VMDiff: Visual Mixing Diffusion for Limitless Cross-Object Synthesis

Anonymous authors
Paper under double-blind review

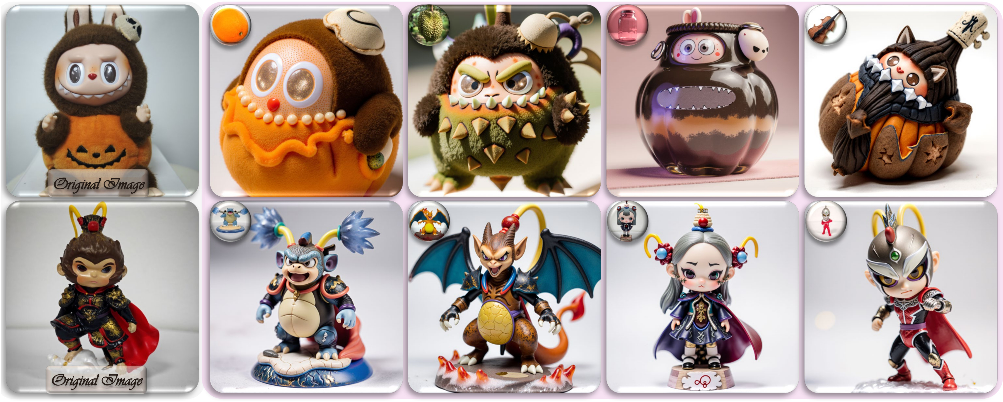


Figure 1: Two groups (rows) illustrating our VMDiff's capability to generate coherent hybrid objects. For each group, images from the 2*nd* to the 5*th* column are the product of fusing the source image in the 1*st* column with the corresponding image in the top left.

Abstract

Creating novel images by fusing visual cues from multiple sources is a fundamental yet underexplored problem in image-to-image generation, with broad applications in artistic creation, virtual reality and visual media. Existing methods often face two key challenges: *coexistent generation*, where multiple objects are simply juxtaposed without true integration, and *bias generation*, where one object dominates the output due to semantic imbalance. To address these issues, we propose **Visual Mixing Diffusion (VMDiff)**, a simple yet effective diffusion-based framework that synthesizes a single, coherent object by integrating two input images at both noise and latent levels. Our approach comprises: (1) a ***hybrid sampling process*** that combines guided denoising, inversion, and spherical interpolation with adjustable parameters to achieve structure-aware fusion, mitigating coexistent generation; and (2) an ***efficient adaptive adjustment*** module, which introduces a novel similarity-based score to automatically and adaptively search for optimal parameters, countering semantic bias. Experiments on a curated benchmark of 780 concept pairs demonstrate that our method outperforms strong baselines in visual quality, semantic consistency, and human-rated creativity. Project.

1 Introduction

Synthesizing novel images by combining visual elements from multiple sources is a fundamental challenge in image-to-image generation, with wide applications in virtual reality (Haque et al., 2023; Chen et al., 2024), digital media (Zheng et al., 2024; Zhao et al., 2024), product design (Ju et al., 2024; Sheynin et al., 2024; Wang et al., 2024) and film and game (Ceylan et al., 2023; Liu et al., 2024). In particular, visual composition methods generate high-fidelity images by composing objects through various strategies, such as combining object words into complex sentences (Liu et al., 2022), merging multiple objects (Liu et al., 2021), or blending scenes and styles (Zou et al., 2025).

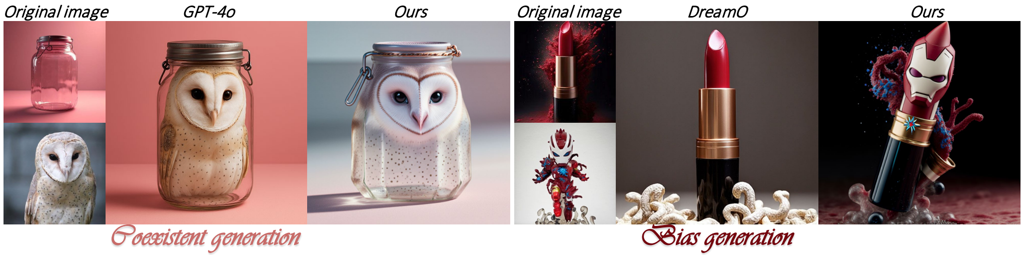


Figure 2: **Failed fusions between two object images.** GPT-4o OpenAI (2025) performs ***coexistent generations*** (left), while DreamO (Mou et al., 2025) exhibits ***bias generations*** (right). In contrast, our method achieves a seamless and harmonious fusion of the two objects.

Although these approaches effectively position different objects or parts within an image, they often struggle to seamlessly integrate distinct elements into a single object. Recent semantic mixing (Li et al., 2024; Xiong et al., 2024) explores novel object synthesis by combining textual descriptions of one object with another images or text. In contrast, this work focuses on visual mixing—directly blending two object images into a single, imaginative, and visually cohesive concept.

However, when existing powerful methods are used to perform this visual mixing task, we identify two key limitations. First, **coexistent generation** (see Fig. 2, left) occurs when different objects merely appear in the same scene—either side-by-side or partially overlapped—without achieving true visual and semantic integration. While the resulting compositions are spatially coherent, they remain conceptually disjoint. For example, OpenAI's recent GPT-4o (OpenAI, 2025) produces an image where the glass jar and owl overlap but fail to meaningfully fuse. Second, **bias generation** (see Fig. 2, right) arises when the model generates only one object while omitting the other. This asymmetry likely stems from imbalanced representations or unresolved semantic conflicts, leading to outputs that disproportionately emphasize one object. For instance, DreamO (Mou et al., 2025) generates the lipstick while entirely neglecting the iron man figurine.

To address these limitations, we develop **Visual Mixing Diffusion (VMDiff)**, a simple yet effective framework for synthesizing novel, coherent objects that seamlessly integrate two input images. VMDiff ensures structural plausibility and semantic balance through two key components: a ***Hybrid Sampling Process (HSP)*** and an ***Efficient Adaptive Adjustment (EAA)***. HSP integrates the two inputs through noise inversion and feature fusion. The inversion refines an initial noise vector conditioned on a concatenated input object embedding with two parameters and their corresponding text prompt, ensuring deep information mixing to prevent mere juxtaposition. Subsequently, feature fusion employs a curvature-respecting interpolation to blend image embeddings, with a scale factor controlling either object from dominating and thus countering bias generation. EAA automates the search for optimal parameters by proposing a novel similarity-based score that measures alignment with both visual/semantic similarity and balance between the fused object and the input object images/their category labels. By maximizing this score, the EAA dynamically adjusts the influence of each input, ensuring semantically coherent and visually faithful fusions across diverse object pairs.

Our contributions are summarized as follows: **(1)** We introduce a *hybrid sampling process* that constructs optimized semantic noise via guided denoising and inversion, combined with a curvature-aware latent fusion strategy using spherical interpolation for smooth and tunable blending. **(2)** We present an *efficient adaptive adjustment* algorithm that adjusts fusion parameters to achieve semantic and visual balance via a lightweight score-driven search. **(3)** By integrating them, we propose VMDiff, a unified and controllable framework for object-level visual concept fusion. Experiments on a curated benchmark of 780 concept pairs demonstrate that our method achieves superior object synthesis, excelling in semantic consistency, visual harmony, and user-rated creativity.

2 Related Work

Multi-Concept Generation. Multi-concept generation seeks to synthesize images representing multiple user-defined concepts, typically from a few reference images per concept. Early works such as Custom Diffusion (Kumari et al., 2023) and SVDiff (Han et al., 2023) extend single-concept personalization by fine-tuning on joint data or merging customized models. Later methods (Gu et al., 2023; Liu et al., 2023b) enhance compositionality by merging LoRA modules or token embeddings

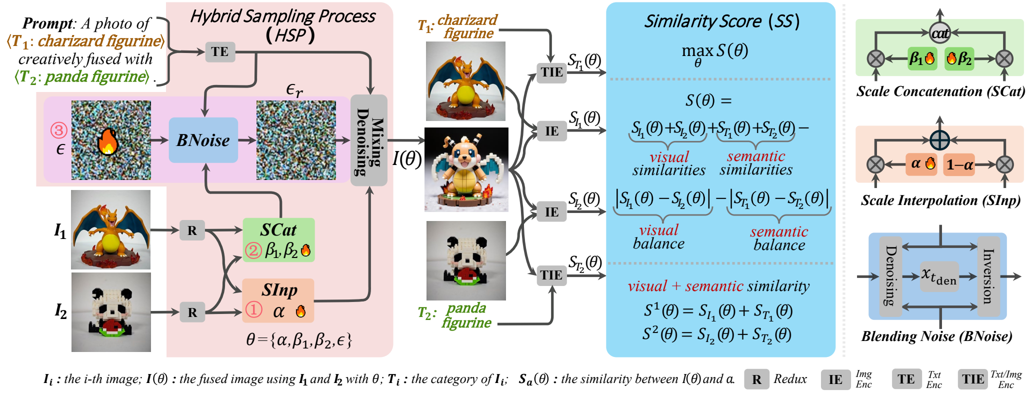

Prompt: A photo of ⟨T1: charizard figurine⟩ creatively fused with ⟨T2: panda figurine⟩ .
Hybrid Sampling Process (HSP)
TE
③ ϵ
BNoise
ϵr
Mixing Denoising
I(θ)
I1
I2
R
R
SCat
② β1,β2
SInp
① α
θ={α,β1,β2,ϵ}
T1: charizard figurine
T2: panda figurine
TIE
IE
IE
TIE
ST1(θ)
SI1(θ)
SI2(θ)
ST2(θ)
Similarity Score (SS)
max_θ S(θ)
S(θ) =
SI1(θ)+SI2(θ)+ST1(θ)+ST2(θ) −
visual similarities
semantic similarities
|SI1(θ) − SI2(θ)| − |ST1(θ) − ST2(θ)|
visual balance
semantic balance
visual + semantic similarity
S1(θ) = SI1(θ) + ST1(θ)
S2(θ) = SI2(θ) + ST2(θ)
cat
β1
β2
Scale Concatenation (SCat)
α
1−α
Scale Interpolation (SInp)
Denoising
x_tden
Inversion
Blending Noise (BNoise)
Ii : the i-th image; I(θ) : the fused image using I1 and I2 with θ; Ti : the category of Ii; Sa(θ) : the similarity between I(θ) and a.
R Redux
IE Img Enc
TE Txt Enc
TIE Txt/Img Enc

Figure 3: **Overview of our VMDiff framework.** Given two input images and their categories, the Hybrid Sampling Process (HSP) fuses them using noise inversion, scale interpolation (SInp) and scale concatenation (SCat). Efficient adaptive adjustment (EAA) optimizes fusion parameters $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$ via a similarity score (SS) that measures visual, semantic, and balance consistency.

via gradient fusion (Gu et al., 2023) or spatial inversion (Zhang et al., 2024). More recent approaches further improve efficiency and flexibility: FreeCustom (Ding et al., 2024) employs multi-reference self-attention and weighted masks for training-free composition, while MIP-Adapter (Huang et al., 2025) mitigates object confusion with a weighted-merge strategy. OmniGen (Xiao et al., 2025) and DreamO (Mou et al., 2025) provide unified instruction-based frameworks for diverse generation tasks. Unlike prior methods that explicitly separate input concepts, our approach introduces a unified fusion framework that integrates two concept inputs into a novel object with coherent structure and balanced semantics.

Semantic Mixing. Creativity, spanning domains from scientific theories to culinary recipes, has long been a key driver of progress in artificial intelligence (Boden, 2004; Maher, 2010; Wang et al., 2023; Xiong et al., 2025b). In this context, semantic mixing has emerged as a promising approach for generating novel objects by fusing features from multiple concepts into a single coherent representation. Unlike traditional style transfer (Zhang et al., 2023; Tang et al., 2023; Ke et al., 2023) or image editing (Avrahami et al., 2025; Dong & Han, 2023; Brooks et al., 2023; Gal et al., 2023)—which emphasize texture transfer or localized modifications while preserving layout—semantic mixing focuses on concept-level integration within a single entity. Conceptlab (Richardson et al., 2024) interpolates token embeddings to synthesize imaginative entities, while TP2O (Li et al., 2024) enhances controllability by aligning and blending prompt embeddings. However, both operate purely in the textual domain and lack support for real visual content. MagicMix (Liew et al., 2022) fuses image latents with text prompts during denoising, preserving spatial structure, while ATIH (Xiong et al., 2024) improves semantic alignment through more coordinated integration of visual and textual inputs. FreeBlend (Zhou et al., 2025) performs staged interpolation in latent space to produce blended objects. In contrast, our method integrates structural and semantic cues from real image concepts, generating hybrid objects that are both visually coherent and semantically balanced.

3 Visual Mixing Diffusion

In this section, we present a Visual Mixing Diffusion (**VMDiff**) for synthesizing novel objects images in Fig. 3. Our method consists of two key components. We introduce a Hybrid Sampling Process (**HSP, §3.1**) that generates a new object image by blending two distinct inputs using learned scale factors and noise. An Efficient Adaptive Adjustment (**EAA, §3.2**) dynamically adjusts the scale factors and noise based on a Similarity Score (**SS**), ensuring high-quality object synthesis.

3.1 Hybrid Sampling Process

Given two distinct images I_1 and I_2, along with their respective category labels T_1 and T_2 (e.g., *Iron Man* and *Duck*), we first construct a guiding prompt P_G: "*A photo of* $< T_1 >$ *creatively fused with* $< T_2 >$." and sample an initial Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. For convenience, we denote an input data $D = \{I_1, I_2, T_1, T_2, P_G\}$. We first employ pretrained image/text encoders $\mathcal{E}_I(\cdot)/\mathcal{E}_T(\cdot)$ of Flux-

Krea (Lee et al., 2025) to project both visual and textual modalities into a unified image-language latent space. Specifically, these embeddings are extracted by $z_1 = \mathcal{E}_I(I_1),\ z_2 = \mathcal{E}_I(I_2),\ z_p = \mathcal{E}_T(P_G)$. Using these embeddings, HSP includes *blending noise* and *mixing denoise*.

Blending Noise (BNoise): Directly sampling standard Gaussian noise to generate a blend of two objects frequently produces incomplete results, with key features such as arms or legs missing (Fig. 4). This occurs because random noise contains no information about the input objects. Our solution is to refine an initial noise vector ϵ, transforming it into a visually and semantically-informed estimate that faithfully represents the source data. Inspired by Rectified Flow (Albergo & Vanden-Eijnden, 2023), this is achieved through a guided denoising and inversion process. Using inputs ϵ, z_1, z_2, z_p, we denoise to an intermediate timestep t_{den}, and invert to a refined noise ϵ_b, which is defined as:

$$\begin{aligned} \hat{x}_t = x_{t_{\text{den}}} &\Leftarrow \overbrace{x_{t-1} = x_t - (\sigma_t - \sigma_{t-1})v_\phi(x_t, t, z_{\text{SCat}}(z_1, z_2; \beta_1, \beta_2), \gamma_{\text{den}}, z_p)}^{\text{denoise: } t \text{ decreases from } T \text{ to } t_{\text{den}}, \text{ starting } x_T = \epsilon}, \\ \epsilon_b = \underbrace{\hat{x}_T}_{\text{BNoise}} &\Leftarrow \underbrace{\hat{x}_{t+1} = \hat{x}_t + (\sigma_{t+1} - \sigma_t)v_\phi(\hat{x}_t, t, z_{\text{SCat}}(z_1, z_2; \beta_1, \beta_2), \gamma_{\text{inv}}, z_p)}_{\text{inversion: } t \text{ increases from } t_{\text{den}} \text{ to } T, \text{ starting } \hat{x}_t = x_{t_{\text{den}}}}, \end{aligned} \tag{1}$$

where x_t and $\hat{x}_t$ are latent variables at timestep t, v_ϕ denotes the noise prediction network, σ_t controls the sampler parameter. For conditioning, we adopt parameters from (Bai et al., 2025): a high denoising strength $\gamma_{\text{den}} = 5$ ensures strong guidance, while an inversion strength of $\gamma_{\text{inv}} = 0$ is used to reduce distortion in the noise space. The total number of timesteps T is 999, with a predefined intermediate denoising timestep at $t_{\text{den}} = 652$. In equation 1, z_p provides the semantic information, while z_{SCat} provides visual information. Here, we introduce two learnable factors $\beta_1, \beta_2 \in \mathbb{R}_+$ to create a ***scale concatenation (SCat)*** of the input latents: $z_{\text{SCat}}(z_1, z_2; \beta_1, \beta_2) = \text{concat}(\beta_1 z_1, \beta_2 z_2)$.

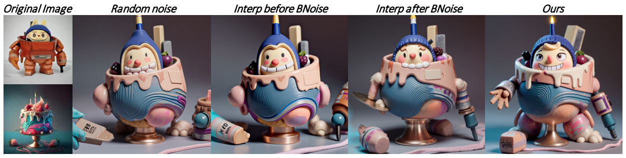


Figure 4: Different BNoise strategies.

Discussion on BNoise: concatenate vs. interpolate. We hypothesize that interpolating mismatched embeddings obscures subtle features, while concatenation preserves them, allowing the inversion process to refine noise containing the full concept. To test this, we compare *Interpolate before BNoise:* Blend embeddings first, then refine the noise, and *Interpolate after BNoise:* Refine noise from each embedding first, then blend the results. Fig. 4 shows that both interpolation methods fail to capture intricate details (e.g., legs), whereas our concatenation yields superior visual quality and faithfulness by preserving input details and ensuring a coherent denoising pathway. ***Quantitative results in Appdx. A***.

Mixing Denoise (MDeNoise): Using the blended noise ϵ_b, we denoise it to finally produces a cross-object fusion by mixing the inputs, z_1, z_2, z_p. Specifically, we formulate this process as:

$$I = \mathcal{D}(x_0),\ \text{where } x_0 \Leftarrow \overbrace{x_{t-1} = x_t - (\sigma_t - \sigma_{t-1})v_\phi(x_t, t, z_{\text{SInp}}(z_1, z_2; \alpha), \gamma_{\text{gen}}, z_p)}^{\text{MDeNoise: } t \text{ decreases from } T \text{ to } 0, \text{ starting } x_T = \epsilon_b}. \tag{2}$$

Here, $\gamma_{\text{gen}} = 4.0$ is a fixed guidance scale, and the decoder $\mathcal{D}(\cdot)$ generate the final fusion image I using the Flux-Krea decoder (Lee et al., 2025). The ***scale interpolation (SInp)***, $z_{\text{SInp}}(z_1, z_2; \alpha)$, mixes the two visual embeddings z_1 and z_2 into a single coherent representation, which is implemented by a spherical interpolation (Shoemake, 1985): $z_{\text{SInp}}(\alpha) = \frac{\sin(\alpha \cdot \delta)}{\sin(\delta)} z_1 + \frac{\sin((1-\alpha)\cdot\delta)}{\sin(\delta)} z_2$, where $\delta = \cos^{-1}(z_1 \cdot z_2)$, and $0 \leq \alpha \leq 1$ is a learnable factor to control the mixing ratio. This MDeNoise process in equation 2 outputs the final fusion image I.

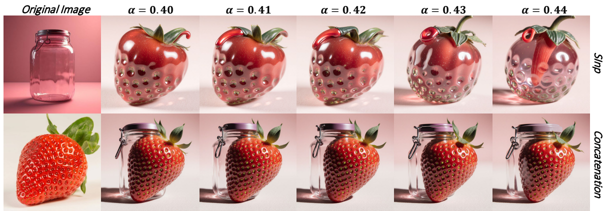


Figure 5: Different MDeNoise generations across α.

Discussion on MDeNoise: interpolate vs. concatenate. MDeNoise prioritizes fusing its two inputs, unlike BNoise which preserves them. While concatenation retains more input information, its rigid separation often creates disjointed representations and generations. However, interpolation enables seamless integration. To demonstrate this, we compare with a concatenation-fusion variant: z_{SInp} is replaced by $z_{\text{SCat}}(\alpha) = \text{concat}(\alpha z_1, (1-\alpha) z_2)$

in equation 2 (Fig. 5), which tends to produce isolated objects rather than a unified hybrid. Our interpolation instead creates a single, coherent entity with harmonious consistency.

HSP: Overall, for a given input D, the hybrid sampling process combines the BNoise (equation 1) and MDeNoise (equation 2). To simplify the notation, we formalize this process as the function:

$$I(\theta) = \mathrm{HSP}(D; \theta, \hat{\theta}) = \mathcal{D}(x_0), \tag{3}$$

where $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$ are learnable parameters, and $\hat{\theta} = \{\gamma_{\text{den}} = 5, \gamma_{\text{inv}} = 0, \gamma_{\text{gen}} = 4, T = 999, t_{\text{den}} = 652\}$ are fixed defaults in this paper.

3.2 Efficient Adaptive Adjustment (EAA)

The HSP process yields distinct fusion results $I(\theta)$ defined in equation 3 with parameters θ, defaults $\hat{\theta}$ and inputs D, making parameter selection critical for high-quality synthesis. We propose an adaptive framework to jointly adjust $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$, aiming to achieve both semantic coherence and visual fidelity. Inspired by prior work (Li et al., 2024; Xiong et al., 2024), we first introduce a **Similarity Score (SS)** to guide this search: (*For simplicity, input D and defaults $\hat{\theta}$ are not shown.*)

$$S(\theta) = \underbrace{S_{I_1}(\theta) + S_{I_2}(\theta)}_{\text{visual similarity}} + \underbrace{S_{T_1}(\theta) + S_{T_2}(\theta)}_{\text{semantic similarity}} - \underbrace{|S_{I_1}(\theta) - S_{I_2}(\theta)|}_{\text{visual balance}} - \underbrace{|S_{T_1}(\theta) - S_{T_2}(\theta)|}_{\text{semantic balance}}, \tag{4}$$

where $S_{I_i}(\theta)$ $(i = 1, 2)$ is the visual similarity between $I(\theta)$ and the source image I_i, computed via a DINO encoder (Oquab et al., 2024), while $S_{T_i}(\theta)$ $(i = 1, 2)$ is the semantic similarity between $I(\theta)$ and the category label T_i, measured using CLIP (Radford et al., 2021). This scoring function is designed to optimize two key objectives for successful fusion: (i) *maximizing similarity*, and (ii) *enforcing balance*. The first two terms ensure that the generated image $I(\theta)$ retains high perceptual and semantic fidelity to both input images and their corresponding category labels. By maximizing similarity to both sources, these terms preserve the core features of the original concepts. The final two terms—penalizing the absolute differences—explicitly enforce *balance*, preventing the model from overfitting to one input and encouraging a fair integration of both objects' features. Together, these components create a unified SS objective that balances fidelity and symmetry, offering a principled framework for optimizing feature fusion parameters.

Our EAA Algorithm. To maximize this objective $S(\theta)$ in equation 4, we present a hierarchical adjustment strategy that learns the parameters $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$ using the acceptance threshold $Th = 2.4$. The key loop iterates from $k = 1$ to $K = 3$, performing these steps:

① **Sample (initial) Gaussian noise:** $\epsilon \sim \mathcal{N}(0, I)$, **initialize the parameters:** $\alpha = 0.5, \beta_1 = \beta_2 = 1.0$.

② **Searching α:** Fixed $\beta_1 = \beta_2 = 1.0$ and ϵ, perform a golden section search (Teukolsky et al., 1992) to find the optimal mixing factor α^*:

$$\alpha^* = \arg\max_{\alpha \in [0,1]} S(\alpha, \beta_1, \beta_2, \epsilon). \tag{5}$$

③ **Adjusting β_1, β_2:** Fixed α^*, ϵ, if $S(\alpha^*, \beta_1, \beta_2, \epsilon) \leq Th$, then update the noise factors:

$$\begin{cases} \beta_1^* = \beta_1 \;\&\; \beta_2^* = \arg\max\limits_{\beta_2 \in \mathbb{R}_+} S(\alpha^*, \beta_1, \beta_2, \epsilon), & \text{if } S_1 > S_2, \\ \beta_2^* = \beta_2 \;\&\; \beta_1^* = \arg\max\limits_{\beta_1 \in \mathbb{R}_+} S(\alpha^*, \beta_1, \beta_2, \epsilon), & \text{otherwise.} \end{cases}, \tag{6}$$

where $S_1 = S_{I_1} + S_{T_1}$, $S_2 = S_{I_2} + S_{T_2}$, and $S_1 > S_2$ indicates that the mixing noise favors the object I_1, and vice versa.

④ **Acceptance criterion:**

$$\begin{cases} \epsilon^* = \epsilon \;\&\; \textbf{return } \theta^* = \{\alpha^*, \beta_1^*, \beta_2^*, \epsilon^*\}, & \text{if } S(\alpha^*, \beta_1^*, \beta_2^*, \epsilon) > Th, \\ \textbf{return } \theta^* = \{\alpha^*, \beta_1^*, \beta_2^*, \epsilon^*\} \;\&\; \textbf{break}, & \text{if } k > K, \\ \textbf{turn to the step ① to resample } \epsilon \;\&\; k{+}{+}, & \text{otherwise.} \end{cases}, \tag{7}$$

where the fused object image $I(\theta)$ is defined in equation 3. Our adaptive loop efficiently explores a low-dimensional yet expressive parameter space $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$, yielding conceptually balanced and perceptually smooth fusion results (Fig. 9). By reusing intermediate predictions and limiting optimization to scalar-level searches (via golden section search), the method enhances sample efficiency—avoiding the computational overhead of gradient-based latent-space backpropagation.

Discussion on resampling ϵ***.*** During our blending process, sampling random Gaussian noise can occasionally yield low-quality or failed fusions. While first-order optimization is an intuitive solution, it offers no significant advantage over simple zero-order resampling for diffusion generation, despite its higher cost (Ma et al., 2025). Consequently, we adopt a zero-order resampling strategy to search for ϵ, and a small number of resamples $K = 3$ proves sufficient for high-quality fusion. ***For fair comparison, this resampling is disabled,*** $K = 1$***, and the random seed is fixed at 42.***

4 Experiments

4.1 Experimental Settings

Datasets. We introduce IIOF (Image-Image Object Fusion), a new benchmark of 780 image pairs derived from 40 objects across four classes (i.e., animals, fruits, artificial objects, and character figurines). Most images are from PIE-Bench (Ju et al., 2024) and Pexels[1]; figurines were self-captured for quality. To evaluate order-sensitive methods, we also generate all ordered pairs (1,560 total), ensuring a comprehensive and fair benchmark. ***More details in Appdx. B.***

Implementation Details. Our method builds upon Flux-Krea (Lee et al., 2025), implementing $\mathcal{E}_I$ with Redux (Black Forest Labs, 2024) for latent-space alignment. We generate all images at 512×512 resolution using the FlowMatchEulerDiscreteScheduler (Lipman et al., 2022) with 20 denoising steps. For the Efficient Adaptive Adjustment (EAA) module, we use Grounded-SAM (Ren et al., 2024) and the query *"most prominent object"* to localize main regions for visual and semantic similarity computation. Each parameter search for α and β involves at most 10 image generations. All experiments are conducted on two NVIDIA RTX 4090 GPUs.

Evaluation Metrics. To evaluate our method, we use two metric families: Semantic Alignment (SA) and Single-entity Coherence (SCE). SA is computed on the generated prompt P_G using VQAScore (Lin et al., 2024b) and LLaVA-Critic (Xiong et al., 2025a). VQAScore employs CLIP-FlanT5 (Roberts et al., 2022) and LLaVA (Liu et al., 2023a), denoted as $\mathrm{VQA}^{\mathrm{SA}}_{\mathrm{T5}}$ and $\mathrm{VQA}^{\mathrm{SA}}_{\mathrm{LLaVA}}$, respectively; the LLaVA-Critic score is $\mathrm{LC}^{\mathrm{SA}}$. SCE assesses if the image forms a unified concept by asking: *"A photo of a seamless fusion of <*T_1*> and <*T_2*> into a single coherent entity."* Its scores are $\mathrm{VQA}^{\mathrm{SCE}}_{\mathrm{T5}}$, $\mathrm{VQA}^{\mathrm{SCE}}_{\mathrm{LLaVA}}$, and $\mathrm{LC}^{\mathrm{SCE}}$. We also compute the SS score and the balance metric $B_{\text{sim}} = |S_{I_1}(\theta) - S_{I_2}(\theta)| + |S_{T_1}(\theta) - S_{T_2}(\theta)|$, where $S_{T_i}(\theta)$ are normalized to $[0, 1]$ using empirical bounds 0.15 and 0.45 to align the scales of visual and textual modalities.

4.2 Main Results

We compare with leading methods across three categories: (i) multi-concept generation (e.g., OmniGen (Xiao et al., 2025), FreeCustom (Ding et al., 2024), MIP-Adapter (Huang et al., 2025), DreamO (Mou et al., 2025)), (ii) mixing-based (e.g., ATIH (Xiong et al., 2024), Conceptlab (Richardson et al., 2024), FreeBlend (Zhou et al., 2025)), and (iii) image editing (e.g., Stable Flow (Avrahami et al., 2025)). We also include qualitative results from GPT-4o (OpenAI, 2025). Inputs vary: multi-concept methods use two images and a text prompt; ATIH and Stable Flow use one image and text; Conceptlab uses text only. ***More examples in Appdx. G.***

Qualitative Comparison. Fig. 6 compares our method with multi-concept generation baselines (e.g., MIP-Adapter, OmniGen, DreamO, GPT-4o), highlighting two observations. First, baselines output often merely overlay features rather than fusing them—for example, *a lime enclosed in a glass jar without integration*—while our method creates a coherent hybrid. Second, baselines frequently favor one concept, such as generating either a doll or a corgi but not a unified blend. In contrast, our approach balances both concepts, producing structurally unified and semantically consistent results. This demonstrates our method's superior ability to achieve fine-grained visual fusion.

[1] `https://www.pexels.com/`

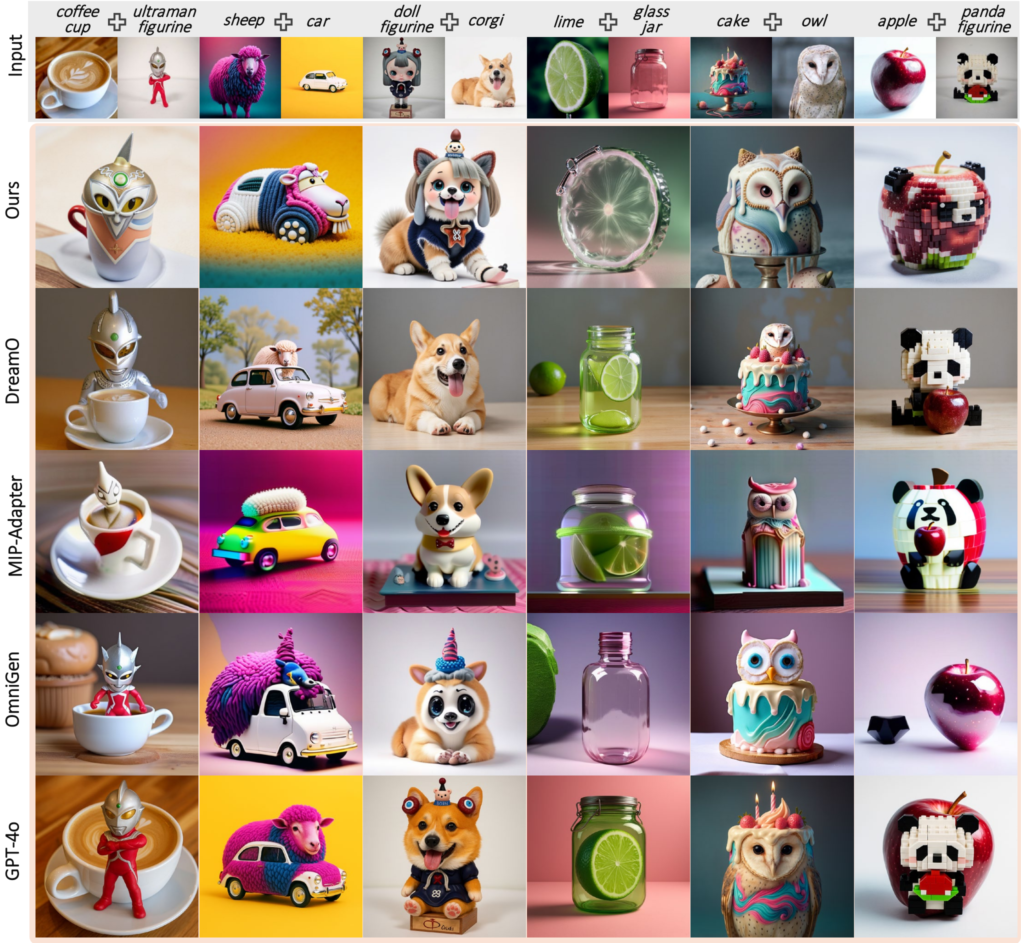


Figure 6: **Comparisons with Multi-Concept Generation Methods.** Our approach yields hybrid objects with improved structural coherence and visual balance over existing methods.

Table 1: Quantitative comparisons on our IIOF dataset.

Models	$\text{VQA}^{\text{SA}}_{\text{T5}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{T5}}\uparrow$	$\text{LC}^{\text{SA}}\uparrow$	$\text{LC}^{\text{SCE}}\uparrow$	$\text{VQA}^{\text{SA}}_{\text{LLaVA}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{LLaVA}}\uparrow$	$SS\uparrow$	$B\text{sim}\downarrow$
Our VMDiff	0.639	0.540	8.372	8.392	0.390	0.413	2.068	0.324
FreeCustom (CVPR (Ding et al., 2024))	0.579	0.452	6.958	6.946	0.360	0.388	1.580	0.776
MIP-Adapter (AAAI (Huang et al., 2025))	0.621	0.512	8.301	8.076	0.389	0.417	1.866	0.483
OmniGen (CVPR (Xiao et al., 2025))	0.570	0.469	7.550	7.233	0.352	0.348	1.705	0.617
Conceptlab (TOG (Richardson et al., 2024))	0.573	0.483	7.589	7.728	0.362	0.395	–	–
ATIH (NeurIPS (Xiong et al., 2024))	0.523	0.465	7.275	6.816	0.317	0.367	–	–
Stable Flow (CVPR (Avrahami et al., 2025))	0.460	0.372	6.020	5.024	0.266	0.294	–	–
DreamO (SIGGRAPH Asia (Mou et al., 2025))	0.591	0.467	7.592	7.013	0.370	0.346	1.793	0.644
FreeBlend (arXiv (Zhou et al., 2025))	0.588	0.507	7.836	7.788	0.341	0.383	1.870	0.479

Fig. 7 qualitatively compares our method with mixing/editing baselines (e.g., Conceptlab, ATIH, FreeBlend, Stable Flow). Conceptlab often biases toward one concept, while Stable Flow and ATIH make only subtle edits, such as color or texture transfer. FreeBlend frequently loses original information and yields fragmented outputs. In contrast, our approach synthesizes novel objects that structurally and visually integrate both concepts, achieving a deeper, more harmonious fusion and demonstrating superior blending capability.

Quantitative Comparison. Table 1 presents quantitative comparisons on key metrics, including $\text{VQA}^{\text{SA}}_{\text{T5}}$, $\text{VQA}^{\text{SA}}_{\text{LLaVA}}$, $\text{VQA}^{\text{SCE}}_{\text{T5}}$, $\text{VQA}^{\text{SCE}}_{\text{LLaVA}}$, LC^{SA}, LC^{SCE}, similarity score (SS), and fusion balance B_{sim}. Although MIP attains the highest $\text{VQA}^{\text{SCE}}_{\text{LLaVA}}$, it ranks only second or below on the other VQA, LC, SS, and B_{sim} metrics, indicating that its improvements are not holistic. In contrast, our method consistently outperforms all baselines on most metrics, demonstrating strong capability in generating coherent and natural blended objects. These results reinforce our qualitative findings and confirm the effectiveness of our approach in achieving high-quality visual fusion.

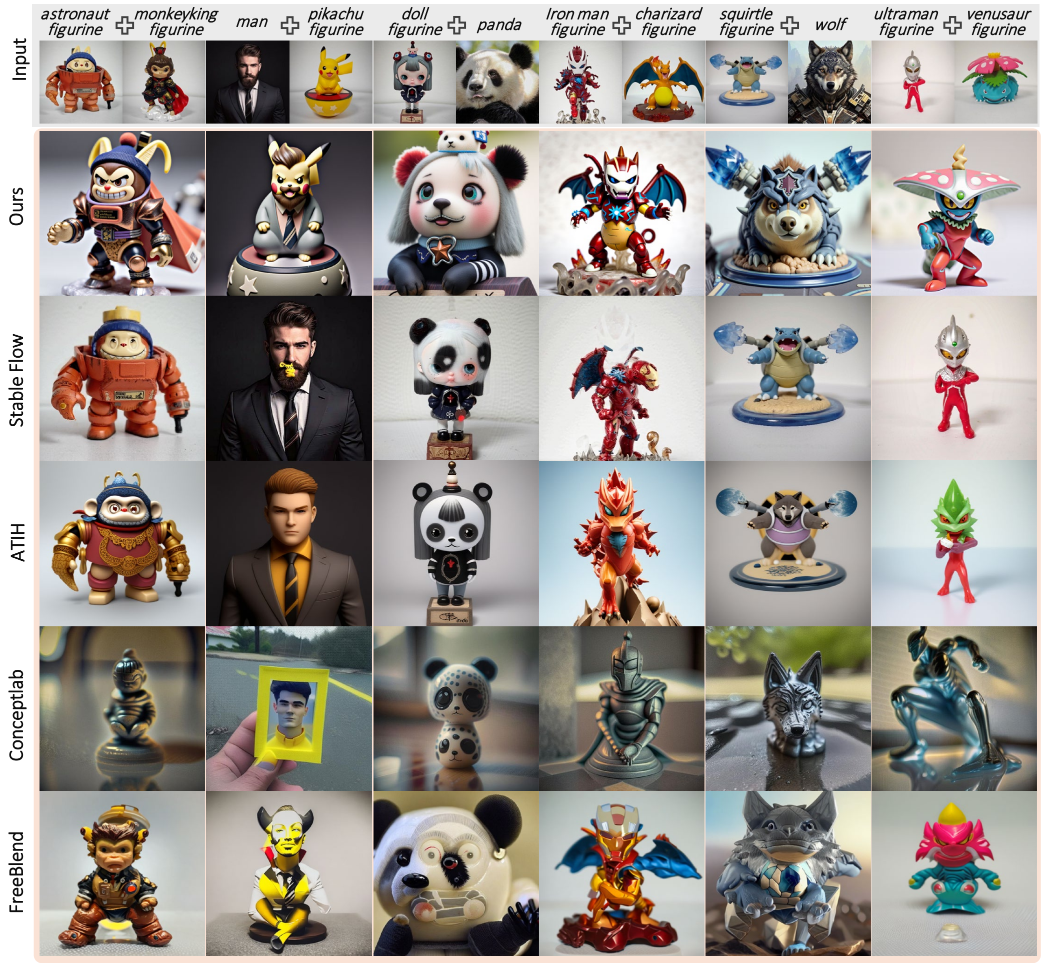


Figure 7: **Comparisons with Mixing and Image Editing Methods.** Our method produces more coherent and balanced hybrids, while baselines often favor one concept or apply minimal edits.

Table 2: Quantitative ablation study on our IIOF dataset.

Models	$\mathrm{VQA}^{\mathrm{SA}}_{\mathrm{T5}}\uparrow$	$\mathrm{VQA}^{\mathrm{SCE}}_{\mathrm{T5}}\uparrow$	$\mathrm{LC}^{\mathrm{SA}}\uparrow$	$\mathrm{LC}^{\mathrm{SCE}}\uparrow$	$\mathrm{VQA}^{\mathrm{SA}}_{\mathrm{LLaVA}}\uparrow$	$\mathrm{VQA}^{\mathrm{SCE}}_{\mathrm{LLaVA}}\uparrow$	$SS\uparrow$	$B\mathrm{sim}\downarrow$
Baseline 1	0.497	0.438	7.261	7.077	0.287	0.314	1.570	0.682
Baseline 2	0.508	0.441	7.426	7.291	0.298	0.325	1.586	0.693
Baseline 2+α-search	0.625	0.532	8.278	8.276	0.382	0.405	2.025	0.358
Baseline 2+α-search+β_1, β_2-search	**0.639**	**0.540**	**8.372**	**8.392**	**0.390**	**0.413**	**2.068**	**0.324**

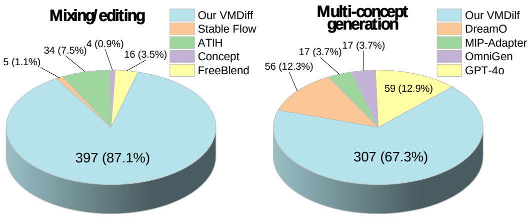


Figure 8: User studies.

User Study. To evaluate the perceptual quality of our fusions, we conducted two user studies (Fig. 8). 76 participants each rated 12 results—6 from *Multi-Concept Generation* and 6 from *Mixing/Editing*—yielding 912 total votes. Our VMDiff received the highest preference in both groups: **67.3%** and **87.1%**, respectively. GPT-4o and ATIH ranked second, but with significantly lower votes (12.9% and 7.5%). These results indicate that our VMDiff aligns better with human preferences in visual coherence and creativity. ***More details in Appdx. C.***

4.3 Ablation Study

We conducted an ablation study to evaluate the contributions of our VMDiff's key components, as shown in Fig. 9 and Table 2. Progressively adding each element—**(i)** *baseline 1:* random

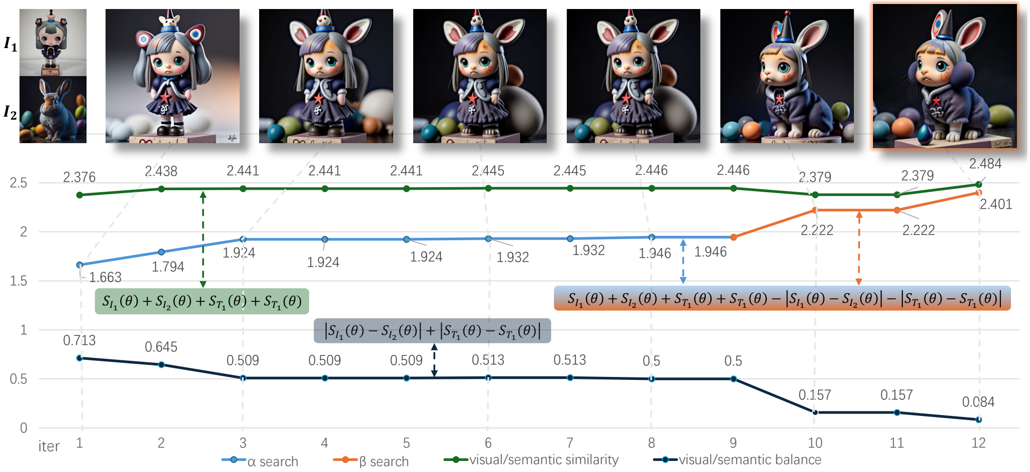


Figure 10: **Visualizing the updated process of our EAA** based on two input images I_1 (*doll figurine*) and I_2 (*rabbit*). The α parameter (blue) improves fusion quality, while β (orange) enhances semantic balance. The green curve (similarity) rises and the dark blue curve (imbalance) falls over iterations. The final output is a coherent hybrid with high similarity and minimal imbalance.

noise+MDeNoise ($\alpha = 0.5$), **(ii)** *baseline 2:* baseline 1+BNoise ($\beta_1 = \beta_2 = 1$), **(iii)** baseline 2 + MDeNoise (α search), and **(iv)** baseline 2 + BNoise (β_1, β_2 search) + MDeNoise (α search)—yielded consistent improvements. Without noise refinement, outputs lacked detail. Its inclusion enhanced structural fidelity and preserved input features. Adaptive α improved fusion balance, while adaptive β refined noise influence for greater visual harmony. Fig. 10 illustrates the optimization process for a representative case (*doll figurine* + *rabbit*). Throughout iterations, similarity $S(\theta)$ (green) increased steadily, while the blending balance metric (dark blue) decreased. The α search (light blue) rapidly boosted similarity, and β search (orange) smoothed visual-textual alignment. These results confirm that our EAA design effectively optimizes both similarity and symmetry for high-quality blending. ***Limitations are discussed in Appdx. D.***

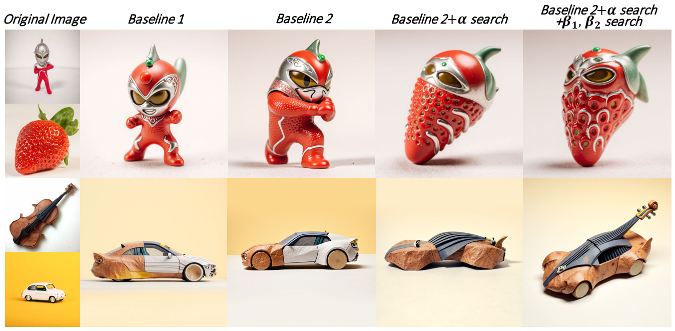


Figure 9: **Ablation study in VMDiff.** *Noise refinement* improves detail and structure, while *adaptive α and β search* progressively enhance semantic balance and visual coherence.

4.4 MULTI-IMAGE FUSION AND BACKBONE GENERALIZATION.

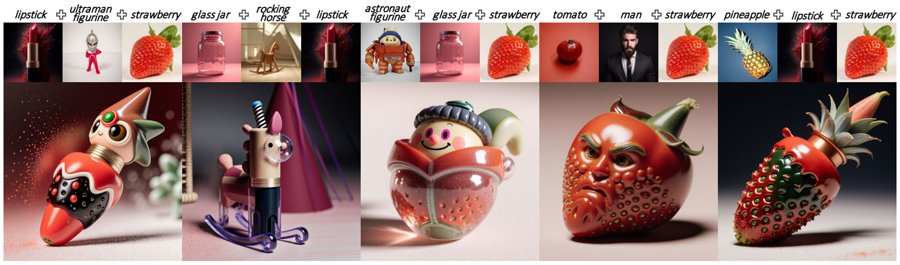

lipstick ✚ ultraman figurine ✚ strawberry
glass jar ✚ rocking horse ✚ lipstick
astronaut figurine ✚ glass jar ✚ strawberry
tomato ✚ man ✚ strawberry
pineapple ✚ lipstick ✚ strawberry

Figure 11: **Multi-image fusion.**

Multi-image fusion. We also explore extending VMDiff beyond pairwise fusion. Figure 11 shows preliminary three-image results obtained by sequentially applying our pipeline (e.g., first fusing (I_1, I_2) and then fusing the hybrid with I_3). The method can still produce single coherent entities that blend attributes from all three categories, indicating that our formulation can, in principle, scale to more inputs. However, compared with the pairwise case, these hybrids exhibit stronger information loss and imbalance across sources, so in this work we focus on image pairs and leave permutation-invariant, learned aggregation of multiple image embeddings to future work.

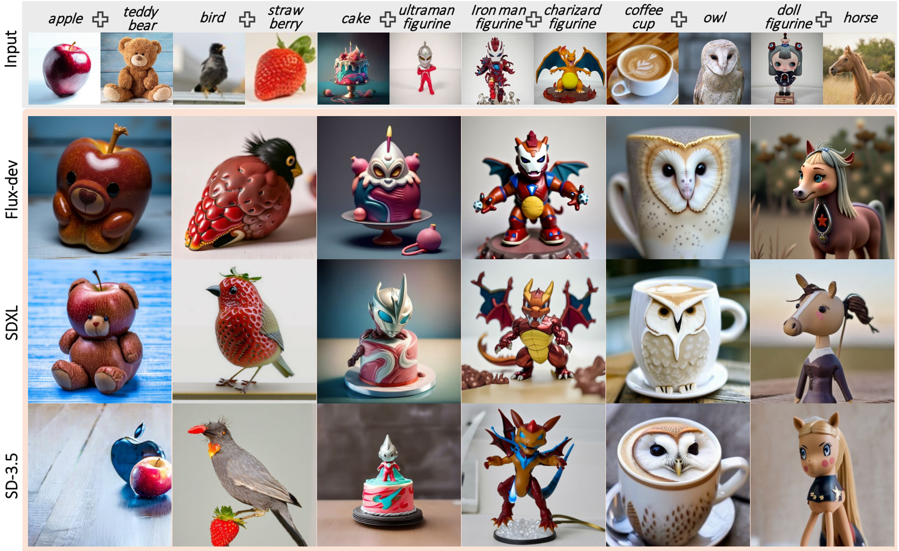


Figure 12: **Each column shows an input pair (top) and the fused outputs when plugging our framework into Flux-1.0-dev+Redux, SDXL+IP-adapter(Ye et al., 2023), and SD-3.5+SD-3.5-IP-adapter(Team, 2024). All backbones use the same images and fusion prompt.**

Backbone generalization. We evaluate VMDiff on three backbones with identical settings: Flux-dev+Redux, SDXL (Lin et al., 2024a)+IP-Adapter (Ye et al., 2023), and SD-3.5 (AI, 2024)+SD-3.5-IP-Adapter (Team, 2024) (Fig. 12). All three run under our HSP+EAA framework, so VMDiff is not tied to Flux-Krea, but the *quality and tendency toward a single hybrid object* strongly depend on how image information is encoded. Flux+Redux maps images into a semantic latent space shared with text, allowing BNoise+SInp to operate directly on rich, text-like image embeddings and thus best preserve instance-level geometry and appearance from both sources. For SDXL and especially SD-3.5, IP-Adapter injects image features as extra attention tokens; interpolating these tokens mainly modulates high-level semantics and, in our results, often weakens retention of input-specific structure. VMDiff therefore benefits most from backbones that preserve detailed instance information in a text-compatible embedding space. Our SDXL and SD-3.5 experiments should be viewed as feasibility checks under this weaker image interface, and we expect that adding Redux-style semantic image encoders to such models would narrow the quality gap to Flux-Krea.

5 CONCLUSION

In this paper, we presented VMDiff, a novel unified and controllable framework for visual concept fusion that synthesizes coherent new objects directly from two input images. Our approach enables fine-grained control by semantically integrating concepts at both the noise and latent levels. VMDiff consists of two core components: (1) a hybrid sampling process that constructs optimized semantic noise through guided denoising and inversion, followed by a curvature-aware latent fusion using spherical interpolation, and (2) an efficient adaptive adjustment algorithm that refines fusion parameters via a lightweight, score-driven search. Experimental results on a curated benchmark demonstrate VMDiff's superior performance, excelling in semantic consistency, visual harmony, and user-rated creativity, thereby establishing a new paradigm for hybrid object synthesis. This work offers practical and valuable insights for professionals developing combinational characters, directly applicable to diverse fields from film and animation to figures and industrial design.

References

Stability AI. Sd3.5 – inference-only reference implementation. https://github.com/Stability-AI/sd3.5, 2024.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7877–7888, 2025.

Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed: 2025-05-07.

Margaret A Boden. *The creative mind: Myths and mechanisms.* Routledge, 2004.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23206–23217, 2023.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21476–21485, 2024.

Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9089–9098, 2024.

Xiaoyue Dong and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7430–7440, 2023.

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)*, 2023.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 36:15890–15902, 2023.

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)*, pp. 7323–7334, 2023.

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19740–19750, 2023.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3707–3714, 2025.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (ICCV)*, pp. 14173–14182, 2023.

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)*, pp. 1931–1941, 2023.

Sangwu Lee, Titus Ebbecke, Erwann Millon, Will Beddow, Le Zhuo, Iker García-Ferrero, Liam Esparraguera, Mihai Petrescu, Gian Saß, Gabriel Menezes, and Victor Perez. FLUX.1 Krea [dev]. `https://github.com/krea-ai/flux-krea`, 2025.

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)*, pp. 92–111, 2024.

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

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024a.

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)*, pp. 366–384, 2024b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 34892–34916, 2023a.

Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23166–23178, 2021.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 423–439, 2022.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8599–8608, 2024.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: customizable image synthesis with multiple subjects. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 57500–57519, 2023b.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534, 2025.

Mary Lou Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pp. 22–28, 2010.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.

OpenAI. Chatgpt: Optimizing language models for dialogue. 2025. URL `https://www.openai.com`. Accessed: 2025-05-07.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

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 the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

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

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8871–8879, 2024.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

H. Tang, L. Yu, and J. Song. Master: Meta style transformer for controllable zero-shot and few-shot artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

InstantX Team. Instantx sd3.5-large ip-adapter page, 2024.

Saul A Teukolsky, Brian P Flannery, W Press, and W Vetterling. Numerical recipes in c. *SMR*, 693 (1):59–70, 1992.

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *Proceedings of the SIGGRAPH Asia 2024 Conference Papers*, 2024.

Renke Wang, Guimin Que, Shuo Chen, Xiang Li, Jun Li, and Jian Yang. Creative birds: Self-supervised single-view 3d style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8775–8784, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304, 2025.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 13618–13628, 2025a.

Zeren Xiong, Ze dong Zhang, Zikun Chen, Shuo Chen, Xiang Li, Gan Sun, Jian Yang, and Jun Li. Novel object synthesis via adaptive text-image harmony. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 139085–139113, 2024.

Zeren Xiong, Zikun Chen, Zedong Zhang, Xiang Li, Ying Tai, Jian Yang, and Jun Li. Category-aware 3d object composition with disentangled texture and shape multi-view diffusion. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2025b.

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

Xulu Zhang, Xiao-Yong Wei, Jinlin Wu, Tianyi Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Compositional inversion for stable diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, number 7, pp. 7350–7358, 2024.

Y. Zhang, L. Wang, and H. Li. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10146–10156, 2023.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25912–25921, 2024.

Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26384–26395, 2024.

Yufan Zhou, Haoyu Shen, and Huan Wang. Freeblend: Advancing concept blending with staged feedback-driven interpolation diffusion. *arXiv preprint arXiv:2502.05606*, 2025.

Xiandong Zou, Mingzhu Shen, Christos-Savvas Bouganis, and Yiren Zhao. Cached multi-lora composition for multi-concept image generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Supplementary Materials

This supplementary material provides additional technical details and extended results to support the main paper. We begin in **Section A** with two key discussions: the necessity of adjusting β_1 and β_2 in our hierarchical parameter search, and a quantitative comparison of BNoise fusion strategies—concatenation versus interpolation. **Section B** describes the construction of our proposed IIOF benchmark dataset, including the criteria for category selection and object pairing strategies. **Section C** presents a comprehensive user study, providing human preference validation of our fusion results. In **Section D**, we outline the current limitations of our method, discuss remaining challenges, and suggest possible directions for future improvement. **Section E** contains our formal statement on the use of LLMs in this work, in accordance with ICLR policy. **Section F** details the full inference pipeline of our VMDiff framework. Finally, **Section G** showcases extensive qualitative results, further demonstrating the effectiveness and generalization ability of our method across diverse fusion scenarios.

A Additional discussions

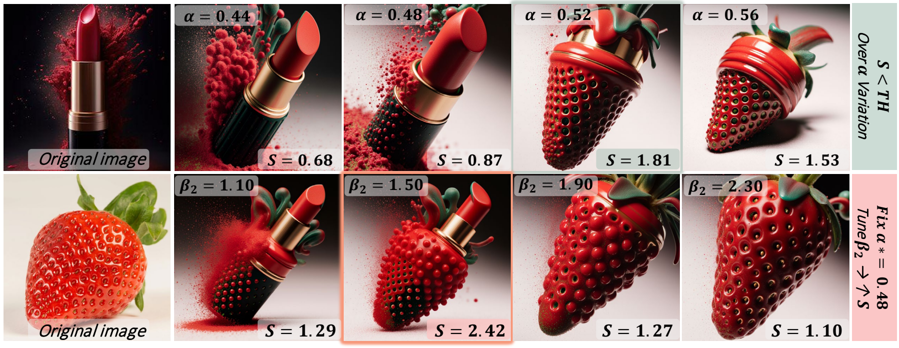


Figure 13: Illustration of our hierarchical parameter adjustment. The top row shows results from searching α; the bottom row refines the fusion by fixing α and adjusting β_2. **Consistent with Sec. 3.2, once the overall score S exceeds the acceptance threshold $T_h = 2.4$, the fusion becomes visually coherent and balanced**; when α-only optimization underperforms, the second-stage β_2 refinement raises S above the threshold.

Discussion on the necessity of adjusting β_1, β_2**.** As shown in Fig. 13, global optimization over α alone occasionally fails to yield well-fused results. To mitigate this, we first fix α^* (corresponding to the best similarity score in Eq. 4) and then perform a local refinement by optimizing β_1, β_2. This adjustment allows the model to precisely calibrate the noise contribution of each object, enhancing both visual coherence and semantic balance in the final output.

Discussion on BNoise. As shown in Table 3 on the IIOF dataset, Ours (Concat before inversion) achieves state-of-the-art performance on most metrics. Although it ranks second on the LC metric, its substantial advantage on SS demonstrates that concatenation more effectively preserves and integrates complementary information from both inputs. In summary, concatenation before inversion yields superior visual quality and semantic faithfulness by retaining fine-grained details and guiding a more coherent denoising pathway, compared with either form of interpolation.

Table 3: Quantitative Evaluation of BNoise Fusion: Concatenation vs. Interpolation.

Models	$\text{VQA}^{\text{SA}}_{\text{T5}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{T5}}\uparrow$	$\text{LC}^{\text{SA}}\uparrow$	$\text{LC}^{\text{SCE}}\uparrow$	$\text{VQA}^{\text{SA}}_{\text{LLaVA}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{LLaVA}}\uparrow$	$SS\uparrow$	$B\text{sim}\downarrow$
Random noise	0.497	0.438	7.261	7.077	0.287	0.314	1.570	0.682
Interp Before Inversion	0.504	0.441	**7.439**	**7.390**	0.293	0.321	1.551	**0.678**
Interp After Inversion	0.486	0.430	7.278	7.112	0.283	0.311	1.532	0.712
Ours(Concat Before Inversion)	**0.508**	**0.442**	7.426	7.291	**0.298**	**0.325**	**1.586**	0.693

Discussion on additional ablation of BNoise and the α/β ***search.*** To better understand the contributions of BNoise and the EAA search, we conduct an additional ablation on 1,184 pairs from IIOF, summarized in Fig. 14 and Table 4. We compare four variants: (i) **Baseline 1**, which uses random noise plus MDeNoise with a fixed $\alpha = 0.5$ (no BNoise, no search); (ii) **Baseline 2**, which augments Baseline 1 with BNoise by setting $\beta_1 = \beta_2 = 1$ (semantic noise injected, no search); (iii) **Baseline 1**

+ β_1, β_2-search, which augments Baseline 1 with BNoise and an EAA search over (β_1, β_2); and (iv) **Random noise + α-search**, which applies EAA only to α without BNoise.

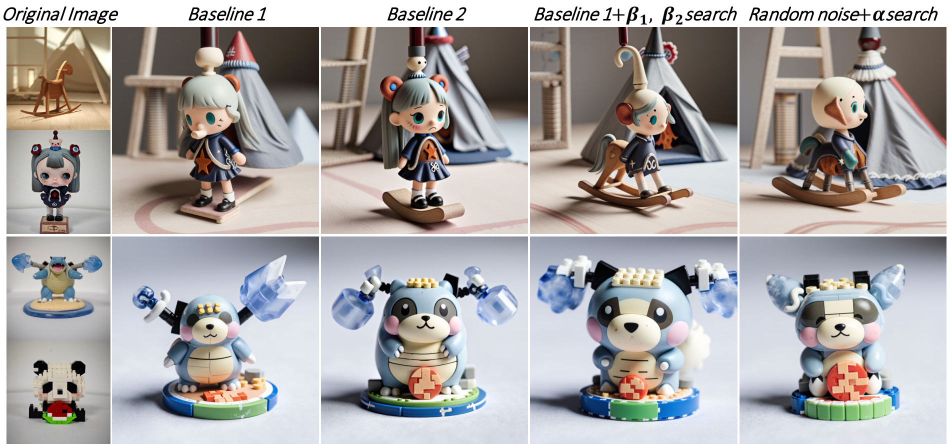

Original Image
Baseline 1
Baseline 2
Baseline 1+β_1, β_2 search
Random noise+α search

Figure 14: **Ablation of BNoise and the α/β search.** Each column shows the original image pair (left) and fused results from different variants: *Baseline 1* (random noise + MDeNoise with fixed α=0.5), *Baseline 2* (Baseline 1 + BNoise with β_1=β_2=1), *Baseline 1 + β-search*, and *Random noise + α-search*. BNoise (columns 2–3) provides a more informative initialization that preserves structures from both sources, while β-search further balances semantic content in the noise. In contrast, random-noise + α-search alone often loses details, confirming the complementary roles of BNoise and the α/β search.

Table 4: Quantitative ablation of BNoise and the α/β search on the IIOF dataset.

Models	$\text{VQA}^{\text{SA}}_{\text{T5}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{T5}}\uparrow$	$\text{LC}^{\text{SA}}\uparrow$	$\text{LC}^{\text{SCE}}\uparrow$	$\text{VQA}^{\text{SA}}_{\text{LLaVA}}\uparrow$	$\text{VQA}^{\text{SCE}}_{\text{LLaVA}}\uparrow$	$SS\uparrow$	$B\text{sim}\downarrow$
Baseline 1	0.496	0.419	7.186	7.065	0.283	0.320	1.563	0.691
Baseline 2	0.503	0.420	7.326	7.191	0.290	0.326	1.580	0.705
Baseline 1+β_1, β_2-search	0.553	0.461	7.723	7.679	0.320	0.359	1.760	0.553
Random noise+α-search	0.603	0.508	8.009	8.017	0.357	0.394	1.972	0.354

Qualitatively, Baseline 1 often loses information from the sources, whereas Baseline 2 preserves more structures from both inputs, confirming that the semantic noise ϵ_b obtained via the denoise–invert cycle provides a more informative initialization than pure Gaussian noise. Adding β-search on top of BNoise further improves all SA/SCE and SS scores and reduces the imbalance metric B_{sim} from 0.691 to 0.553, indicating that (β_1, β_2) effectively rebalance how each source contributes to the noise. The random-noise+α-search variant achieves higher SA/SCE and lower B_{sim} than Baseline 1, but still misses fine details and parts from the inputs, as seen in Fig. 14, due to the lack of a semantically informed noise initialization. Taken together, these results highlight complementary roles: BNoise produces a conditional, information-carrying noise ϵ_b, while the EAA search over α and (β_1, β_2) adjusts the contributions of the two sources in the mixed embeddings and in the noise, respectively. This motivates our full HSP+EAA design, which combines both components to obtain the most faithful and balanced hybrids.

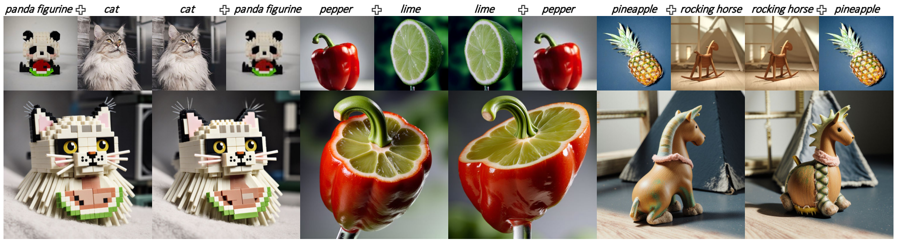


Figure 15: **Effect of swapping the order of T_1 and T_2 in the prompt.**

Discussion on name order. Fusion is, in principle, sensitive to the order of the category names in the guiding prompt, since the text encoder need not be strictly commutative. To probe this effect, we fix the image pair (I_1, I_2) and all hyperparameters, and only swap the order of the category tokens T_1

and T_2 in the prompt, using the "Random noise + MDeNoise ($\alpha = 0.5$)" baseline (Fig. 15). For most pairs (left and middle examples), the two orderings produce almost identical hybrids, indicating that our fusion behaves approximately symmetric with respect to name order. In a few harder cases (right), the leading token receives slightly more emphasis and extra attributes may appear, but both generations remain single, coherent hybrids rather than collapsing to one source. In the full VMDiff pipeline, this mild asymmetry is further reduced by the symmetric fusion score $S(\theta)$ and EAA search, which explicitly discourage strong bias toward a single category.

Discussion on fusion strategy. To better understand why we favor interpolation over concatenation, we also test a weighted concatenation variant $z_{\text{cat}}(\alpha) = \text{concat}(\alpha z_1,\ (1-\alpha)z_2)$, and fix the source images while varying α from 0.1 to 0.8 (Fig. 16). As the figure shows, relatively large changes in α are required to noticeably alter the result, confirming that α exerts much weaker control in the concatenation space than in the interpolated space. More importantly, across all settings the fusion remains *stitching-like*: one region of the image is dominated by the strawberry and the other by the jar, with a clear boundary between them. This suggests that separating z_1 and z_2 into distinct blocks encourages the network to treat them as two pieces to be glued together, rather than a single coherent object. In contrast, our MDeNoise stage mixes z_1 and z_2 via spherical interpolation within the *same* latent subspace, leading to much more integrated hybrids with smoothly shared geometry and appearance (see Fig. 21). These observations support our choice of slerp-based mixing in MDeNoise rather than concatenation-based fusion.

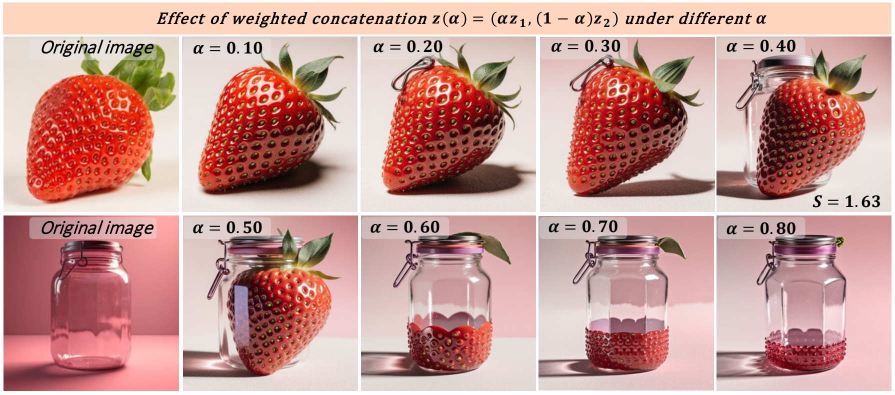


Figure 16: **Behaviour of weighted concatenation $z_{\text{cat}}(\alpha) = \text{concat}(\alpha z_1, (1-\alpha)z_2)$ under different α.** We fix the source images and vary α from 0.1 to 0.8. Large changes in α are required to noticeably alter the result, and across all settings the fusion remains stitching-like: one region is dominated by the strawberry and the other by the jar, with a clear boundary between them.

Why MIP-Adapter scores higher on SA/SCE. At first glance, MIP-Adapter appears visually weaker than DreamO and OmniGen, yet it achieves higher SA/SCE scores in Table 1. This is because our metrics are explicitly designed to measure *semantic fusion quality* rather than photo-realism. SA and SCE are LMM-based scores that reward (i) strong alignment with the fusion prompt and (ii) the presence of a *single* fused entity that simultaneously reflects both source categories. As illustrated in Fig. 17, MIP-Adapter typically produces one coherent object that clearly contains cues from both inputs, even though many fine-grained instance details are washed out. DreamO and OmniGen, on the other hand, often generate highly realistic and aesthetically pleasing images, but they frequently either omit one concept or render two separate objects instead of a single hybrid. Such behaviours are explicitly penalized by SA/SCE (and SS), which explains why MIP-Adapter scores higher in Table 1 despite being less visually appealing than DreamO and OmniGen in Fig. 17 and receiving lower user preference in Table 6.

B Datasets

To systematically evaluate our fusion framework, we construct a comprehensive benchmark dataset named **IIOF** (Image-Image Object Fusion), specifically tailored for assessing diverse and semantically rich visual concept mixing.

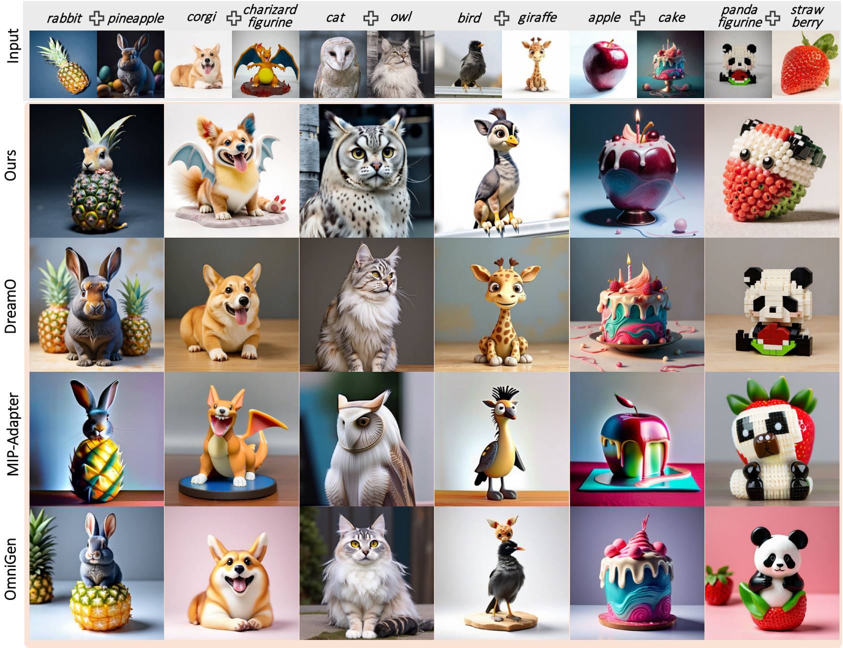


Figure 17: **Additional qualitative comparisons on IIOF.**

We meticulously selected **40 distinct object categories**, strategically organized into four semantic groups: *Animals*, *Fruits*, *Artificial Objects*, and *Character Figurines*. Each group comprises 10 unique classes, a design choice that ensures both intra-group consistency and ample inter-group diversity. A complete list of all selected categories is provided in Table 5.

For each chosen class, we sourced one high-quality, representative image. The majority of these images were obtained from established public benchmarks such as PIE-Bench (Ju et al., 2024) and popular stock image platforms like Pexels[2]. Recognizing the scarcity of high-quality, publicly available data for character figurines, we self-captured these images under controlled conditions, ensuring consistent lighting and resolution to maintain visual quality and diversity across the dataset. Figure 18 showcases all the selected images, providing a visual overview of the dataset's content. Additionally, each selected image is paired with its corresponding **textual category name**, as detailed in Table 5, to facilitate evaluations for prompt-based fusion methods. Initially, we derived **780 unique image pairs** by combining each of the 40 objects with every other object once, without considering input order. However, to ensure a comprehensive evaluation and enable fair comparison across all methods, particularly those sensitive to input order (e.g., ATIH (Xiong et al., 2024)), we further expanded IIOF to include **all possible ordered pairs** among the 40 categories. This expansion yielded a total of **1,560 image pairs**, where each combination (A, B) is present alongside its reverse (B, A). This exhaustive pairing strategy allows us to rigorously assess fusion performance across a wide spectrum of semantic relationships—ranging from semantically close concepts to challenging distant combinations, such as fusing a ’violin’ with a ’panda’ or a ’horse’ with ’lipstick’. This also critically highlights our model’s ability to generalize and compose novel concepts effectively across diverse domains.

Table 5: List of Objects in the IIOF Dataset by Category.

Category	**Object Names**
Animals	wolf, panda, owl, rabbit, horse, giraffe, corgi, cat, bird, sheep
Fruits	apple, orange, strawberry, durian, lime, pear, pineapple, watermelon, tomato, pepper
Artificial Objects	lipstick, violin, coffee cup, rocking horse, glass jar, car, teapot, cake, man, teddy bear
Character Figurines	iron man figurine, monkey king figurine, doll figurine, pikachu figurine, charizard figurine, ultraman figurine, astronaut figurine, venusaur figurine, panda figurine, squirtle figurine

[2] `https://www.pexels.com/`

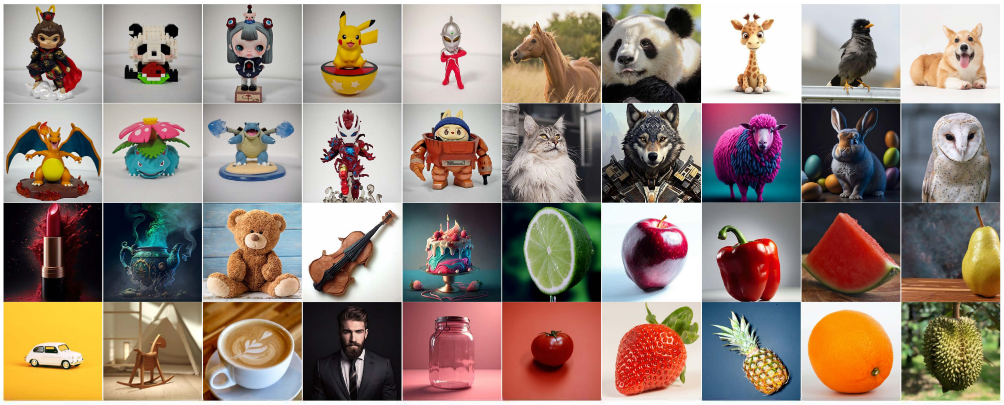
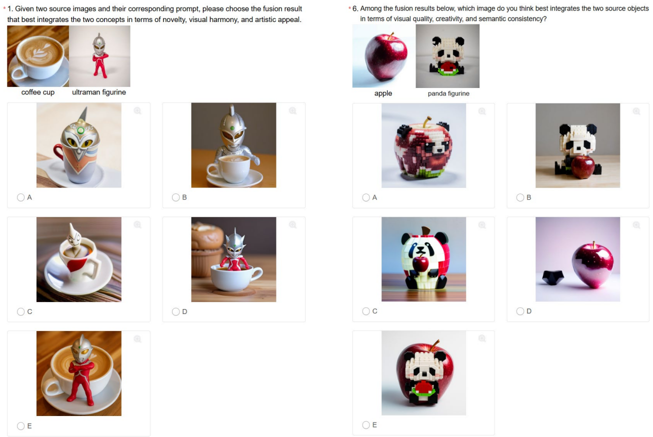


Figure 18: Original Object Image Set.

C User Study

To evaluate the perceptual quality and human preference for the novel images generated by our fusion framework, we conducted two user studies. These studies assessed our method, **VMDiff**, against state-of-the-art baselines in two main categories: *Multi-Concept Generation* methods and *Mixing and Image Editing* methods. The overall vote distributions are visualized in Fig. 8, while detailed per-example preferences are presented in Table 6 and Table 7. An example user study question for the *Multi-Concept Generation* group and the *Mixing and Image Editing* group is provided in Fig. 19. A total of 76 participants completed the survey, each evaluating 12 fused results (6 from each group), contributing a total of 912 votes. Participants were asked to select the fusion result that best integrated the given concepts in terms of visual quality, creativity, and semantic consistency. As shown in Fig. 8, our method consistently received the highest number of votes in both evaluation groups. In the *Mixing and Image Editing* category (left pie chart), VMDiff garnered a significant **397 votes (87.1%)** of the total. This considerably surpassed other methods such as Stable Flow (Avrahami et al., 2025) (5 votes, 1.1%), ATIH (Xiong et al., 2024) (34 votes, 7.5%), Conceptlab (Richardson et al., 2024) (4 votes, 0.9%) and FreeBlend (Zhou et al., 2025) (16 votes, 3.5%). For instance, as illustrated in Fig. 19, for the "astronaut figurine-monkey king figurine" fusion, our method obtained 81.58% of the votes, demonstrating its strong capability in seamlessly integrating distinct visual elements.



* 1. Given two source images and their corresponding prompt, please choose the fusion result that best integrates the two concepts in terms of novelty, visual harmony, and artistic appeal.

* 6. Among the fusion results below, which image do you think best integrates the two source objects in terms of visual quality, creativity, and semantic consistency?

Figure 19: An example of a user study comparing various multi-concept generation, mixing and image editing methods.

In the *Multi-Concept Generation* category (right pie chart), **VMDiff** led with **307 votes (67.3%)**, significantly outperforming GPT-4o (OpenAI, 2025), which ranked second with 59 votes (12.9%). Other baselines—DreamO (56 votes, 12.3%), MIP-Adapter (17 votes, 3.7%), and OmniGen (17 votes, 3.7%)—received notably fewer votes. In the "doll figurine–corgi" case, VMDiff earned **78.95%** of preferences. Even in more challenging cases like "apple–panda figurine" (see Fig. 19), it maintained an edge with **75.00%** over GPT-4o's **5.26%**. These results indicate that **VMDiff** better aligns with human preferences for visual coherence, creativity, and concept integration, consistently outperforming existing methods across diverse fusion scenarios.

Table 6: User study with multi-concept generation methods.

image-image	A(Our VMDiff)	B(DreamO)	C(MIP-Adapter)	D(OmniGen)	E(GPT-4o)
coffee cup-ultraman figurine	43(56.58%)	11(14.47%)	7(9.21%)	3(3.95%)	12(15.79%)
sheep-car	57(75.00%)	4(5.26%)	1(1.32%)	2(2.63%)	12(15.79%)
doll figurine-corgi	60(78.95%)	1(1.32%)	3(3.95%)	3(3.95%)	9(11.84%)
lime-glass jar	45(59.21%)	22(28.95%)	1(1.32%)	0(0.00%)	8(10.53%)
cake-owl	45(59.21%)	5(6.58%)	3(3.95%)	9(11.84%)	14(18.42%)
apple-panda figurine	57(75.00%)	13(17.11%)	2(2.63%)	0(0.00%)	4(5.26%)

Table 7: User study with mixing and image editing methods.

image-image	A(Our VMDiff)	B(Stable Flow)	C(ATIH)	D(Conceptlab)	E(FreeBlend)
astronaut figurine-monkey king figurine	62(81.58%)	2(2.63%)	7(9.21%)	1(1.32%)	4(5.26%)
man-pikachu figurine	68(89.47%)	0(0.00%)	4(5.26%)	1(1.32%)	3(3.95%)
doll figurine-panda	62(81.58%)	0(0.00%)	13(17.11%)	1(1.32%)	0(0.00%)
iron man figurine-charizard figurine	69(90.79%)	3(3.95%)	3(3.95%)	0(0.00%)	1(1.32%)
squirtle-wolf	66(86.84%)	0(0.00%)	4(5.26%)	1(1.32%)	5(6.58%)
ultraman figurine-venusaur figurine	70(92.11%)	0(0.00%)	3(3.95%)	0(0.00%)	3(3.95%)

D LIMITATIONS

Our method effectively fuses two input images into a coherent hybrid object that captures broad conceptual information; however, it has two main limitations. First, inference relies on iterative optimization, which increases computational cost and latency (Table 8). A promising remedy is to train a lightweight prediction/refinement module that guides the fusion in a single forward pass, thereby reducing runtime while maintaining—or even improving—visual quality and semantic balance. Second, in a small fraction of cases the fused outputs do not fully align with human preferences (Fig. 20), exhibiting semantic inconsistencies or stylistic imbalance. Although repeated noise resampling and selection can mitigate these failures, this heuristic has limited controllability. In future work, we will pursue more controllable, preference-aligned fusion via explicit human feedback, aesthetic priors, or learned alignment objectives, enabling results that more reliably reflect human intent and aesthetics.

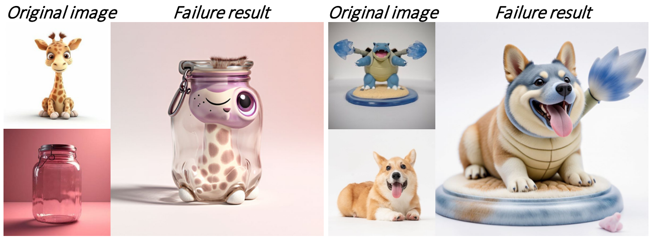


Figure 20: Examples of failure cases where our method produces fused outputs with suboptimal semantic or stylistic coherence.

Table 8: Runtime comparison across methods.

Methods	**Avg. Time / Pair**
Ours	2 min 46 sec
ATIH	10 sec
Stable Flow	27 sec
Conceptlab	13 min 45 sec
FreeCustom	22 sec
OmniGen	53 sec
Freeblend	12 sec
MIP-Adapter	12 sec
DreamO	8 sec

E STATEMENT ON LLM USAGE

In accordance with the ICLR policy on the use of Large Language Models (LLMs), we hereby declare that an LLM (ChatGPT, GPT-5) was used solely to aid or polish the writing of this paper, such as improving grammar and wording. All ideas, technical content, and experimental results are entirely our own. Further details are described within the paper. The authors take full responsibility for the accuracy and integrity of the content.

Algorithm 1: VMDiff with Efficient Adaptive Adjustment (VMDiff-EAA)

```
Input: images $I_1, I_2$, labels $T_1, T_2$, prompt $P_G$, threshold $TH$, max rounds $K$
Output: fused image $I^*$ and parameters $\theta^* = \{\alpha^*, \beta_1^*, \beta_2^*, \epsilon^*\}$
Compute embeddings $z_1 = \mathcal{E}_I(I_1),\ z_2 = \mathcal{E}_I(I_2),\ z_p = \mathcal{E}_T(P_G)$;
Initialize $\alpha = 0.5,\ \beta_1 = \beta_2 = 1.0;\quad S_{\text{best}} = -\infty,\ \theta_{\text{best}} = \varnothing$;
for $k = 1$ to $K$ do
    Sample noise $\epsilon \sim \mathcal{N}(0, I)$;
    $z_{\text{SCat}} = \text{concat}(\beta_1 z_1, \beta_2 z_2)$, $x_T = \epsilon$;
    for $t = T$ to $t_{den}$ do
        $x_{t-1} = x_t - (\sigma_t - \sigma_{t-1}) v_\phi(x_t, t, z_{\text{SCat}}, \gamma_{\text{den}}, z_p)$
    for $t = t_{den}$ to $T$ do
        $x_{t+1} = \hat{x}_t + (\sigma_{t+1} - \sigma_t) v_\phi(\hat{x}_t, t, z_{\text{SCat}}, \gamma_{\text{inv}}, z_p)$
    $\epsilon_r = \hat{x}_T$;
    $\alpha^* = \text{GoldenSearch}(\alpha \in [0, 1], f(\alpha) = S(\alpha, \beta_1, \beta_2, \epsilon_r))$;
    $(S, S_{I_1}, S_{I_2}, S_{T_1}, S_{T_2}) = \text{Score}(\alpha^*, \beta_1, \beta_2, \epsilon_r)$;
    if $S > S_{best}$ then
        $S_{\text{best}} = S$; $\theta_{\text{best}} = \{\alpha^*, \beta_1, \beta_2, \epsilon_r\}$
    if $S \geq TH$ then
        return $I(\theta^*),\ \theta^*$
    $S_1 = S_{I_1} + S_{T_1}, S_2 = S_{I_2} + S_{T_2}$;
    if $S_1 > S_2$ then
        $\beta_2^* = \text{GoldenSearch}(\beta_2 \in [\beta_{\min}, \beta_{\max}], f(\beta_2))$
    else
        $\beta_1^* = \text{GoldenSearch}(\beta_1 \in [\beta_{\min}, \beta_{\max}], f(\beta_1))$
    $(S', \cdot) = \text{Score}(\alpha^*, \beta_1^*, \beta_2^*, \epsilon_r)$;
    if $S' > S_{best}$ then
        $S_{\text{best}} = S'$; $\theta_{\text{best}} = \{\alpha^*, \beta_1^*, \beta_2^*, \epsilon_r\}$
    if $S' \geq TH$ then
        Normalize $z_1, z_2$ and compute spherical interpolation $z_{\text{SInp}}(\alpha^*)$;
        $x_T = \epsilon_r$;
        for $t = T$ to $0$ do
            $x_{t-1} = x_t - (\sigma_t - \sigma_{t-1}) v_\phi(x_t, t, z_{\text{SInp}}(\alpha^*), \gamma_{\text{gen}}, z_p)$
        $I = \mathcal{D}(x_0)$;    return $I,\ \theta^*$
if $\theta_{best} \neq \varnothing$ then
    Decode best parameters $\theta_{\text{best}}$ via MixingDenoise;
    return $I,\ \theta_{\text{best}}$
return $\varnothing$;
```

F Algorithm

Algorithm 1 outlines the complete inference process of our proposed framework, **VMDiff**, which integrates a noise refinement step and an efficient adaptive adjustment (EAA) loop. Given two input images I_1, I_2 and their category labels T_1, T_2, we construct a prompt P_G and initialize the fusion parameters $\theta = \{\alpha, \beta_1, \beta_2, \epsilon\}$.

The algorithm begins by sampling initial Gaussian noise ϵ, which is refined through a denoising-inversion procedure to produce a structure-aware latent representation ϵ_r. The core loop involves:

- **Searching** for the optimal interpolation factor α using Golden Section Search to maximize the similarity score $S(\theta)$.

- **Conditionally adjusting** the noise scaling factors β_1, β_2 when the current fusion score is below a threshold TH, guiding the fusion toward balance between the two source objects.
- **Returning** a fused image $I(\theta^*)$ once a satisfactory similarity score is achieved.

This design ensures a lightweight and interpretable optimization routine over a low-dimensional parameter space. The algorithm reliably produces perceptually and semantically coherent hybrid images, as validated in our experiments.

G More Results

In this section, we present additional qualitative results with **resampling disabled**, to evaluate VMDiff under a deterministic setting and further demonstrate its effectiveness and generalization. Fig. 1 shows generations at 1024×1024 resolution. Figs. 21, 22, 23, 24, 25, 26, 27, 28, and 29 provide diverse fusion examples spanning animals, fruits, artificial objects, and character figurines. In all figures, the leftmost column displays the source images, and the adjacent columns show the fused outputs.

These examples are generated from our IIOF dataset and cover a wide range of visual appearances and semantic attributes. Across varied fusion types—such as person–fruit, animal–object, and object–object—the results consistently exhibit structural coherence, balanced integration, and high visual fidelity. This indicates that VMDiff can integrate symbolic and structural cues into stylistically consistent hybrids, regardless of whether the source concepts are semantically similar or dissimilar.

Overall, these results substantiate the strong generalization of VMDiff, yielding novel, imaginative, and structurally plausible hybrid objects from diverse real-world inputs, even without resampling or seed variation.

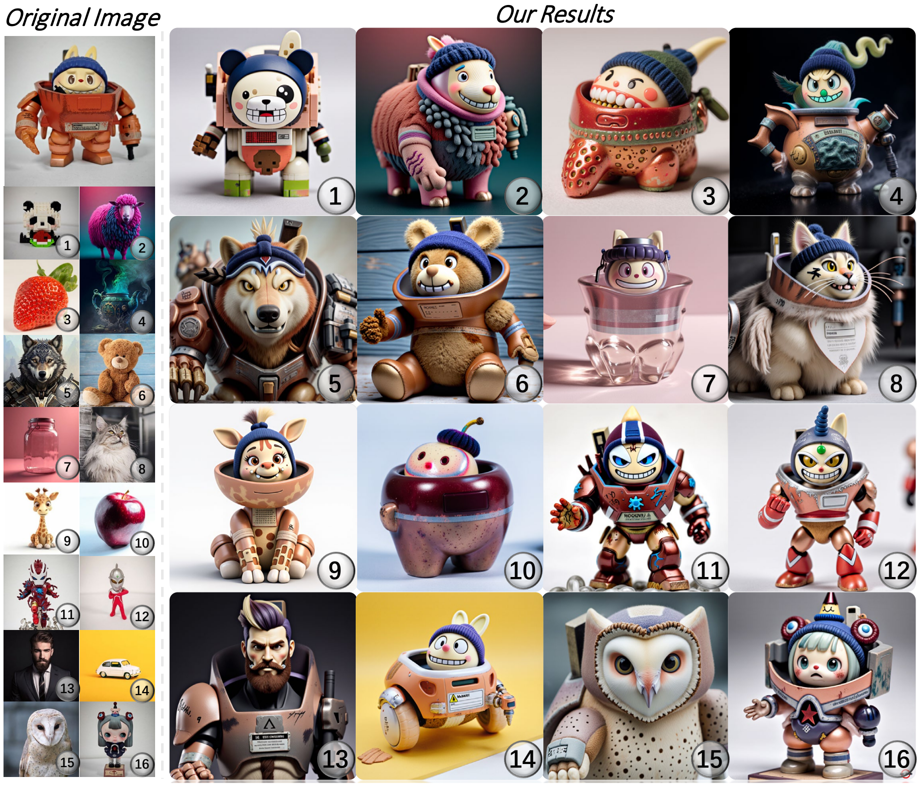


Figure 21: **More Results.** The primary source (*astronaut figurine*, top-left) is fused with secondary inputs (left column), with results shown on the right.

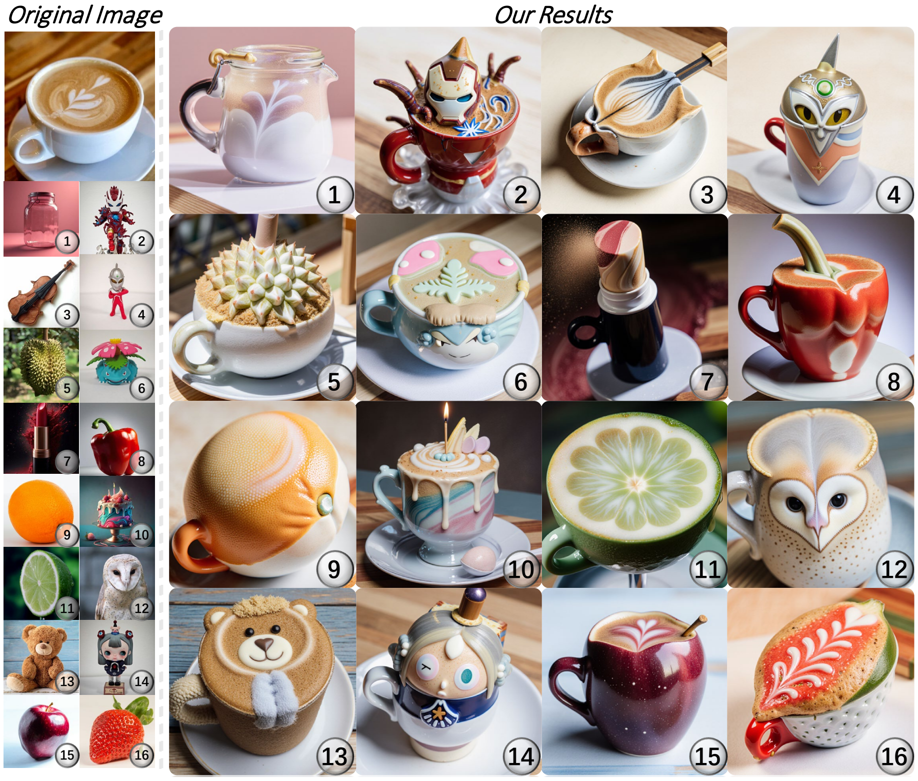


Figure 22: **More Results.** The primary source (*coffee cup*, top-left) is fused with secondary inputs (left column), with results shown on the right.

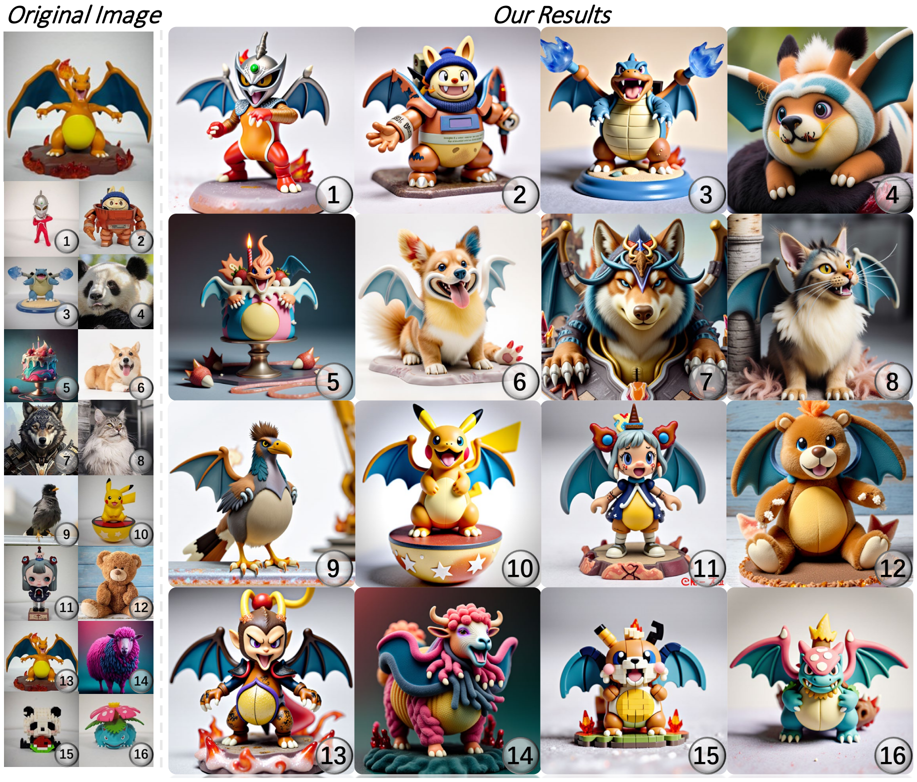


Figure 23: **More Results.** The primary source (*charizard figurine*, top-left) is fused with secondary inputs (left column), with results shown on the right.

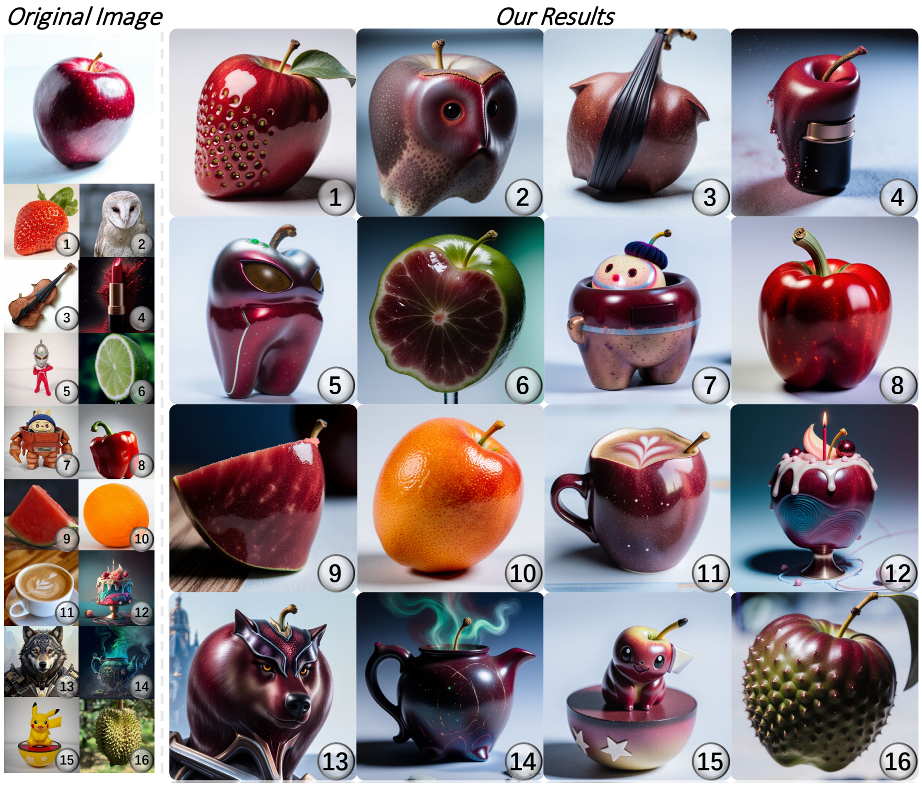


Figure 24: **More Results.** The primary source (*apple*, top-left) is fused with secondary inputs (left column), with results shown on the right.

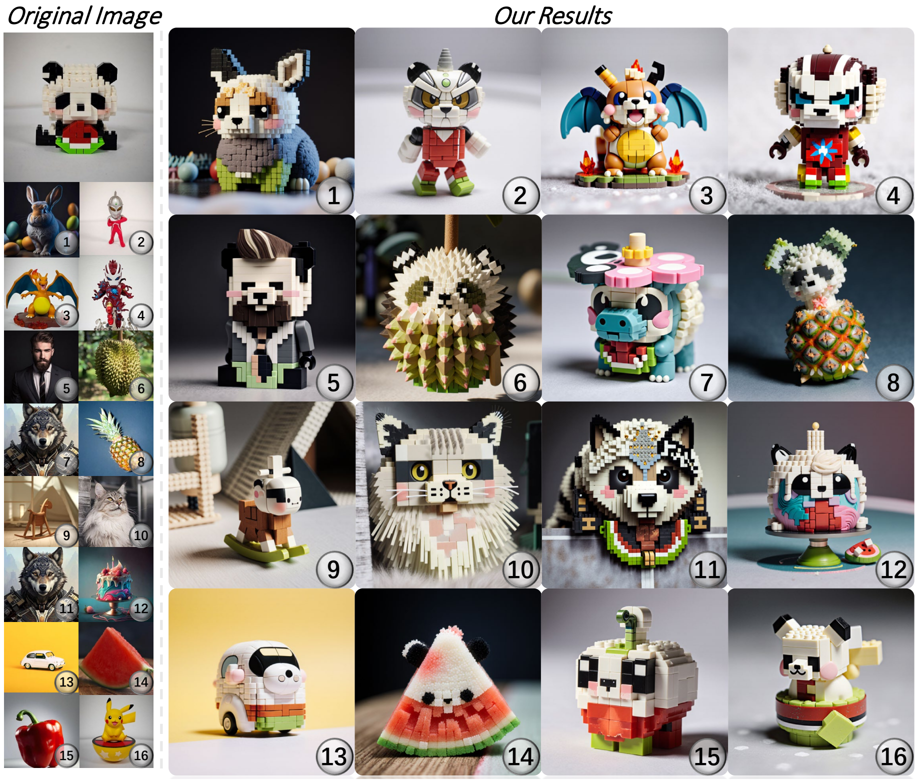


Figure 25: **More Results.** The primary source (*panda figurine*, top-left) is fused with secondary inputs (left column), with results shown on the right.

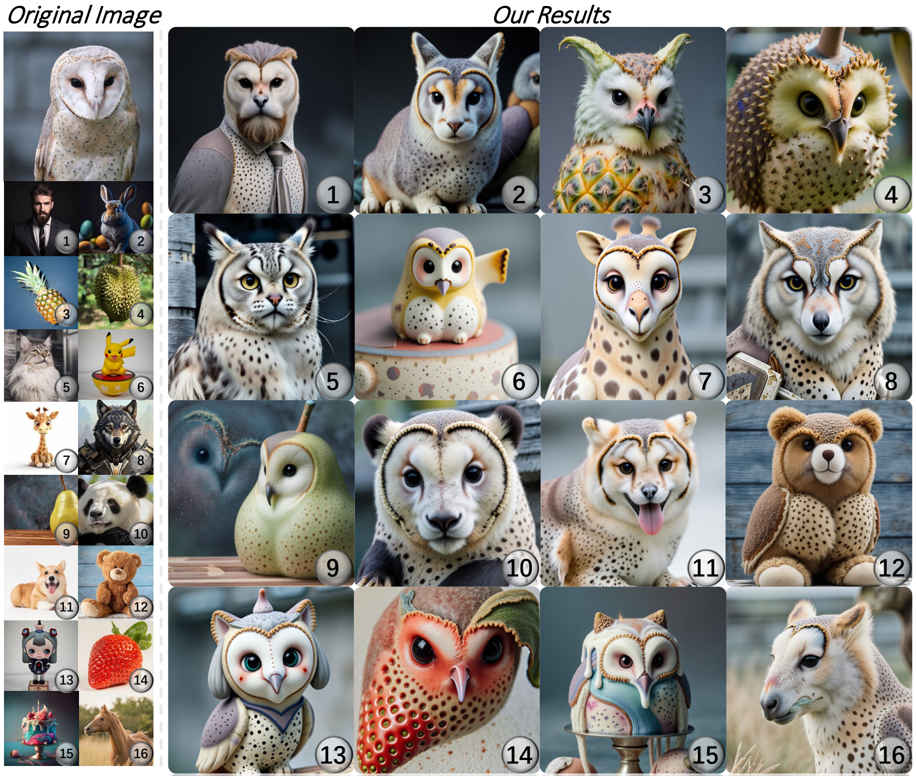


Figure 26: **More Results.** The primary source (*owl*, top-left) is fused with secondary inputs (left column), with results shown on the right.

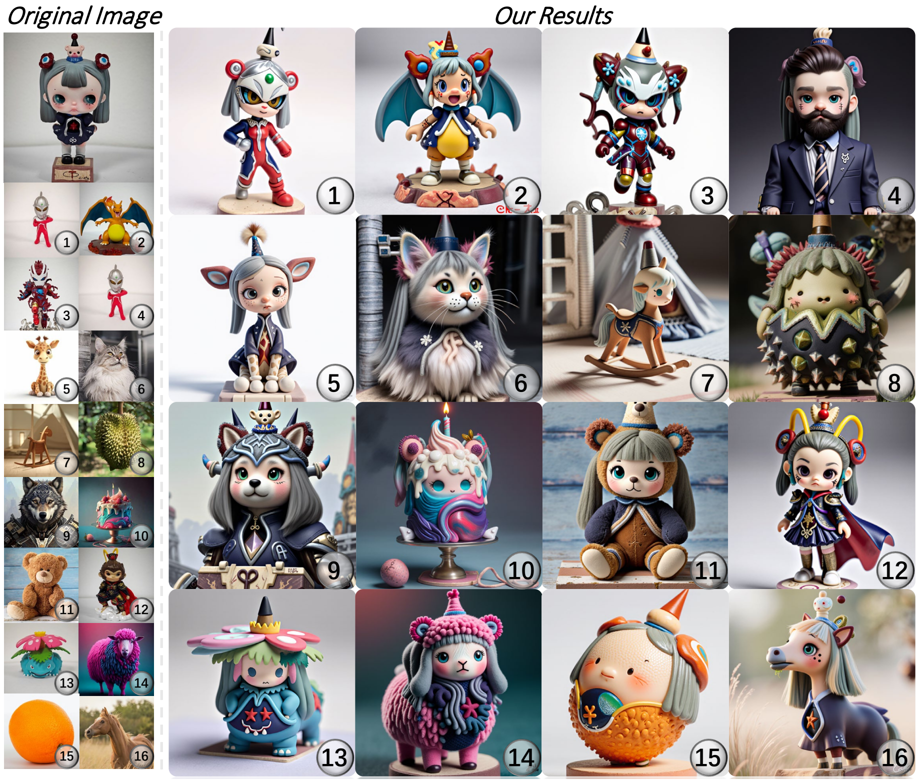


Figure 27: **More Results.** The primary source (*doll figurine*, top-left) is fused with secondary inputs (left column), with results shown on the right.

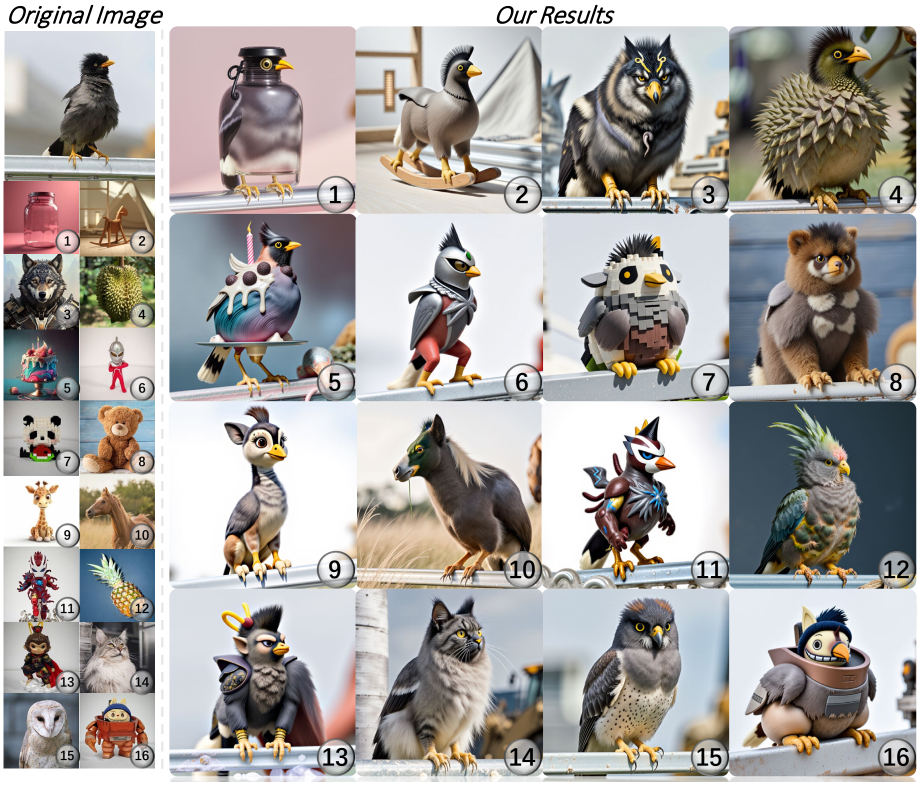


Figure 28: **More Results.** The primary source (*bird*, top-left) is fused with secondary inputs (left column), with results shown on the right.

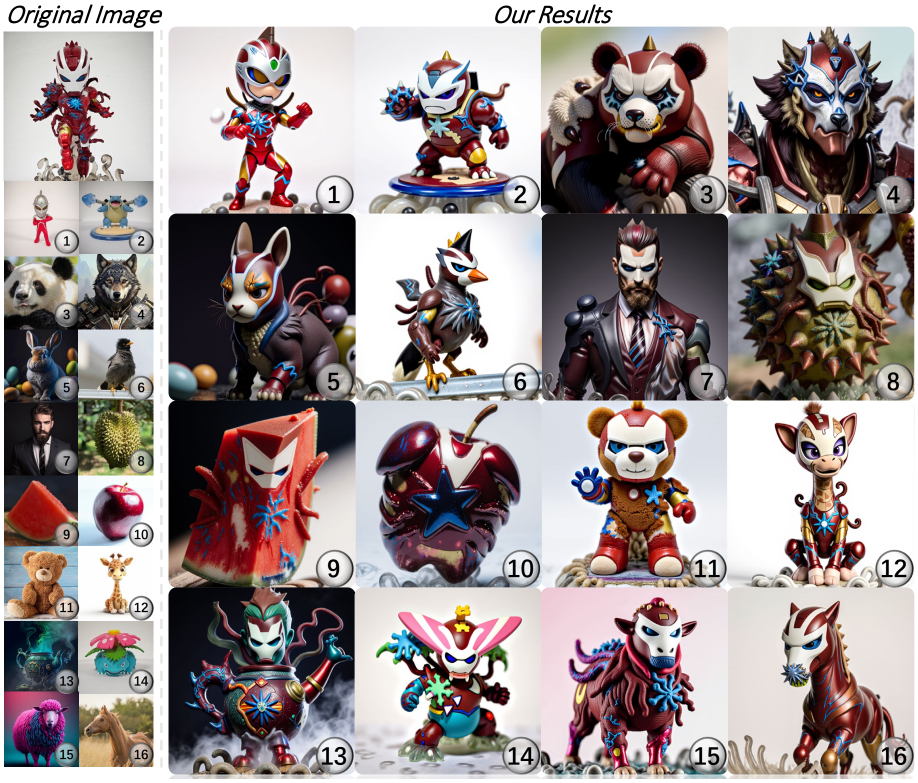


Figure 29: **More Results.** The primary source (*Iron man figurine*, top-left) is fused with secondary inputs (left column), with results shown on the right.