RMLer: Synthesizing Novel Objects Across Diverse Categories via Reinforcement Mixing Learning

Jun Li[1*], Zikun Chen[1], Haibo Chen[1*], Shuo Chen[2], Jian Yang[2]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]School of Intelligence Science and Technology, Nanjing University, China
{junli, zikunchen, hbchen}@njust.edu.cn, {shuo.chen, csjyang}@nju.edu.cn

Abstract

Novel object synthesis by integrating distinct textual concepts from diverse categories remains a significant challenge in Text-to-Image (T2I) generation. Existing methods often suffer from insufficient concept mixing, lack of rigorous evaluation, and suboptimal outputs—manifesting as conceptual imbalance, superficial combinations, or mere juxtapositions. To address these limitations, we propose **Reinforcement Mixing Learning (RMLer)**, a framework that formulates cross-category concept fusion as a reinforcement learning problem: mixed features serve as states, mixing strategies as actions, and visual outcomes as rewards. Specifically, we design an MLP-policy network to predict dynamic coefficients for blending cross-category text embeddings. We further introduce visual rewards based on (1) semantic similarity and (2) compositional balance between the fused object and its constituent concepts, optimizing the policy via proximal policy optimization. At inference, a selection strategy leverages these rewards to curate the highest-quality fused objects. Extensive experiments demonstrate RMLer's superiority in synthesizing coherent, high-fidelity objects from diverse categories, outperforming existing methods. Our work provides a robust framework for generating novel visual concepts, with promising applications in film, gaming, and design.

Introduction

The rise of large-scale Text-to-Image (T2I) synthesis, driven primarily by advances in diffusion models (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2023; Esser et al. 2024), has revolutionized digital content creation. These systems now support a wide range of applications, from artistic design (Montenegro 2024; Jin and Chua 2025; Xiong et al. 2024, 2025b,a; Chen et al. 2025) and virtual reality (Yin et al. 2024; Behravan, Matković, and Gračanin 2025) to film production and game development (Zhou et al. 2024; Sun et al. 2024). Recent improvements in photorealistic fidelity (OpenAI 2024a; Chen et al. 2024) and output diversity (Bau et al. 2023) have pushed the boundaries of what these systems can achieve. The next frontier lies in enhancing compositional reasoning and fine-grained control (Zhang, Rao, and Agrawala 2023; Meng et al. 2022), particularly in synthesizing novel objects by combining features from multiple concepts across different categories.

Figure 1: We propose a simple yet effective reinforcement mixing learning approach for generating novel object images by fusing distinct categories. For instance, our method seamlessly combines the *Venom* character with diverse animal categories—such as *bulldog*, *crocodile*, *turtle*, *kangaroo*, and *frog*—effectively blending their features to demonstrate its versatility.

Current T2I diffusion models employ two primary approaches to fuse multiple distinct textual concepts into a single, coherent object: general-purpose foundational models (e.g., SDXL-Turbo (Podell et al. 2023), DALL·E 3 (OpenAI 2024a), Flux (Labs 2024), GPT-Image-1 (OpenAI 2024b)) and specialized fusion techniques (e.g., BASS (Li, Zhang, and Yang 2024), ConceptLab (Richardson et al. 2024)). Despite their capabilities, these methods exhibit three key limitations: (1) **Conceptual Imbalance**—The generated image predominantly represents one object category, significantly overshadowing the other (left, Fig. 2). This bias stems from imbalanced prompt features, allowing one concept to dominate the composition. (2) **Superficial Combination**—The two concepts are merely overlapped without meaningful integration (middle, Fig. 2). Due to imbalanced local prompt features, the model exhibits a bias toward certain concepts in different spatial regions, disrupting coherent integration. (3) **Juxtaposition Generation**—The objects are placed separately in the image rather than being fused (right, Fig. 2). Without precise spatial control, the model generates multiple objects rather than a unified composition. Fundamentally,

*Corresponding authors.

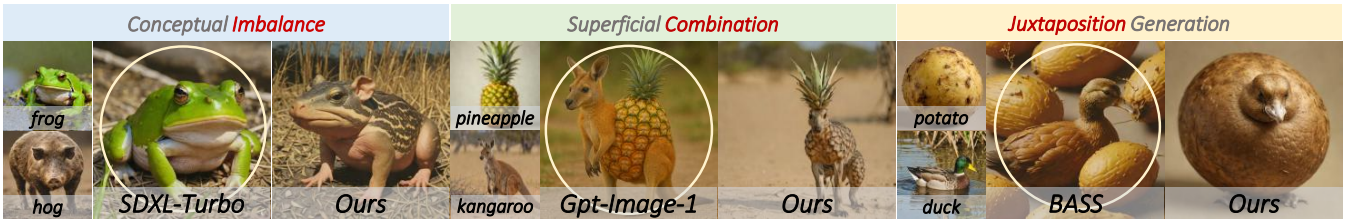


Figure 2: Failures in concept fusion by existing methods. **Left** (SDXL-Turbo (Podell et al. 2023)): Severe imbalance (e.g., *frog* + *hog* $\rightarrow$ dominant frog). **Middle** (GPT-Image-1): Superficial combination (e.g., *pineapple* + *kangaroo*). **Right** (BASS (Li, Zhang, and Yang 2024)): Simple juxtaposition (e.g., *owl* + *snail*). Our approach (rightmost) aims for more balanced and coherent fusions.

these issues arise from insufficient mixing and control over the characteristic features of the two categories.

To address these limitations, we propose **Reinforcement Mixing Learning (RMLer)**, a novel framework that formulates cross-category concept fusion as a reinforcement learning (RL) problem. Given two category labels, we first extract their text embeddings or features using simple prompt: *A photo of <category label>*, and define element-wise interpolation between these embeddings as a mixing strategy. The core idea of RMLer is to treat mixed text features as states, mixing strategies as actions, and the resulting visual outputs as rewards. Specifically, we design an MLP-policy network to predict dynamic interpolation coefficients for blending cross-category text embeddings. To guide learning, we introduce visual rewards that measure both semantic similarity and compositional balance between the fused object and its constituent concepts. These rewards ensure that the mixed features effectively integrate (local) prompt features, mitigating conceptual imbalance and superficial combination. Additionally, we leverage a foreground segmentation model to isolate objects in generated images, avoiding unintended juxtapositions. The policy network is optimized via Proximal Policy Optimization (PPO) (Schulman et al. 2017). During inference, a principled post-selection mechanism—guided by metrics aligned with our reward functions—refines the outputs to select the most compelling fused objects. The key strength of RMLer lies in its ability to learn adaptive mixing strategies through direct optimization of complex fusion objectives, enabling sophisticated embedding manipulation. Extensive experiments demonstrate RMLer's effectiveness, showing that it generates novel fused objects that standard text-to-image baselines struggle to produce (see Fig. 4). Overall, our main contributions are as follows.

- 1) We propose Reinforcement Mixing Learning, a framework that learns an adaptive policy to dynamically manipulate text embeddings across diverse categories for generating novel objects. To the best of our knowledge, this is the first work to effectively formulate cross-category fusion as a reinforcement learning problem.
- 2) Extensive experiments (Fig. 2, Table 2) demonstrate that RMLer synthesizes harmoniously fused objects from disparate categories, outperforming standard T2I techniques in quality, coherence, and compositional fidelity.

Related Work

Text-to-image (T2I) synthesis has advanced rapidly, but controllable and coherent fusion of multiple concepts remains challenging. We review related work in three areas: T2I synthesis, alignment of diffusion models, and object fusion.

Text-to-Image Synthesis. Diffusion models have significantly advanced T2I synthesis in fidelity and diversity (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2023; Zhang, Rao, and Agrawala 2023; Zhang, Wang, and Li 2023; Gu et al. 2022; Gong et al. 2024). Recent architectures like MMDiT (Esser et al. 2024) further improve multi-entity and stylistic generation. However, these models are primarily designed for holistic scene synthesis from a single prompt and often fail when fusing distinct concepts into a coherent entity. Challenges such as attribute leakage (Roy and Boddeti 2019) and semantic imbalance (Ma et al. 2022) arise, especially for out-of-distribution combinations (Madan et al. 2022). In contrast, RMLer presents a policy-driven control over input conditioning, learning to optimally merge concept embeddings for improved compositional fusion.

Alignment of Diffusion Models. To enhance controllability in diffusion models, recent works leverage reinforcement learning from human feedback (RLHF) (Liu et al. 2024b,a), widely adopted in large language model alignment (Ouyang et al. 2022; Bai et al. 2022). Reward models (Schuhmann et al. 2022; Xu et al. 2023; Kirstain et al. 2023; Wu et al. 2023) have enabled learning-based guidance in image generation. Building on this, DDPO (Black et al. 2023), DPOK (Fan et al. 2023), DiffusionDPO (Wallace et al. 2024), and others (Clark et al. 2023; Prabhudesai et al. 2023) formulate diffusion sampling as an MDP and apply policy gradients or reward backpropagation for alignment. While effective at attribute control, these methods typically modify the backbone. In contrast, RMLer introduces a lightweight policy over conditioning embeddings, enabling fine-grained fusion without altering the diffusion network.

Object Fusion. There has been increasing interest in generating fused images (Liew et al. 2022; Yi et al. 2024; Zhang et al. 2025) from multiple concepts, a task that holds great potential for creative applications such as digital art and design. ConceptLab (Richardson et al. 2024) employs diffusion models to synthesize unique visual concepts but its optimization-based approach is computationally expensive and often struggles to semantically integrate real-world concepts. BASS (Li, Zhang, and Yang 2024) introduces a more controllable frame-

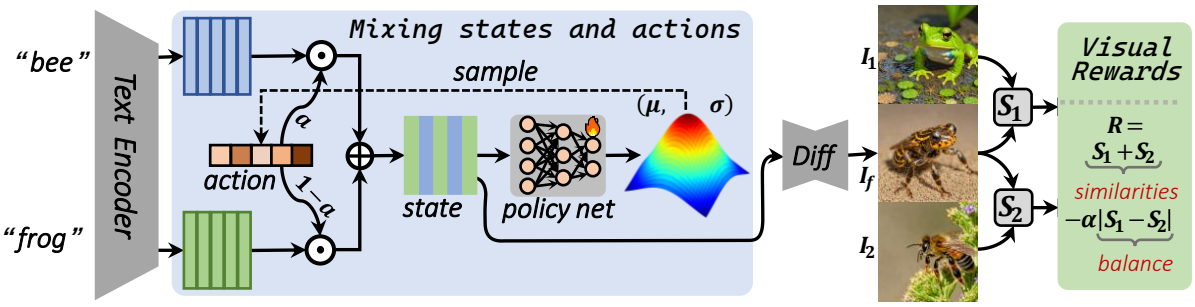


Figure 3: Pipeline of our Reinforcement Mixing Learning (RMLer). Given CLIP embeddings for two concepts $(\mathbf{e}_1, \mathbf{e}_2)$ extracted from labels (c_1, c_2), our policy network π_θ generates an action vector $\mathbf{a}$ that mixs $\mathbf{e}_1$ and $\mathbf{e}_2$ into a fused embedding $\mathbf{e}_f$. This embedding conditions a diffusion model $\mathcal{G}$ to synthesize the image I_f. A visual reward R, computed from CLIP similarity and balance between I_f and references I_1 and I_2 generated by $\mathbf{e}_1$ and $\mathbf{e}_2$ respectively, guides the PPO algorithm to update π_θ.

work for concept fusion by learning balance-aware token swapping. However, the swapped regions can sometimes lead to non-meaningful or visually chaotic results. In contrast, our RMLer framework offers a more efficient and adaptive solution for concept fusion. By learning a policy to directly manipulate embeddings, RMLer enables faster generation of semantically coherent and well-balanced fused images.

Preliminaries

Markov Decision Process (MDP) (Garcia and Rachelson 2013) provides a mathematical framework for modeling decision-making under uncertainty. An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \rho_0)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, P is the state transition probability, R is the reward function, and ρ_0 is the initial state distribution. At each step, an agent selects an action $\mathbf{a}_t \sim \pi(\mathbf{a}_t|\mathbf{s}_t)$, receives a reward $R(\mathbf{s}_t, \mathbf{a}_t)$, and transitions to a new state $\mathbf{s}_{t+1}$. The goal of reinforcement learning is to find a policy π^* that maximizes the expected cumulative reward:

$$\mathcal{J}_{\mathrm{RL}}(\pi) = \mathbb{E}_{\tau\sim p(\tau|\pi)}\left[\sum_{t=0}^{T} R(\mathbf{a}_t, \mathbf{s}_t)\right], \qquad (1)$$

where τ is the trajectory generated by a policy π over T.

Denoising Diffusion Policy Optimization (DDPO) (Black et al. 2023) reformulates the iterative denoising process of diffusion models as a multi-step MDP to enable fine-tuning via reinforcement learning. Each denoising step is treated as an action, and the policy π corresponds to the reverse diffusion kernel $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$, conditioned on time t and context $\mathbf{c}$. The MDP components are defined as:

$$\mathbf{s}_t \triangleq (\mathbf{c}, t, \mathbf{x}_t), \qquad \pi(\mathbf{a}_t \mid \mathbf{s}_t) \triangleq p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}),$$

$$\mathbf{a}_t \triangleq \mathbf{x}_{t-1}, \qquad \rho_0(\mathbf{s}_T) \triangleq (p(\mathbf{c}), \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})),$$

$$P(\mathbf{s}_{t-1} \mid \mathbf{s}_t, \mathbf{a}_t) \triangleq (\delta_{\mathbf{c}}, \delta_{t-1}, \delta_{\mathbf{x}_{t-1}}),$$

$$R(\mathbf{s}_t, \mathbf{a}_t) \triangleq \begin{cases} r(\mathbf{x}_0, \mathbf{c}) & \text{if } t = 0, \\ 0 & \text{otherwise,} \end{cases}$$

where δ_y denotes the Dirac delta function.

Methodology

In this section, we present a Reinforcement Mixing Learning (RMLer) framework for multi-concept fusion in Fig. 3. Our approach consists of three key components. *Problem Formulation:* We formulate multi-concept fusion as a reinforcement learning (RL) task. *Visual Reward Function:* We introduce a reward function based on visual similarity and balance, ensuring high-quality and harmonious outputs. *Two-Stage Sampling Strategy:* To enhance efficiency, we propose a two-stage sampling method that selects the most representative fused object from candidate generations.

Problem Formulation of RMLer

Cross-category concept fusion (CCF) is a challenging task that combines two distinct textual concepts, c_1 and c_2 into a single novel and coherent object image I_f. Our RMLer formulates this task as a multi-step Markov Decision Process (MDP), shown in Figure 3. Before detailing our method, we first formally define the CCF task.

The CCF Task. Given two distinct category labels c_1 and c_2, we construct simple text prompts: p_1 : *A photo of* $< c_1 >$ and p_2 : *A photo of* $< c_2 >$. These prompts are fed into the T2I diffusion model to generate their corresponding original images: $I_1 \sim \mathcal{G}(\mathbf{e}_1, \epsilon) \in \mathbb{R}^{H\times W}$ and $I_2 \sim \mathcal{G}(\mathbf{e}_2, \epsilon) \in \mathbb{R}^{H\times W}$, where $\mathbf{e}_1 = \mathcal{E}(p_1) \in \mathbb{R}^{h\times w}$, $\mathbf{e}_2 = \mathcal{E}(p_2) \in \mathbb{R}^{h\times w}$ and ϵ is a sampling noise. The CCF task involves fusing $\mathbf{e}_1$ and $\mathbf{e}_2$ into a mixing text embedding $\mathbf{e}_f$, which is then used to generate a novel and coherent fused image, $I_f = \mathcal{G}(\mathbf{e}_f) \in \mathbb{R}^{H\times W}$. In our implementation, we use a pretrained Stable Diffusion model (Podell et al. 2023) as our baseline, where $\mathcal{E}(\cdot)$ denotes the text encoder and $\mathcal{G}(\cdot)$ represents the diffusion-based generator. Our framework is model-agnostic and can be adapted to other diffusion models.

CCF as a multi-step MDP. We formulate the CCF task as a multi-step (MDP). In each fusion episode, consisting of T steps $(t = 0, \ldots, T-1)$, the agent interacts with the environment as follows:

- **State $\mathbf{s}_t$:** the current fused embedding $\mathbf{e}_f^{(t)} \in \mathbb{R}^{h\times w}$.
- **Action $\mathbf{a}_t$:** the column-wise interpolation coefficient $\mathbf{a}_t \in \mathbb{R}^w$. The initial state $\mathbf{s}_0$ is computed as the average of the source embeddings, $\mathbf{s}_0 = \mathbf{e}_f^{(0)} = \frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2)$.

–	**RMLer**	PPO	–	**RMLer**	PPO
HPSv2 ↑	**0.2774**	0.2746	VQAScore ↑	**0.4287**	0.4155

Table 1: Comparing PPO and our RMLer in CangJie-200.

- **Policy** $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$**:** a stochastic policy parameterized by an MLP with weights θ,which outputs a distribution over possible actions given the current state $\mathbf{s}_t$. An action is sampled as $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathbf{s}_t)$.
- **Transition:** Updates the state $\mathbf{s}_t$ to the next state $\mathbf{s}_{t+1}$ via a fusion function $f_{\text{fuse}}(\cdot)$:

$$\begin{aligned}\mathbf{e}_f^{(t+1)} =& f_{\text{fuse}}(\mathbf{a}_t, \mathbf{e}_1, \mathbf{e}_2) \\ =& \mathbf{e}_1 \times \text{diag}(\mathbf{a}_t) + \mathbf{e}_2 \times \text{diag}(1 - \mathbf{a}_t), \end{aligned} \quad (2)$$

  where $\text{diag}(\mathbf{a}_t)$ denotes converting the vector $\mathbf{a}_t$ into a diagonal matrix.
- **Reward:** an evaluation score $R_{t+1} = r(I_f^{(t+1)}, c_1, c_2)$, where $I_f^{(t+1)} \sim \mathcal{G}(\mathbf{e}_f^{(t+1)}, \epsilon)$.

Formal MDP at timestep t:

$$\begin{aligned} \mathbf{s}_t &\triangleq \mathbf{e}_f^{(t)}, & \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) &\triangleq P(\mathbf{a}_t \mid \mathbf{s}_t; \theta), \\ \mathbf{a}_t &\sim \pi_\theta(\cdot \mid \mathbf{s}_t), & \mathbf{s}_{t+1} &\triangleq f_{\text{fuse}}(\mathbf{a}_t, \mathbf{e}_1, \mathbf{e}_2), \\ I_f^{(t+1)} &\sim \mathcal{G}(\mathbf{s}_{t+1}, \epsilon) & R_{t+1}(\mathbf{s}_t, \mathbf{a}_t) &\triangleq r(I_f^{(t+1)}, c_1, c_2). \end{aligned} \quad (3)$$

The RMLer objective is to learn π_θ that maximizes the quality of the best fused result encountered within an T-step trajectory $\{\mathbf{s}_0, \mathbf{a}_0, \ldots, \mathbf{s}_T\}$. While the process yields a sequence of images with rewards $\{R_1, \ldots, R_T\}$, our primary goal is to maximize the highest single-step reward, i.e., $\max_t R_t$, where we use $\sum_{t=1}^{T} \gamma_t R_t$ with $\gamma_t = 1$ (no discounting) to aggregate rewards. To guide policy learning, we retain intermediate rewards at each step and optimize π_θ using Proximal Policy Optimization (PPO) (Schulman et al. 2017), with the surrogate objective:

$$\begin{aligned}\mathcal{L}^{\text{PPO}}(\theta) =& \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}} [-\min\left(k_t(\theta) \cdot R(s_t, a_t),\right. \\ & \left.\text{clip}\left(k_t(\theta),\, 1-\xi, 1+\xi\right) \cdot R(s_t, a_t)\right)], \end{aligned} \quad (4)$$

where $k_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is a probability ratio, and $\xi = 0.2$ is a hyperparameter. Unlike standard PPO, our formulation in Eq. (4) eliminates the critic network, relying solely on policy optimization. Table 1 presents preliminary experiments showing that standard PPO suffers from performance degradation due to unstable training dynamics. To address this issue, we introduce an intrinsic visual reward mechanism in the following subsection.

Visual Reward Function

The reward $r(I_f, c_1, c_2)$ plays a key role in evaluating our method for the CCF task. We present a visual reward function that is based on CLIP similarity and balance between I_f and reference exemplars I_1 and I_2, which are generated by $\mathbf{e}_1(c_1)$ and $\mathbf{e}_2(c_2)$, respectively.

Specifically, we first extract foreground segments—$I_{f\text{seg}}$, $I_{1\text{seg}}$, and $I_{2\text{seg}}$—from I_f, I_1 and I_2 using a foreground segmentation model (Oquab et al. 2024). We then compute their CLIP image embeddings $\mathbf{f}_{I_{f\text{seg}}}$, $\mathbf{f}_{I_1}$ and $\mathbf{f}_{I_2}$ via a pretrained CLIP image encoder $E_{\text{CLIP-I}}$ (Radford et al. 2021). The **visual fusion reward** is defined as:

$$R = (S_1 + S_2) - \alpha \cdot |S_1 - S_2|, \quad (5)$$

where $S_1 = \text{sim}(\mathbf{f}_{I_{f\text{seg}}}, I_{1\text{seg}})$ and $S_2 = \text{sim}(\mathbf{f}_{I_{f\text{seg}}}, I_{2\text{seg}})$ denote the cosine similarities between the generated image and the two concept exemplars. The first two terms ensure that the fused image I_f maintains maximum similarity with both I_1 and I_2, indicating that I_f retains more characteristics from the distinct categories c_1 and c_2. The last term promotes balanced alignment with both concepts. The scale factor $\alpha > 0$ mitigates excessive dominance of one concept over the other.

This formulation encourages the RMLer policy π_θ to explore embedding manipulations that produce visually coherent and semantically balanced fusion results. Empirically, we find that image-based CLIP similarity offers stronger guidance than text-based reward signals, a finding further supported by our ablation studies (***refer to Appx. A***).

Two-Stage Sampling Strategy

After learning the policy π_{θ^*} (typically the best-performing checkpoint), the stochasticity of both policy sampling and diffusion generation leads to variability in inference outputs. To identify representative examples that reliably capture the capabilities of our RMLer framework—particularly for qualitative evaluation and visualization—we introduce a principled two-stage selection strategy.

Candidates Selected via Fusion Criteria. In the first stage, we filter a larger pool of generated images $\mathcal{I}_f = \{I_f\}$ to obtain a candidate set that meets two core criteria: *Concept Presence* means that the fused image must clearly exhibit the semantic attributes of the input categories; and *Fusion Balance* shows that the composition should harmoniously integrate all relevant elements. An image I_f is retained as a candidate only if it satisfies both conditions:

1. *Dual Concept Presence:* $S_1 > \tau_{\text{presence}}$ and $S_2 > \tau_{\text{presence}}$.
2. *Fusion Balance:* $|S_1 - S_2| < \tau_{\text{balance}}$.

where τ_{presence} and τ_{balance} are empirically set to 0.63 and 0.05, respectively. Therefore, we have a candidate set:

$$\begin{aligned}\mathcal{I}_{\text{can}} = \{I_f \mid & S_1 > \tau_{\text{presence}} \;\&\; S_2 > \tau_{\text{presence}} \;\& \\ & |S_1 - S_2| < \tau_{\text{balance}}, I_f \in \mathcal{I}_f\}. \end{aligned} \quad (6)$$

Top-1 Ranking. From the set $\mathcal{I}_{\text{can}}$, we select the top-1 image with the highest total semantic alignment score, computed as the sum of its similarities to both source concepts:

$$I_f^* = \max_{I_f \in \mathcal{I}_{\text{can}}} S_1 + S_2, \quad (7)$$

where the top-1 image with the highest score is selected as the final representative exemplar. Of course, the top-K images can also be provided for user selection.

By decoupling fusion balance (enforced during candidate qualification) from concept preservation strength, our two-stage selection ensures that the chosen exemplars are both semantically balanced and highly representative. This approach simplifies scoring while avoiding redundancy.

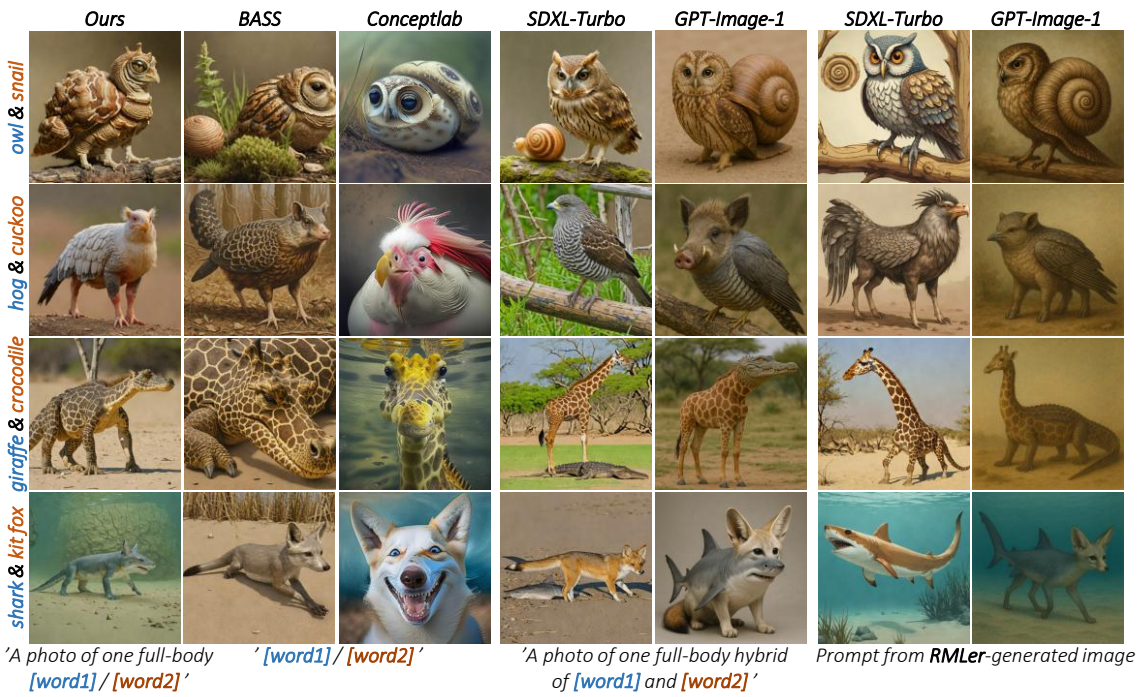


Figure 4: Comparisons with different methods on **ImageNet-200**. The complex prompts are created from RMLer-generated image using GPT-4o. For instance, *A hybrid creature combining an owl and a snail, with an owl-like head, sharp eyes and a curved beak, and a body covered by a spiral shell texture, standing on a wooden branch with bird-like legs and claws.*

Experiments

Experimental Settings

Dataset. We evaluate on a benchmark of 400 diverse concept pairs. This includes **ImageNet-200**, a set of 200 manually curated pairs from the ImageNet (Russakovsky et al. 2015) vocabulary, selected to maximize semantic and visual dissimilarity. In addition, we incorporate the **CangJie** dataset proposed in CreTok (Feng et al. 2024), which contains 200 concept pairs designed to test compositional creativity in the TP2O (Li, Zhang, and Yang 2024) task.

Details. Our method is implemented based on the **SDXL-Turbo** model (Podell et al. 2023) for efficient text-to-image generation. For semantic feature extraction—used in both reward computation and image selection—we employ the **CLIP ViT-H/14** model (Radford et al. 2021). Foreground segmentation is performed using the **RMBG-2.0** model (Zheng et al. 2024) to isolate salient content for evaluation. All generated and processed images are standardized to a resolution of 512×512 pixels. Experiments were conducted on a system equipped with two NVIDIA GeForce RTX 4090 GPUs.

Evaluation Metrics. We evaluate the performance of our RMLer framework using a comprehensive set of automated metrics that assess both fusion quality and perceptual realism. Specifically, we report the following five metrics:

- **Avg. Sim (I→I / I→T)**↑: Mean CLIP similarity between the generated image and either exemplar images (I→I) or text prompts (I→T), measuring overall concept alignment.
- **Balance (I→I / I→T)**↓: Absolute difference between CLIP similarities to the two source concepts (images or texts); lower values indicate more balanced fusion.
- **Reward**↑: Our reward score computes $(S_1+S_2)-\alpha|S_1-S_2|$ in Eq. 5, balancing concept presence and symmetry.
- **HPSv2**↑ (Wu et al. 2023) estimates human preference alignment for generated images, capturing overall aesthetic and alignment qualities.
- **VQAScore**↑ (Lin et al. 2024) evaluates visual-text alignment in complex compositional prompts.

These metrics provide a comprehensive assessment across fusion accuracy, conceptual balance, and perceptual quality.

Main Results

We conducted a comprehensive comparison of our RMLer with existing methods: BASS (Li, Zhang, and Yang 2024), ConceptLab (Richardson et al. 2024), SDXL-Turbo (Podell et al. 2023), and GPT-Image-1 (OpenAI 2024b).

Qualitative Results. Figs. 4 and 5 present qualitative comparisons between our method and several baselines across both ImageNet-200 and Cangjie. These examples reflect the key challenges we highlighted earlier: *Conceptual Imbalance*, *Superficial Combination*, and *Juxtaposition Generation*. As observed, methods like BASS, ConceptLab, and SDXL-Turbo often exhibit strong bias toward one of the source concepts, resulting in imbalanced or unintegrated outputs. In contrast, our approach consistently produces semantically balanced results that preserve salient features from both inputs. Additionally, GPT-Image-1 frequently suffers from superficial fusion. For example, in the *owl & snail* case in Fig. 4, it

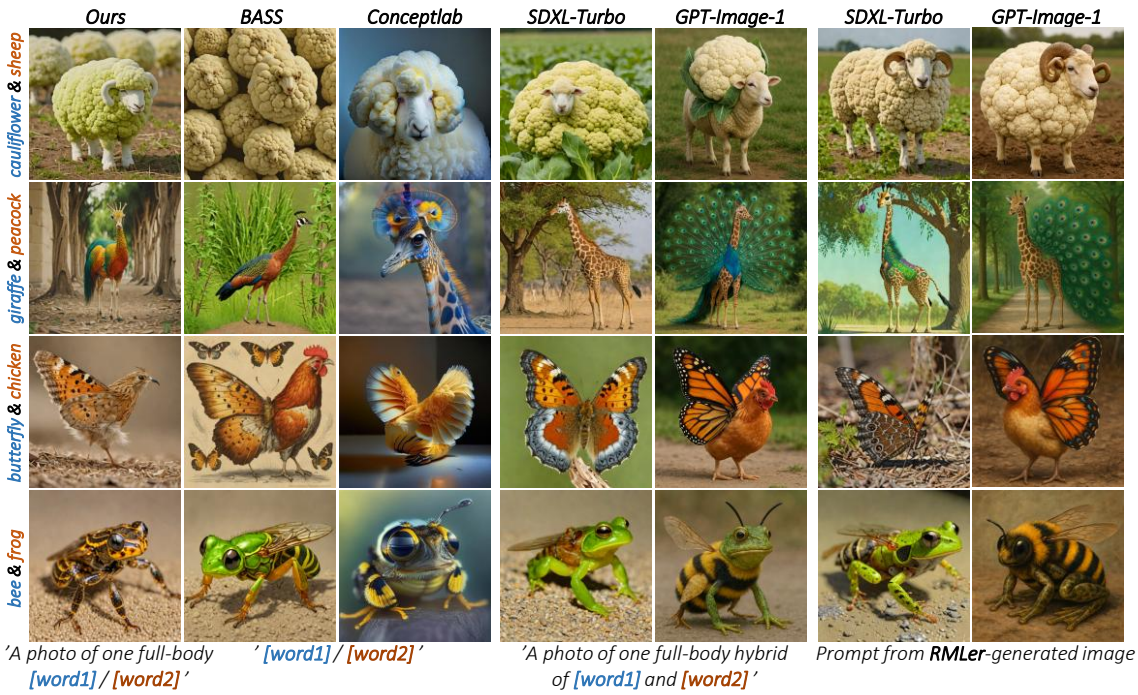


Figure 5: Comparison with different methods on the **CangJie-200**. The complex prompts are created from RMLer-generated images using GPT-4o. For instance, *A hybrid creature combining a butterfly and a chicken, with a compact, feathery bird body, thin legs and claws, and large, vibrant butterfly wings extending from its back, standing on dry forest ground.*

Model	Avg. Sim (I→I)↑		Avg. Sim (I→T)↑		Balance (I→I)↓		Balance (I→T)↓		Reward↑		HPSv2↑		VQAScore↑	
	Img	CJ	Img	CJ	Img	CJ	Img	CJ	Img	CJ	Img	CJ	Img	CJ
Our RMLer	0.7324	0.7193	0.2272	0.2452	**0.0080**	**0.0070**	**0.0394**	**0.0364**	**1.4244**	**1.4034**	0.2737	**0.2774**	**0.3301**	**0.4287**
BASS (Li, Zhang, and Yang 2024)	0.7026	0.6595	0.2223	0.2219	0.0918	0.1309	0.0659	0.0830	0.9459	0.6640	**0.2756**	0.2750	0.3055	0.3069
ConceptLab (Richardson et al. 2024)	0.5991	0.6021	0.2211	0.2434	0.0908	0.1112	0.0701	0.0662	0.7440	0.6480	0.2636	0.2714	0.2671	0.3440
SDXL-Turbo (Podell et al. 2023)	**0.7647**	**0.7413**	0.2410	0.2432	0.2380	0.2205	0.1463	0.1232	0.3394	0.3797	–	–	–	–
GPT-Image-1 (OpenAI 2024b)	0.7308	0.6927	**0.2608**	**0.2625**	0.1080	0.0853	0.0680	0.0451	0.9215	0.9585	–	–	–	–

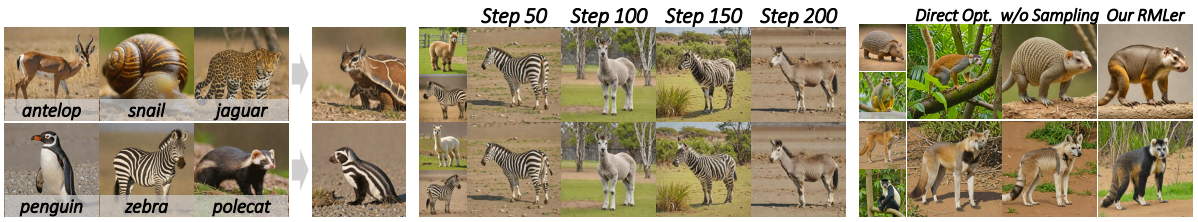
Table 2: Quantitative comparison on the ImageNet-200 and CangJie-200 benchmarks. For Avg. Sim and Balance, we report both image-to-image (I→I) and image-to-text (I→T) variants.

merely overlays a snail shell onto the owl's back, rather than synthesizing a cohesive hybrid entity. Our method, by comparison, generates a more seamless and conceptually blended composition that better reflects the intent of fusion.

In practice, manually conceptualizing prompts that effectively fuse two categories proves challenging. To address this, we employ GPT-4o to produce complex prompts based on RMLer-generated images. Our evaluation reveals mixed success rates when implementing these prompts with both SDXL-Turbo and GPT-Image-1. These results highlight the inherent difficulty of the C3F task, even when using state-of-the-art models like GPT-Image-1. Moreover, Fig. 6(a) demonstrates that our method can handle more than two categories.

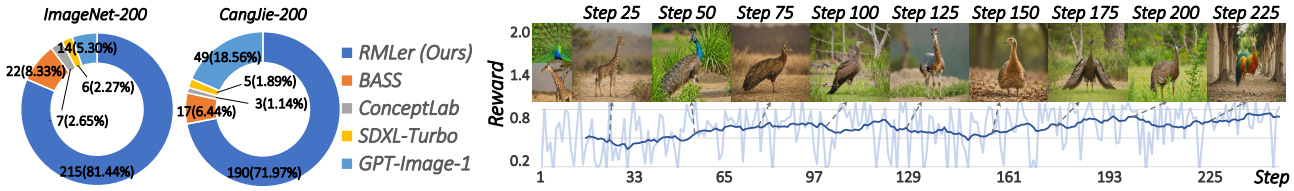
Quantitative Results. Table 2 reports quantitative comparisons across both ImageNet-200 and CangJie-200 benchmarks. Our method achieves state-of-the-art performance on key metrics, including *Balance*, *Reward*, *HPSv2*, and *VQAScore*, consistently outperforming the related approaches. These results demonstrate that RMLer not only generates more semantically balanced fusion images but also produces outputs with stronger visual appeal and better alignment with human preferences. Our Avg. Sim (I→I) scores are slightly lower than SDXL-Turbo's, this is because SDXL-Turbo generates high-quality composites of both categories (see Fig. 4) rather than true concept fusion. Similarly, our Avg. Sim (I→T) scores are lower than GPT-Image-1's, as GPT-Image-1 generates object-spliced partial semantic information instead of genuine concept fusion. Due to this discrepancy, we do not use image-text similarity as a reward signal. Additionally, we exclude HPSv2 and VQAScore for GPT-Image-1 and SDXL-Turbo, as these metrics assess image-text alignment using the input prompt. Since both models generate images directly from the evaluation prompt, their scores would be artificially inflated and incomparable.

User Study. To evaluate perceptual preference for fused images, we conducted a user study on both ImageNet-200 and CangJie-200, comparing our RMLer with four existing methods: BASS, ConceptLab, SDXL-Turbo, and GPT-Image-



(a) Generalizations using three categories. (b) Ablation study of different examples. (c) Ablation study of our RMLer.

Figure 6: Generalization and ablation analysis.



(a) User study on ImageNet-200 and CangJie-200. (b) Reward curve of our RMLer on the concept pair *giraffe* and *peacock*.

Figure 7: Human evaluation and training dynamics of RMLer.

1. A total of 66 participants cast 528 votes by selecting the most harmonious fusion result per concept pair. As shown in Fig. 7(a), RMLer received the highest preference on both benchmarks, achieving 81.44% on ImageNet-200 and 71.97% on CangJie, significantly outperforming all baselines. GPT-Image-1 ranked second on CangJie with 18.56%, while other methods received substantially lower preference rates.

Ablation Study and Parameter Analysis

Ablation Study. To compute similarity-based rewards during training, we use pre-generated exemplar images for each concept. To evaluate whether exemplar selection affects fusion quality, we analyze variations in exemplar sets. In Fig. 6(a), the visual appearance of generated results remains largely consistent, demonstrating robustness to exemplar choice. However, minor fluctuations in quantitative metrics (e.g., similarity scores) can occur. ***See Appx. B for further analysis.***

We further conduct ablation studies to evaluate RMLer's core components: the RL-trained policy and stochastic sampling. In Fig. 6(c), disabling either component (e.g., substituting reward-guided optimization for the RL policy or using deterministic sampling) degrades fusion quality and introduces semantic imbalance. This confirms that adaptive policy learning and controlled stochasticity are both critical for coherent, balanced concept fusion. ***See Appx. C for extended analysis and failure cases.***

Parameter Analysis. Key hyperparameters in our RMLer include the balance factor α in the reward function (Eq. (5)), the training steps and the thresholds τ_{presence} and τ_{balance} in the candidate image selection process.

Reward Balance α. We selected 10 diverse concept pairs from our ImageNet-200 benchmark to evaluate the reward balance factor $\alpha \in \{0, 1, 3, 5, 7\}$, training a separate RMLer agent for each configuration. For each α, we generated 100 fused images (10 per pair) and assessed them using HPSv2 and VQAScore. As Table 3 shows, $\alpha = 5$ achieves the optimal balance between fusion quality and concept preservation.

α	0	1	3	**5**	7
HPSv2 ↑	**0.2753**	0.2747	0.2741	0.2748	0.2736
VQAScore ↑	0.1773	0.1751	0.1836	**0.2148**	0.1994

Table 3: Parameter analysis of the reward balance factor α.

Training Steps. Fig. 7(b) illustrates RMLer's training dynamics for the *giraffe-peacock* concept pair. The reward curve demonstrates consistent improvement across training iterations, reflecting progressively better concept fusion quality. This optimization process enables the framework to ultimately generate high-fidelity fused outputs.

Selection Thresholds $\tau_{presence}$ *and* $\tau_{balance}$. For filtering fused results, we empirically set $\tau_{\text{presence}} = 0.63$ to ensure both source concepts are sufficiently present, and $\tau_{\text{balance}} = 0.05$ to encourage highly symmetric fusion. ***Please refer to Appx. D for further sensitivity analysis.***

Conclusion

In this work, we proposed RMLer, the first reinforcement learning (RL) framework for concept fusion in text-to-image synthesis. Leveraging PPO, our method learns an adaptive policy to dynamically manipulate text embeddings, enabling precise control over concept fusion in diffusion models. We designed a CLIP-based visual reward function that ensures semantically coherent and well-balanced generations, along with a novel selection mechanism to identify the most representative fused output. Extensive experiments on two diverse benchmarks show that RMLer outperforms the related fused methods by a significant margin—both in quantitative metrics and human evaluations—particularly when mixing semantically dissimilar concepts. ***See Appx. E for limitations.***

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. U24A20330, 62361166670, 62502208 and the Youth Science Foundation of Jiangsu Province under Grant BK20230924.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bau, D.; Song, J.; Chen, X.; Belanger, D.; Ho, J.; Vedaldi, A.; and Zhou, B. 2023. Editing Implicit Assumptions in Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Behravan, M.; Matković, K.; and Gračanin, D. 2025. Generative AI for context-aware 3D object creation using vision-language models in augmented reality. In *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, 73–81. IEEE.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Chen, H.; Chen, Z.; Zhao, L.; Li, J.; and Yang, J. 2025. PaintDiffusion: Towards text-driven painting variation via collaborative diffusion guidance. *Neurocomputing*, 620: 129284.

Chen, Z.; Su, R.; Zhu, J.; Yang, L.; Lai, J.-H.; and Xie, X. 2024. VividDreamer: Towards High-Fidelity and Efficient Text-to-3D Generation. *arXiv preprint arXiv:2406.14964*.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 79858–79885.

Feng, F.; Xie, Y.; Yang, X.; Wang, J.; and Geng, X. 2024. Redefining¡ creative¿ in dictionary: Towards an enhanced semantic understanding of creative generation. *arXiv preprint arXiv:2410.24160*.

Garcia, F.; and Rachelson, E. 2013. Markov decision processes. *Markov Decision Processes in Artificial Intelligence*, 1–38.

Gong, C.; Dai, Y.; Li, R.; Bao, A.; Li, J.; Yang, J.; Zhang, Y.; and Li, X. 2024. Text2Avatar: Text to 3D Human Avatar Generation with Codebook-Driven Body Controllable Attribute. *arXiv preprint arXiv:2401.00711*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; and Zhang, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10696–10706.

Jin, Z.; and Chua, T.-S. 2025. Compose Your Aesthetics: Empowering Text-to-Image Models with the Principles of Art. *arXiv preprint arXiv:2503.12018*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, J.; Zhang, Z.; and Yang, J. 2024. TP2O: Creative Text Pair-to-Object Generation Using Balance Swap-Sampling. In *European Conference on Computer Vision*, 92–111. Springer.

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

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 366–384. Springer.

Liu, L.; Du, C.; Pang, T.; Wang, Z.; Li, C.; and Xu, D. 2024a. Improving Long-Text Alignment for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2410.11817*.

Liu, Z.; Xiao, T. Z.; Liu, W.; Bengio, Y.; and Zhang, D. 2024b. Efficient Diversity-Preserving Diffusion Alignment via Gradient-Informed GFlowNets. *arXiv preprint arXiv:2412.07775*.

Ma, Y.; Jiao, L.; Liu, F.; Li, Y.; Yang, S.; and Liu, X. 2022. Delving into semantic scale imbalance. *arXiv preprint arXiv:2212.14613*.

Madan, S.; Henry, T.; Dozier, J.; Ho, H.; Bhandari, N.; Sasaki, T.; Durand, F.; Pfister, H.; and Boix, X. 2022. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2): 146–153.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*.

Montenegro, N. 2024. Integrative analysis of Text-to-Image AI systems in architectural design education: pedagogical innovations and creative design implications. *Journal of Architecture and Urbanism*, 48(2): 109–124.

OpenAI. 2024a. DALL·E 3: AI System for Generating Images from Text. https://www.openai.com/dall-e-3. Accessed: 2024-05-16.

OpenAI. 2024b. GPT-4 with Vision (gpt-image-1). https://platform.openai.com/docs/models/gpt-image-1.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.;

Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Prabhudesai, M.; Goyal, A.; Pathak, D.; and Fragkiadaki, K. 2023. Aligning Text-to-Image Diffusion Models with Reward Backpropagation. *arXiv preprint arXiv:2310.03739*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Richardson, E.; Goldberg, K.; Alaluf, Y.; and Cohen-Or, D. 2024. ConceptLab: Creative Concept Generation using VLM-Guided Diffusion Prior Constraints. *ACM Transactions on Graphics (TOG)*, 43(2).

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roy, P. C.; and Boddeti, V. N. 2019. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2586–2594.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Proceedings of Advances in Neural Information Processing Systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 25278–25294.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, Q.; Luo, Q.; Ni, Y.; and Mi, H. 2024. Text2AC: A Framework for Game-Ready 2D Agent Character (AC) Generation from Natural Language. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*.

Xiong, Z.; Chen, Z.; Zhang, Z.; Li, X.; Tai, Y.; Yang, J.; and Li, J. 2025a. Category-Aware 3D Object Composition with Disentangled Texture and Shape Multi-view Diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9832–9841.

Xiong, Z.; Yu, Y.; Zhang, Z.; Chen, S.; Yang, J.; and Li, J. 2025b. VMDiff: Visual Mixing Diffusion for Limitless Cross-Object Synthesis. *arXiv preprint arXiv:2509.23605*.

Xiong, Z.; Zhang, Z.; Chen, Z.; Chen, S.; Li, X.; Sun, G.; Yang, J.; and Li, J. 2024. Novel object synthesis via adaptive text-image harmony. *Advances in Neural Information Processing Systems*, 37: 139085–139113.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Yi, X.; Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2024. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110: 102450.

Yin, Z.; Wang, Y.; Papatheodorou, T.; and Hui, P. 2024. Text2vrscene: Exploring the framework of automated text-driven generation system for vr experience. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 701–711. IEEE.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3813–3824.

Zhang, Y.; Wang, L.; and Li, H. 2023. Inversion-Based Style Transfer With Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Tai, Y.; Qian, J.; Yang, J.; and Li, J. 2025. AGSwap: Overcoming Category Boundaries in Object Fusion via Adaptive Group Swapping. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 1–12.

Zheng, P.; Gao, D.; Fan, D.-P.; Liu, L.; Laaksonen, J.; Ouyang, W.; and Sebe, N. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research*.

Zhou, H.; Zhu, J.; Mateas, M.; and Wardrip-Fruin, N. 2024. The Eyes, the Hands and the Brain: What can Text-to-Image Models Offer for Game Design and Visual Creativity? In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, 1–13.