{(t)}$ is interpolated with $L_a^{(t,k)}$, the latter should simultaneously undergo a feedback update using $L_b^{\prime(t)}$ as follows:

$$\begin{aligned} L_a^{\prime(t,k)} &= p \cdot L_b^{\prime(t)} + (1-p) \cdot L_a^{(t,k)} \\ &= \underbrace{(1-p) \cdot L_a^{(t,k)}}_{\text{inherits the original appearance}} + \underbrace{p^2 \cdot L_b^{(t)}}_{\text{integrates into the subject}} + \underbrace{p \cdot (1-p) \cdot \left[\frac{1}{N} \cdot \sum_{n=1}^{N} \gamma_n \cdot L_a^{(t,n)} \right]}_{\text{maintains balance}}. \end{aligned} \tag{8}$$

After interpolation, L_a should also be denoised normally, with $\mathbf{c}_{\text{stage}}$ being replaced by $\mathbf{c}_{\text{up}}$ or $\mathbf{c}_{\text{down}}$ according to the category of it.

As shown in Equation (8), as the timesteps decrease, p increases, which causes L_b to take up a larger proportion of L_a. Consequently, in the later stages of blending, the interpolated L_a increasingly resembles L_b, thus maintaining consistency over time. The average of all previous L_a values contributes to the incorporation of general features, particularly during the middle stages of blending, due to the coefficient $p \cdot (1-p)$. Regarding $L_a^{(t,k)}$, it gradually declines in the final stages, thereby reducing the proportion of the original features. Through this approach, the auxiliary latents are updated by incorporating both the preceding auxiliary latents and the main blending latent.

4 Experimental Results

4.1 Experimental Settings

Datasets Construction. To ensure a fair comparison between text and image-conditioned methods while minimizing the deviation caused by external datasets, we create the Concept Text-Image Reference (CTIR) dataset. This dataset consists of 20 categories, each containing 30 images. The categories have been carefully selected to represent a wide variety of real-world objects, showcasing the model's ability to generate content across different species and types. For evaluation purposes, we develop the Concept Text-Image Blending (CTIB) dataset, which includes 5,700 text-image pairs. These pairs are drawn from a set of 190 text prompts, distributed across the 20 categories, with each category containing 30 images.

Implementation Details. We use SD-2-1[1] as the backbone model for the baseline comparisons. For pipeline, we employ SD-2-1-unCLIP[2], which enables image-based conditioning as an input. To maintain a uniform resolution throughout the experiments, all images are resized to 768×768 pixels. The experiments are conducted on a node equipped with 4 NVIDIA GeForce RTX A100 GPUs.

Metrics. Given the unique nature of our task compared to other generation tasks, we use four key metrics to evaluate the blending performance: (1) **CLIP Blending Similarity (CLIP-BS)** is the primary metric, which measures the CLIP [54] distance and similarity between the blending results and the original concepts. (2) **DINO Blending Similarity (DINO-BS)** [55] assesses the detection score of blending objects or the combination of one object with features from another. (3) **CLIP Image Quality Assessment (CLIP-IQA)** [56] evaluates the image quality and the degree of match for the blending objects. (4) **Human Preference Score (HPS)** [57] measures human preferences for the blending results based on blending object prompts. Further explanation is provided in Appendix G.

[1] `https://huggingface.co/stabilityai/stable-diffusion-2-1`

[2] `https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip`

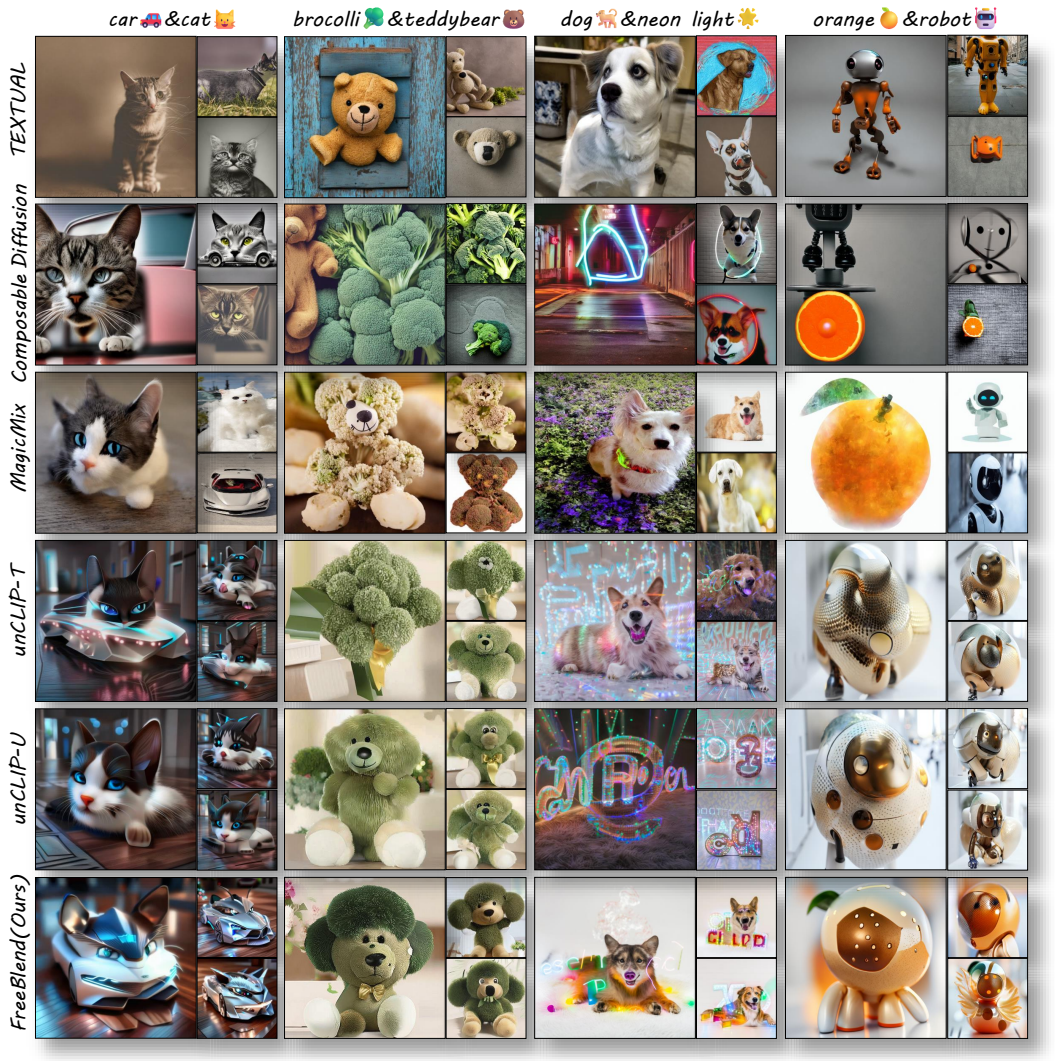

Figure 3: At the top are the concepts, and on the left are the methods we compare. Each row shows the results, with three images per method and concept pair, evaluating our method against five blending methods. Unlike others, which can suffer from rigid splicing, discordant compositions, or concept bias, our method smoothly integrates features from different concepts into a cohesive new object.

4.2 Qualitative Comparisons

The qualitative comparison of our method with five existing approaches applicable to the concept blending task is shown in Figure 3. We observe that interpolated text embeddings [16] tend to favor one category, failing to integrate both concepts effectively. Composable Diffusion [14] introduces both categories but often results in them co-occurring rather than blending harmoniously. MagicMix [18] heavily relies on shape similarity between the reference image and the categories, struggling when the categories are dissimilar, such as "dog-neon light" or "orange-robot". unCLIP-T and unCLIP-U [21] blend the two concepts well, but their blending direction and extent are suboptimal, failing to generate a meaningful blend when the categories lack visual similarity. In contrast, our method effectively merges both categories, producing well-integrated results even with highly distinct concepts like "car-cat" or "dog-neon light".

4.3 Quantitative Comparisons

We evaluate our method on our designed CTIB dataset with four key metrics, with the results presented in Table 1. The results demonstrate that our method significantly outperforms all other

Table 1: Quantitative comparisons with other methods. Our method outperforms all other methods across all metrics, including the main blending effect (CLIP-BS), the blending objects detection metric DINO-BS, image quality (CLIP-IQA), and better human preference (HPS).

Methods	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
MagicMix [18]	$8.3063_{\pm 2.4757}$	$0.2485_{\pm 0.1575}$	$0.4435_{\pm 0.0980}$	$0.2710_{\pm 0.0290}$
Composable Diffusion [14]	$6.1374_{\pm 1.9454}$	$0.2441_{\pm 0.1655}$	$0.4270_{\pm 0.1082}$	$0.2903_{\pm 0.0277}$
unCLIP-T [21]	$8.7433_{\pm 3.1892}$	$0.2214_{\pm 0.1689}$	$0.4436_{\pm 0.1073}$	$0.2384_{\pm 0.0285}$
unCLIP-U [21]	$8.7346_{\pm 3.1577}$	$0.2190_{\pm 0.1703}$	$0.4450_{\pm 0.1061}$	$0.2385_{\pm 0.0280}$
TEXTUAL [16]	$7.8102_{\pm 2.6852}$	$0.2366_{\pm 0.1537}$	$0.4164_{\pm 0.1143}$	$0.2399_{\pm 0.0230}$
UNET [17]	$8.5080_{\pm 2.7110}$	$0.2481_{\pm 0.1628}$	$0.3544_{\pm 0.2785}$	$0.2405_{\pm 0.0326}$
AID [32]	$7.0438_{\pm 3.0395}$	$0.2355_{\pm 0.1655}$	$0.4213_{\pm 0.1132}$	$0.2551_{\pm 0.0309}$
FreeBlend(Ours)	$\mathbf{9.1555}_{\pm \mathbf{2.7134}}$	$\mathbf{0.2743}_{\pm \mathbf{0.1586}}$	$\mathbf{0.5238}_{\pm \mathbf{0.0975}}$	$\mathbf{0.2932}_{\pm \mathbf{0.0316}}$

Table 2: Ablation study with different interpolation strategies. The results demonstrate that our increase strategy outperforms both the invariant and decline strategies.

Increase	Invariant	Decline	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
✓			$\mathbf{9.1555}_{\pm \mathbf{2.7134}}$	$\mathbf{0.2743}_{\pm \mathbf{0.1586}}$	$\mathbf{0.5238}_{\pm \mathbf{0.0975}}$	$\mathbf{0.2932}_{\pm \mathbf{0.0316}}$
	✓		$7.8891_{\pm 2.8732}$	$0.1970_{\pm 0.1489}$	$0.4861_{\pm 0.0990}$	$0.2769_{\pm 0.0264}$
		✓	$8.5355_{\pm 3.0644}$	$0.2222_{\pm 0.1622}$	$0.4981_{\pm 0.0836}$	$0.2712_{\pm 0.0308}$

approaches across all metrics. This indicates that the quality and the blending effect of our synthetic images achieved by our method is the most visually pleasing.

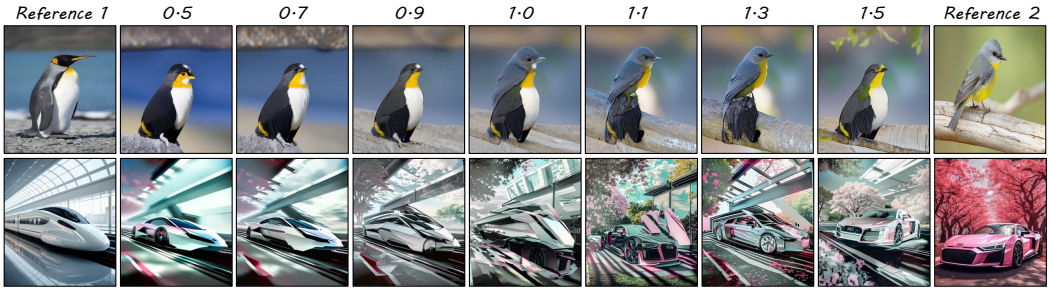

Figure 4: Ablation study on the impact of γ. The results show that, in the first row, better blending is achieved on the left side, while the right side appears more spliced. In the second row, both sides exhibit a more visually appealing effect. The blending process, however, is inherently subjective, and users can adjust the parameter γ to tailor the output according to their preferences. By adjusting γ, users can control the contribution of each concept, thereby mitigating associated biases.

4.4 Ablation Study

Mitigation of bias. Different concepts inherently exhibit biases in Stable Diffusion, likely due to the training datasets and their relevance to specific concepts or content. So we introduce an additional parameter, γ, to reduce the bias toward specific categories in the generated output image. As shown in Figure 4, we vary γ within the range $[0.5, 1.5]$. Our observations show that as γ moves from 0 to 2, the generated image gradually transitions from resembling reference 1 to favoring reference 2. This provides us with the flexibility to adjust the interpolation strength, allowing for precise control over the image's characteristics and enabling the achievement of optimal results.

Interpolation Strategy. Table 2 illustrates the impact of different interpolation strategies. In the blending stage, we vary the blending ratio between the interpolated auxiliary latents and the blending latent. From this table, we observe that the "increase" strategy yields a higher CLIP-BS and CLIP-IQA score. As p increases, the blending latent gains more weight relative to the auxiliary latents, resulting in the final output exhibiting features of both kinds. However, the "invariant" method leads to a rigid adjustment, yielding the lowest performance. On the other hand, the "decline" strategy prioritizes the auxiliary latent space and avoids blending, which clearly results in ineffective blending.

Table 3: Ablation study with staged denoising process. The results demonstrate that our staged denoising process effectively enhances both the blending quality and the scores. Both the initialization and refinement stages contribute to the overall improvement.

Initialization	Blending	Refinement	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
	✓		$8.3716_{\pm 2.8559}$	$0.2520_{\pm 0.1542}$	$\mathbf{0.5288}_{\pm \mathbf{0.0965}}$	$8.9552_{\pm 2.7259}$
	✓	✓	$8.3382_{\pm 2.8628}$	$0.2655_{\pm 0.1562}$	$0.5282_{\pm 0.0934}$	$8.9832_{\pm 2.8083}$
✓	✓		$8.5026_{\pm 2.8505}$	$0.2617_{\pm 0.1501}$	$0.5222_{\pm 0.1006}$	$9.1313_{\pm 2.6941}$
✓	✓	✓	$\mathbf{9.1555}_{\pm \mathbf{2.7134}}$	$\mathbf{0.2743}_{\pm \mathbf{0.1586}}$	$0.5238_{\pm 0.0975}$	$\mathbf{9.1555}_{\pm \mathbf{2.7134}}$

Staged Denoising Process. We validate the impact of our staged denoising process in Table 3. The results demonstrate that both the initialization and refinement stages enhance the stability and quality of the generation. This highlights the effectiveness of the initialization stage in shaping the structure of the initial latents, while the refinement stage successfully adds finer details to the images.

Feedback-Driven Mechanism. We conduct an ablation study in Table 7 to validate the effect of the feedback-driven mechanism. The results demonstrate the impact of our feedback mechanism in blending different concepts. The DINO-BS score, however, is not optimal, which may be due to the surprising overlap of blended objects, thus boosting the score. We also present an intuitive visual comparison in Figure 5, where our effect is clearly significant.

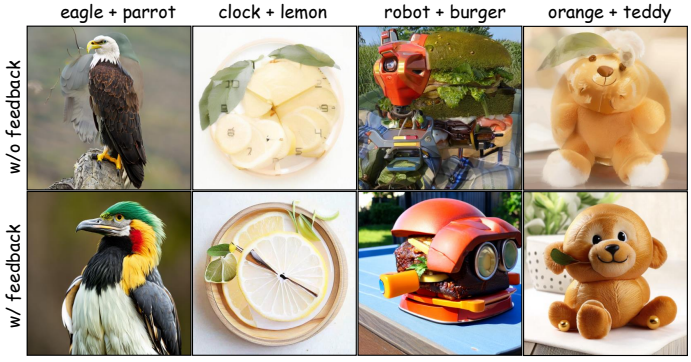

Figure 5: Ablation study of the feedback mechanism: removing the feedback module causes image overlap, disrupting blending and preventing integration.

4.5 User Study

We also conduct a user study to evaluate human preference for the blending results with images generated by different methods. Our survey comprises 50 sets of blending pairs. With the active participation of 26 volunteers, we successfully collected a total of 1300 votes. The results are displayed in Table 4. These findings not only affirm the effectiveness of our approach but also provide valuable insights into user preferences in the context of concept blending.

Table 4: User study results comparing FreeBlend with three other methods. Higher values indicate better user preference.

Method	MagicMix	TEXTUAL	Composable Diffusion	FreeBlend(Ours)	Total
User Votes (↑)	49(3.77%)	42(3.23%)	127(9.77%)	**1082(83.2%)**	1300

5 Conclusion

Blending multiple concepts into a single object based on diffusion models is both interesting and valuable for 2D content creation. However, it is quite challenging, as it requires controlling the diffusion model to generate objects that have never been seen before. In this paper, we introduce *FreeBlend*, a *training-free* method to blend concepts with stark inherent semantic dissimilarities. At its core, FreeBlend employs the unCLIP model with images as conditions and utilizes a novel proposed interpolation strategy with feedback. Extensive qualitative and quantitative studies demonstrate that our method effectively blends disparate concepts, significantly surpassing the prior SoTA counterparts. Moving forward, exploring alternative ways to address the preference dynamics between different concepts may further improve the performance.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[2] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *ICLR*, 2024.

[3] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," in *ICML*, 2024.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022.

[5] A. Lugmayr, M. Danelljan, A. Romero, and R. Timofte, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022.

[6] C. Corneanu *et al.*, "Latentpaint: Image inpainting in latent space with diffusion models," in *WACV*, 2024.

[7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[8] J. Shi *et al.*, "Instantbooth: Personalized text-to-image generation without test-time finetuning," in *CVPR*, 2023.

[9] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, "Anydoor: Zero-shot object-level image customization," in *CVPR*, 2024.

[10] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," in *NeurIPS*, 2023.

[11] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *CVPR*, 2024.

[12] G. Fauconnier and M. Turner, "Conceptual integration networks," *Cognitive science*, vol. 22, no. 2, pp. 133–187, 1998.

[13] L. D. Ritchie, "Lost in" conceptual space": Metaphors of conceptual integration," *Metaphor and symbol*, vol. 19, no. 1, pp. 31–50, 2004.

[14] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *ECCV*, 2022.

[15] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[16] S. Melzi, R. Peñaloza, A. Raganato *et al.*, "Does stable diffusion dream of electric sheep?" in *CEUR Workshop Proceedings*, 2023.

[17] L. Olearo, G. Longari, S. Melzi, A. Raganato, and R. Peñaloza, "How to blend concepts in diffusion models," *arXiv preprint arXiv:2407.14280*, 2024.

[18] J. H. Liew, H. Yan, D. Zhou, and J. Feng, "Magicmix: Semantic mixing with diffusion models," *arXiv preprint arXiv:2210.16056*, 2022.

[19] E. Richardson, K. Goldberg, Y. Alaluf, and D. Cohen-Or, "Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints," *ACM Transactions on Graphics*, 2024.

[20] Z. Xiong, Z. dong Zhang, Z. Chen, S. Chen, X. Li, G. Sun, J. Yang, and J. Li, "Novel object synthesis via adaptive text-image harmony," in *NeurIPS*, 2024.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[22] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021.

[23] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[24] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *CVPR*, 2022.

[25] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in *CVPR*, 2023.

[26] G. Couairon, M. Careil, M. Cord, S. Lathuiliere, and J. Verbeek, "Zero-shot spatial layout conditioning for text-to-image diffusion models," in *ICCV*, 2023.

[27] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[28] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," in *ICLR*, 2023.

[29] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[30] Y. Shi, C. Xue, J. H. Liew, J. Pan, H. Yan, W. Zhang, V. Y. Tan, and S. Bai, "Dragdiffusion: Harnessing diffusion models for interactive point-based image editing," in *CVPR*, 2024.

[31] C. Wang and P. Golland, "Interpolating between images with diffusion models," in *ICML Workshop*, 2023.

[32] H. Qiyuan, J. Wang, Z. Liu, and A. Yao, "Aid: Attention interpolation of text-to-image diffusion," in *NeurIPS*, 2024.

[33] M. Hamazaspyan and S. Navasardyan, "Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models," in *CVPR*, 2023.

[34] Z. Wang, L. Zhao, and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," in *ICCV*, 2023.

[35] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and versatile 3d generation," *arXiv preprint arXiv:2412.01506*, 2024.

[36] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," *arXiv preprint arXiv:2212.05032*, 2022.

[37] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," in *ICCV*, 2023.

[38] S. Mo, F. Mu, K. H. Lin, Y. Liu, B. Guan, Y. Li, and B. Zhou, "Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition," in *CVPR*, 2024.

[39] N. Liu, S. Li, Y. Du, J. Tenenbaum, and A. Torralba, "Learning to compose visual relations," in *NeurIPS*, 2021.

[40] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," in *AAAI*, 2024.

[41] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *CVPR*, 2023.

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[43] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[44] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[45] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS Workshop*, 2021.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[47] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023.

[48] Z. Zhan, D. Chen, J.-P. Mei, Z. Zhao, J. Chen, C. Chen, S. Lyu, and C. Wang, "Conditional image synthesis with diffusion models: A survey," *arXiv preprint arXiv:2409.19365*, 2024.

[49] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[50] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI*, 2024.

[51] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *ICLR*, 2023.

[52] H. Lin, "Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation," in *CVPR*, 2024.

[53] J. Ackermann and M. Li, "High-resolution image editing via multi-stage blended diffusion," *arXiv preprint arXiv:2210.12965*, 2022.

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[55] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *ECCV*, 2025.

[56] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

[57] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.

[58] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[59] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan, "Diffmorpher: Unleashing the capability of diffusion models for image morphing," in *CVPR*, 2024.

A T-SNE Analysis

In this section, we explore the CLIP embedding space of blended images. Given the high dimensionality of the embeddings, we apply t-SNE [58] for dimensionality reduction and visualize the embeddings in 2D to illustrate their spatial relationships with the corresponding concepts.

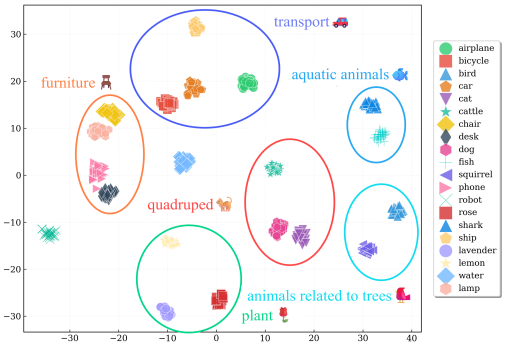

Figure 6: Visualization of the CLIP feature distribution for various image categories after dimensionality reduction via t-SNE.

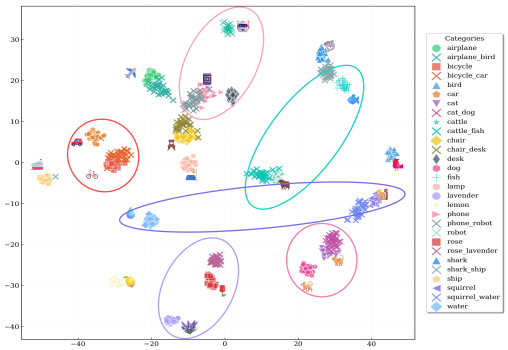

Figure 7: Visualization of the CLIP feature distribution for original class images and blending class images after dimensionality reduction via t-SNE.

In Figure 6, it is evident that semantically related categories tend to be positioned closer to each other, while unrelated categories are spaced farther apart. This is reasonable given the contrastive training objective of CLIP. In Figure 7, we observe that the blending images exhibit a strong relationship with the original classes. Notably, these blending images are generally biased towards one of the original classes, which reflects the inherent bias in the concept blending task.

Specifically, we notice that the blending images are not located at the midpoint between the original classes. We hypothesize that this observation can be attributed to the following factors: a) t-SNE is a non-linear dimensionality reduction technique that aims to map similar points in high-dimensional space to nearby points in low-dimensional space, while preserving local structures rather than global relationships. Consequently, the position of points in the t-SNE plot may not necessarily correspond to the relative position of categories in the original feature space. b) The feature vector of the blended class may not be positioned directly between the two original classes. This could be due to: i) The blended class containing a mixture of features from original classes, with the resulting feature vector being closer to one of the original classes in the feature space. ii) The blended feature may have a direction in the feature space that is relatively distant from both original classes. In certain cases, this could cause the blended class's feature vector to extend further from the center, leading to its placement behind the original classes in the t-SNE plot, rather than being positioned between them.

B Comparison on TIF Dataset

To provide a convincing comparison with other image-conditioned methods, we conduct experiments on the TIF dataset [20]. Table 5 demonstrates that our method consistently achieves superior results.

Table 5: Quantitative comparisons with other image-conditioned methods on TIF dataset.

Methods	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
unCLIP-T [21]	$11.2156_{\pm 4.3520}$	$0.2277_{\pm 0.0415}$	$0.4162_{\pm 0.1594}$	$0.2020_{\pm 0.1954}$
unCLIP-U [21]	$11.2767_{\pm 4.1601}$	$0.1687_{\pm 0.1889}$	$0.4399_{\pm 0.1611}$	$0.2163_{\pm 0.0445}$
AID [32]	$11.2288_{\pm 3.6646}$	$0.1378_{\pm 0.1792}$	$0.4378_{\pm 0.1630}$	$0.2191_{\pm 0.0438}$
FreeBlend(Ours)	$\mathbf{11.9740}_{\pm \mathbf{4.0038}}$	$\mathbf{0.1784}_{\pm \mathbf{0.1715}}$	$\mathbf{0.5029}_{\pm \mathbf{0.1663}}$	$\mathbf{0.2500}_{\pm \mathbf{0.0465}}$

C More Ablation Study

Conditions for Denoising Process. Table 6 presents the results of our ablation study under different conditions. For text conditions, we use the TEXTUAL and UNET text prompt blending methods for comparison. For image conditions, we apply the unCLIP method along with either unCLIP-T (TEXTUAL-like) or unCLIP-U (UNET-like) image blending methods. We also explore the method of simultaneously using text and images. The results demonstrate that the unCLIP-U method yields the best performance compared to other conditions, while text-based conditions tend to reduce the effectiveness of the generation.

Table 6: Ablation study for our method with different conditions. We find that the unCLIP-U method, which incorporates two embeddings—one for downsampling and the other for upsampling—achieves the best blending results.

TEXTUAL	UNET	unCLIP-T	unCLIP-U	CLIP-BS(↑)
		✓		$8.5577_{\pm 2.6566}$
			✓	**$9.1555_{\pm 2.7134}$**
✓				$7.5187_{\pm 2.5852}$
✓		✓		$7.9888_{\pm 2.5823}$
✓			✓	$8.8170_{\pm 2.7085}$
	✓			$7.4068_{\pm 2.5855}$
	✓	✓		$6.9255_{\pm 2.4354}$
	✓		✓	$9.0260_{\pm 2.7408}$

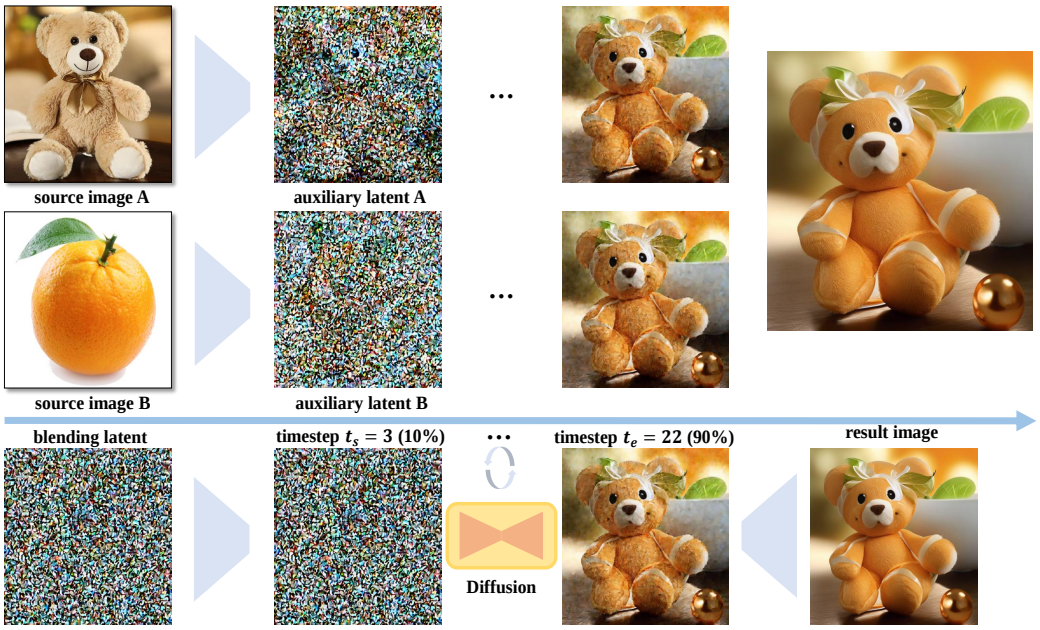

Figure 8: Visualization of our staged denoising process. From the image, we can observe the effects of the three stages. From the start to t_s, the image establishes the basic layout of the blending. During the blending stage, both the blending latent and the auxiliary latents are iteratively updated. Finally, in the refinement stage, the quality of the blended image is significantly improved.

Visualization of Staged Denoising Process. Here, we provide a more detailed description of the staged denoising process employed in our method. To facilitate understanding, we visualize the intermediate latent representations, including the blending latent L_b and the auxiliary latents $L_a^{(t,n)}$ $(n = 0, 1, 2, ..., N)$ at various stages of the process. The results are presented in Figure 8. Specifically, we focus on timesteps 3 and 22, as they correspond to the initial and final stages where our method begins to exhibit its effects. These visualizations demonstrate how the latents evolve over timesteps,

gradually converging to a consistent structure. Through the feedback interpolation module, the blending latent and auxiliary latents seamlessly merge, ultimately resulting in a coherent output image.

Table 7: Ablation study with our feedback mechanism.

Feedback	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
	$8.9211_{\pm 3.0076}$	$\mathbf{0.2860}_{\pm \mathbf{0.1663}}$	$0.4874_{\pm 0.1035}$	$0.2866_{\pm 0.0305}$
✓	$\mathbf{9.1555}_{\pm \mathbf{2.7134}}$	$0.2743_{\pm 0.1586}$	$\mathbf{0.5238}_{\pm \mathbf{0.0975}}$	$\mathbf{0.2932}_{\pm \mathbf{0.0316}}$

D Details of Dataset Construction

Table 8: Categories of Concepts. Categories of these selected concepts for blending are designed to encompass the primary objects commonly found in the real world.

Category	Names
Transports	airplane, bicycle, car, ship
Animals	bird, cat, cattle, dog, fish, squirrel, shark
Common objects	chair, desk, phone, robot, lamp
Nature	rose, lavender, lemon, water

The designed category list can be divided into four distinct groups based on their characteristics, as shown in Table 8. This categorization provides a clear organization of items based on their respective categories and reflects common concepts encountered in daily life.

E Why Vanilla Stable Diffusion Fails

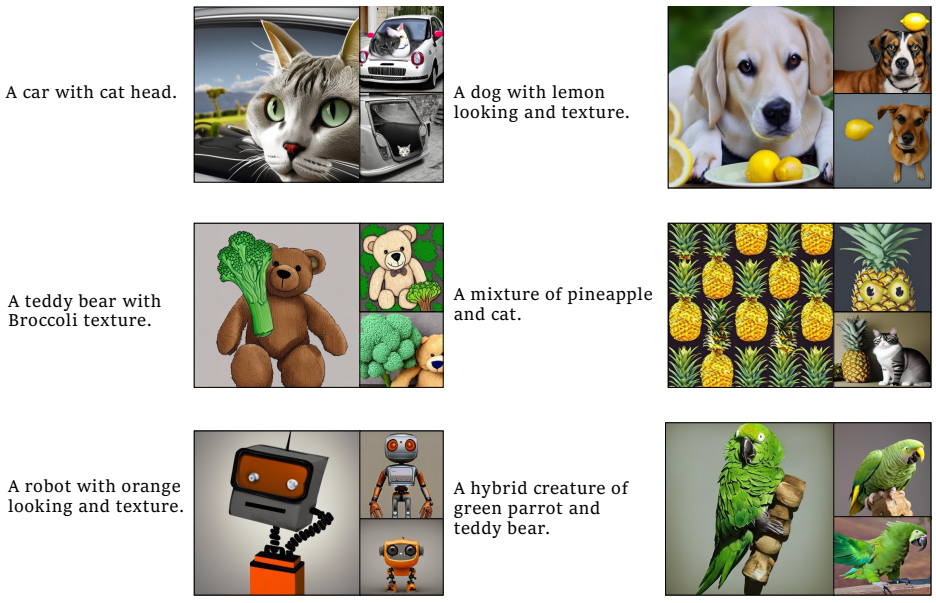

Figure 9: Relying solely on prompts for concept blending, we find that Stable Diffusion 2.1 often fails to effectively merge the concepts as instructed. Instead, it typically depicts the two concepts as separate, coexisting elements within the image.

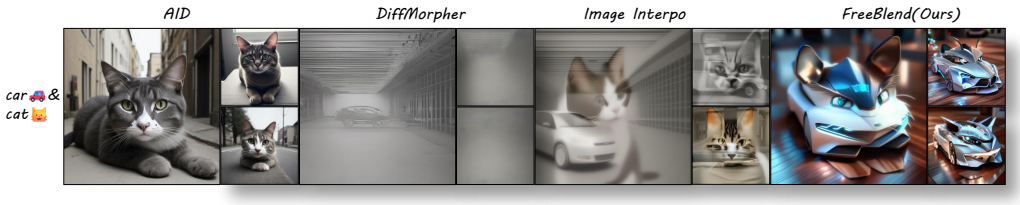

Figure 10: When compared to other image interpolation methods, we find that AID [32] exhibits a stronger attention bias towards maintaining the quality of generated images, but it does not effectively blend concepts together. DiffMorpher [59] is designed to smoothly transfer one image to another, but when applied to unrelated concepts, it generates white noise. On the other hand, Image Interpo [31] interpolates between two different images but suffers from significant overlapping issues.

F More Style Tasks

Image Interpolation Task. Figure 10 presents a qualitative comparison with other image interpolation methods, demonstrating that our method achieves the best results.

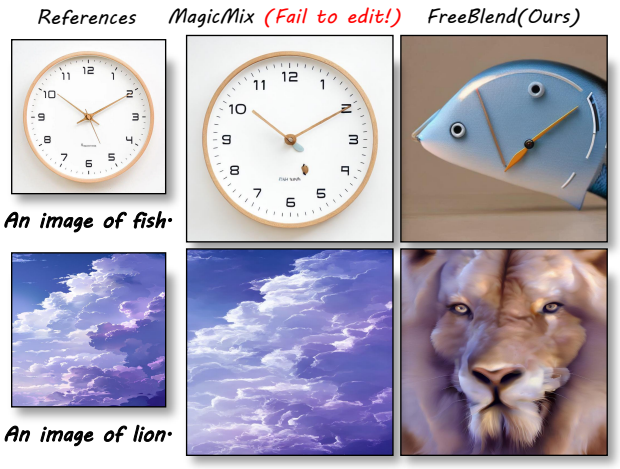

Figure 11: Comparison with the setting of MagicMix [18]. Given a reference image and a prompt, the model can transfer the reference image to another concept. From the results, we observe that when there is a significant semantic and shape gap between the original and target concepts, MagicMix struggles to modify the original image while preserving its features, without introducing new ones. This highlights that our method is capable of handling concepts with substantial semantic and shape differences.

Switching One Condition from Image to Text. Figure 11 presents the qualitative comparison between MagicMix [18] and our method.

Style Transfer Task. Figure 12 illustrates the impressive visual performance of our method on the style transfer task. Given a reference image and a content image, our training-free approach effortlessly transfers the content into the style of the reference, demonstrating both the adaptability and compatibility of our method for this task.

G Details of Metrics

We calculated the corresponding scores for each image within a category, then computed the average score for that category, and finally determined the overall average score for the entire dataset.

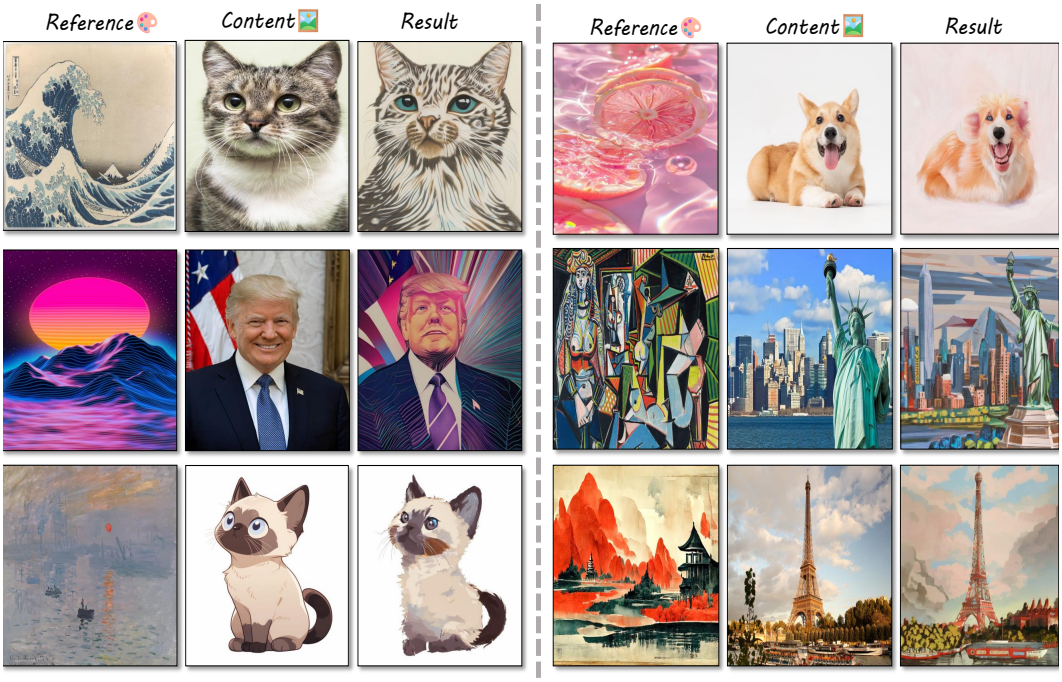

Figure 12: Adapting our method to the style transfer task.

CLIP-BS. For class A and class B, we calculate the CLIP blending score using the following equation:

$$\text{Score}_{\text{CLIP-BS}} = \sum_{i=1}^{n} \left[\text{Score}(\text{Image}_{\text{blending}}, \text{Prompt}_i) - \text{Score}(\text{Image}_{\text{original}_i}, \text{Prompt}_i) \right], \quad (9)$$

where $\text{Score}(\cdot, \cdot)$ is the cosine similarity score computed by the CLIP model using the input image and prompt, $\text{Image}_{\text{blending}}$ represents the generated blended images, $\text{Image}_{\text{original}_i}$ denotes the original images of class i, and n is the number of concepts blended. Prompt_i refers to the prompt for class i, formatted as "a photo of a {class}" to minimize noise in the image generation process. This metric can also be interpreted as the high-dimensional distance between the blended concepts and the original ones. It represents the total distance between the blended concept and each of the original concepts. A larger score generally indicates the dimensional center of the original concepts. Note that irrelevant content could potentially receive a high score using this metric due to unrelated features. However, this scenario is unlikely to occur, as our generated images are based on relevant concepts.

DINO-BS [55]. Grounding DINO is an open-set object detection model designed to improve the ability to interpret human language inputs through tight modality fusion and large-scale grounded pre-training. The model excels at detecting unseen objects and can perform effective zero-shot transfer learning, even for categories not encountered during training, by leveraging language prompts. To guide the model in detecting blended objects, we designed two specific prompts: "a {concept_a} with blending features from {concept_b}" and "a {concept_b} with blending features from {concept_a}". These prompts help the model appropriately recognize and identify the blended objects.

CLIP-IQA [56]. This metric employs a thoughtfully crafted prompt pairing strategy to reduce linguistic ambiguity and effectively harness CLIP's pretrained knowledge for assessing visual perception. In our task, we design specific prompt pairs like ("mixed", "dull"), ("blending features from two different objects", "natural object from one object").

HPS [57]. This metric more accurately predicts human preferences for generated images. The designed prompt is "a photo of a blended object combining mixed features from {concept_a} and {concept_b}".

H Baseline Methods

In this section, we describe the baseline methods used in our study:

- **MagicMix [18].** The original MagicMix method modifies the input image using a text prompt from a different class to transfer the image's characteristics. In our implementation, we first generate the original image using a class text prompt with Stable Diffusion, and then pass it into the original MagicMix pipeline.
- **Composable Diffusion [14].** This method decomposes the text description into components, each processed by a different encoder to produce a latent vector capturing its semantic information. Multiple diffusion models are then used to independently generate images for each component. The Conjunction operation combines these components into a single image. This approach excels in zero-shot compositional generation, allowing the model to create novel combinations *but not blending* of objects and their relationships, even without prior training on such combinations.
- **unCLIP-T [56].** The unCLIP model leverages images as conditioning inputs, replacing traditional text prompts, by training several linear layer adapters. In this implementation, image embeddings are treated similarly to text embeddings, with their average embedding being computed to serve as the conditions.
- **unCLIP-U [56].** This implementation is akin to unCLIP-T, with the key difference being that the image embeddings are integrated using the UNET method.
- **TEXTUAL [16].** This method encodes the text prompts into text embeddings and computes their average embedding to input into Stable Diffusion for generating blended images. However, it is unstable within the embedding space and constrained by the semantic scope of the concepts involved.
- **UNET [17].** This method uses the first text embedding for the conditions of downsampling of U-Net [46] and the second one for the upsampling of the U-Net.
- **AID [32].** This paper introduces Attention Interpolation via Diffusion (AID), a training-free method for improving conditional image interpolation in text-to-image models. AID enhances consistency and smoothness, with a variant (PAID) enabling user-guided interpolation. Experiments show AID outperforms traditional methods and enhances control for image editing.

I User Study

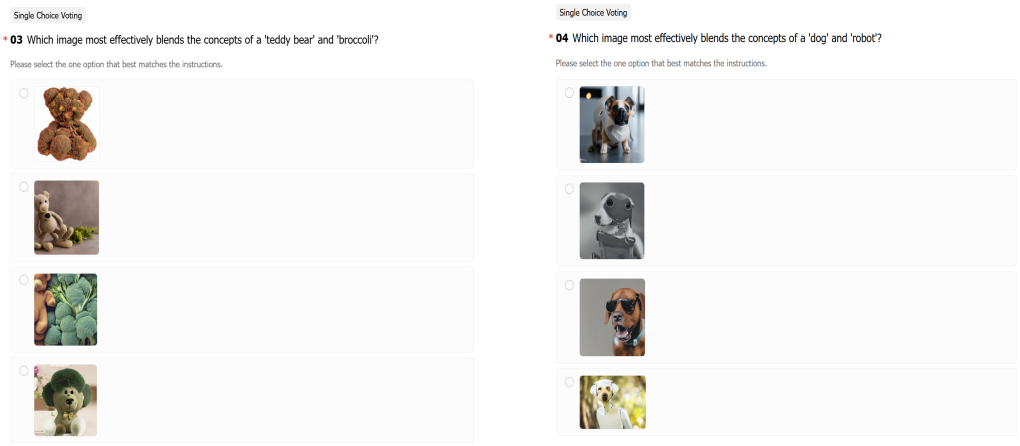

Figure 13: Screenshots of user study.

The screenshots in Figure 13 display our questionnaires used to assess the validation of different methods.

J More Visual Results

Figure 14 and Figure 15 present several of our blending results, highlighting the strong capabilities of our method for both concept blending and generation tasks.

K More Analyses of Blending Results

Blending is a multifaceted process that occurs across various domains, each exhibiting distinct characteristics and behaviors. In the study of embedding blending, it is evident that different categories impose constraints on the resulting blends.

K.1 Results from Different Categories of Concepts

Table 9: Mean scores for different categories.

Category	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
Animals	**9.4410**	0.2460	0.5160	0.2800
Common objects	7.9810	0.4160	0.5270	0.3000
Nature	9.2250	**0.4190**	0.3330	0.2760
Transports	9.3100	0.2280	**0.5430**	**0.3060**

Table 10: Mean scores for blending category pairs.

Category Pair	CLIP-BS(↑)	DINO-BS(↑)	CLIP-IQA(↑)	HPS(↑)
Animals & Common	8.7110	0.3308	0.5212	0.2908
Animals & Nature	9.3330	0.3323	0.4244	0.2780
Animals & Transports	**9.3760**	0.2369	0.5294	0.2930
Common & Nature	8.6030	**0.4170**	0.4298	0.2880
Common & Transports	8.6450	0.3217	**0.5348**	**0.3030**
Nature & Transports	9.2680	0.3232	0.4380	0.2910

Table 9 and Table 10 show the blending scores of four metrics for different category pairs. The CLIP-BS in Table 9 indicates that similar types of objects, such as animals, nature, and transports, achieve higher scores. However, for common objects, the score is lower due to the larger diversity within this category. Regarding DINO-BS, categories in nature tend to blend more successfully. In the case of HPS, we observe that transport categories generate higher-quality images. In Table 10, the pattern is similar to that in Table 9, where the best scores align with the original categories. Additionally, compared to nature categories, animals tend to blend better, which we hypothesize is due to biological reasons discussed below.

K.2 Factors of Blending Concepts

Distance between Concepts. Distance is a crucial factor that determines the effectiveness of the blending. This principle is illustrated through t-SNE analysis in Appendix A, which demonstrates how the proximity of two concepts within an embedding space influences their ability to blend. When two concepts are close to each other in this space, they blend easily, forming a cohesive blend. However, as the distance between the concepts increases, the blending becomes less seamless. For example, concepts that are close in proximity, such as cats and dogs, tend to blend effortlessly, with the final output lying somewhere between both concepts. When concepts are moderately distant, such as a cactus and a rose, partial blending occurs, with the resulting image leaning more towards one concept while still maintaining some elements of the other. As the distance between concepts becomes even greater, such as in the case of an airplane and a bird, the blending becomes limited, and the final result often favors one concept, with the other remaining a secondary influence.

L Limitations and Future work

In the case of the Stable Diffusion model, the bias towards different concepts requires adjusting the parameter γ, a process that is cumbersome and inefficient. Moreover, when blending three or more concepts, it becomes challenging because the UNET architecture can only support two conditions. Therefore, our goal is to explore more stable, cost-effective, and training-free generation methods, while also seeking a more structured approach for blending multiple concepts.

M Impact Statement

This paper presents a contribution aimed at advancing the field of concept blending. We explicitly state that our work is intended for academic research and commercial applications, provided such uses are authorized by the author. However, we strongly emphasize that we do not condone, nor do we support, the use of our research to generate harmful, unethical, or unlawful content. The generation of unsafe or offensive images that violate ethical standards, legal regulations, or societal norms is strictly prohibited. We use the Safe Checker from Stable Diffusion [1] to prevent the generation of unsafe content. As researchers, we believe it is our responsibility to ensure that our work is used for the betterment of society and in a manner that aligns with fundamental ethical principles.

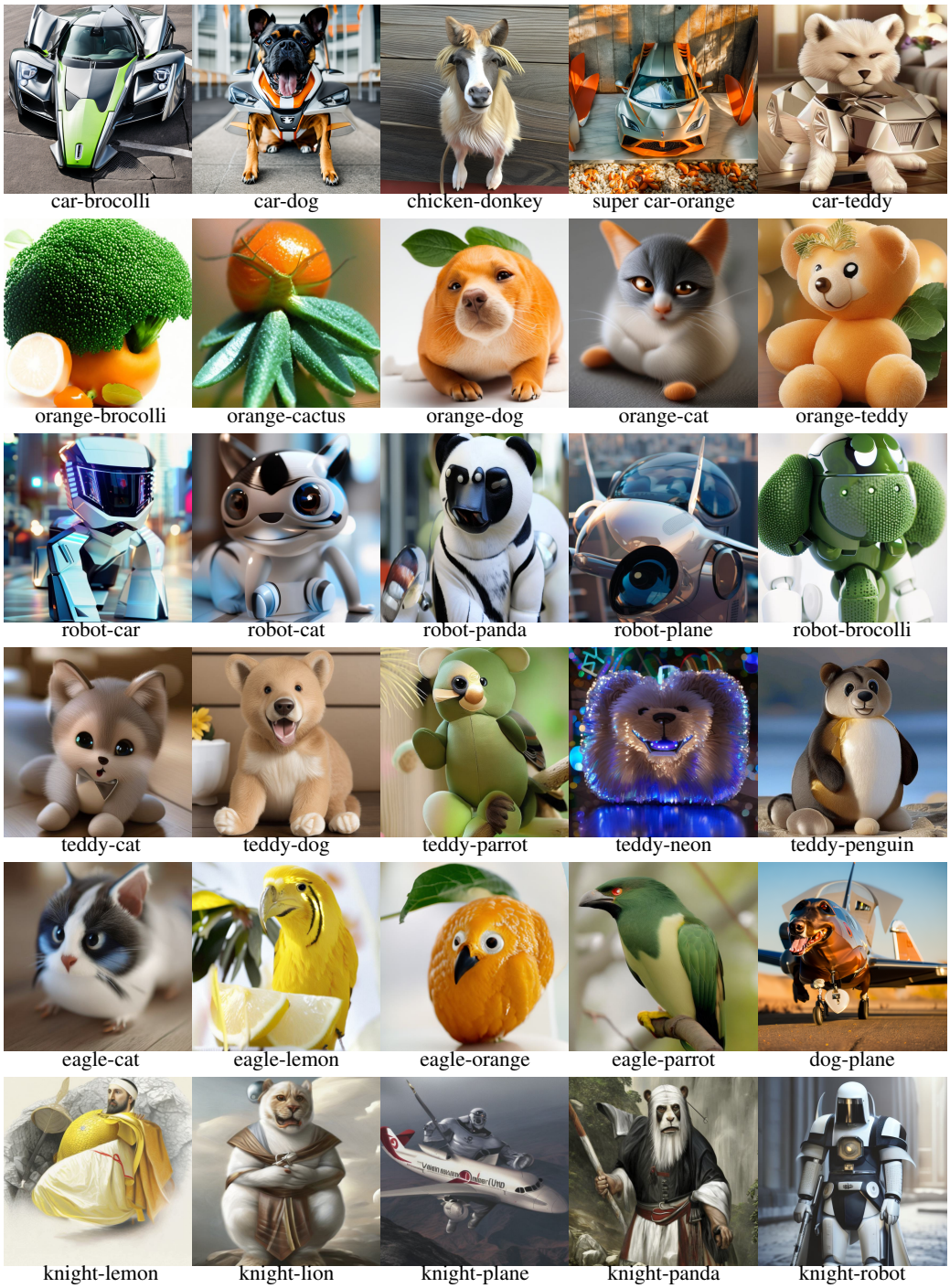

Figure 14: More blending results. Given two concepts, our approach can blend them into a novel object that the model has never seen before, generating high-quality images.

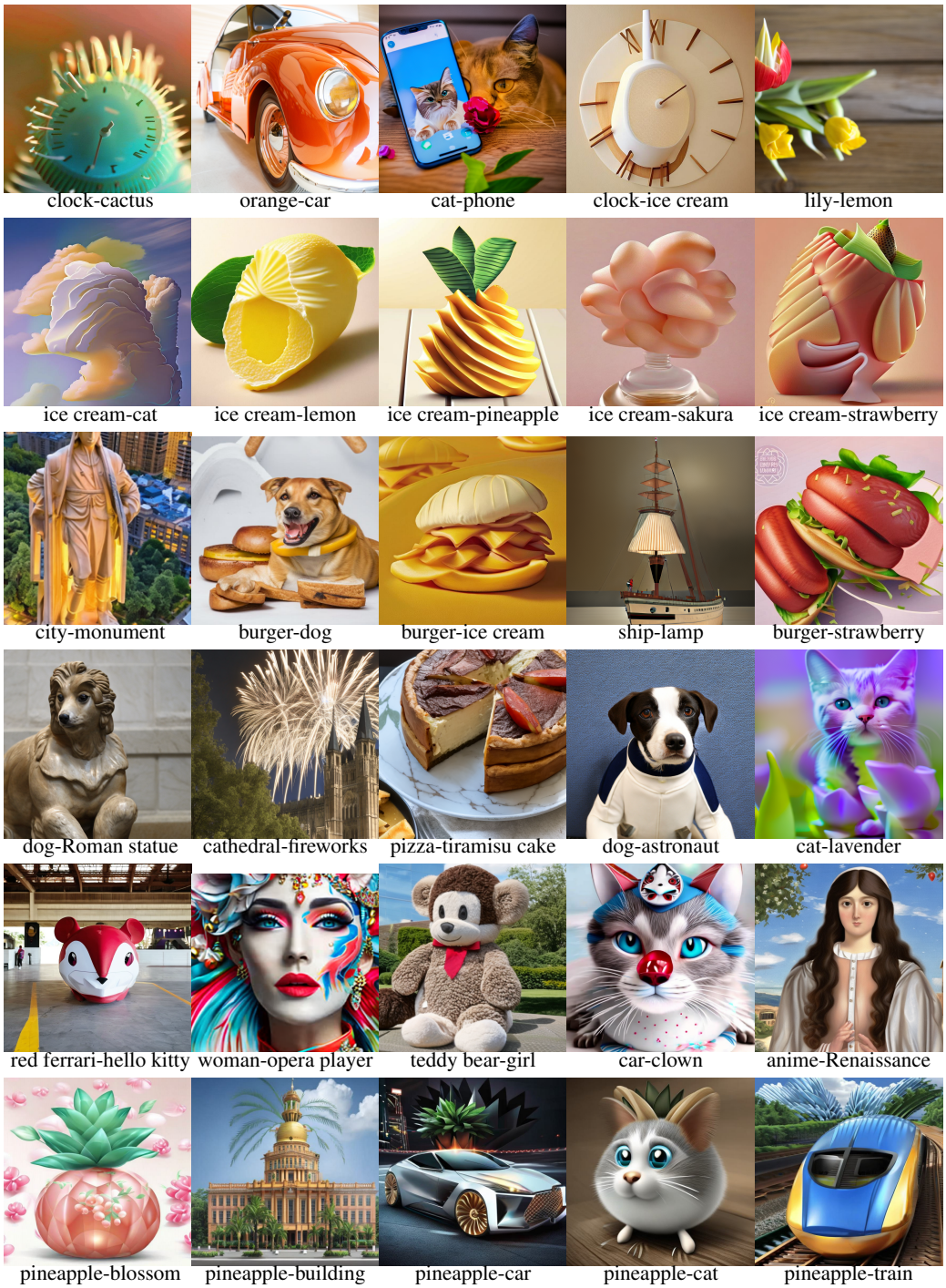

Figure 15: More blending results. Given two concepts, our approach can blend them into a novel object that the model has never seen before, generating high-quality images.