CIF: COMMONALITY-INDIVIDUALITY FUSION FOR CREATIVE TWO-OBJECT COMPOSITION

Anonymous authors
Paper under double-blind review

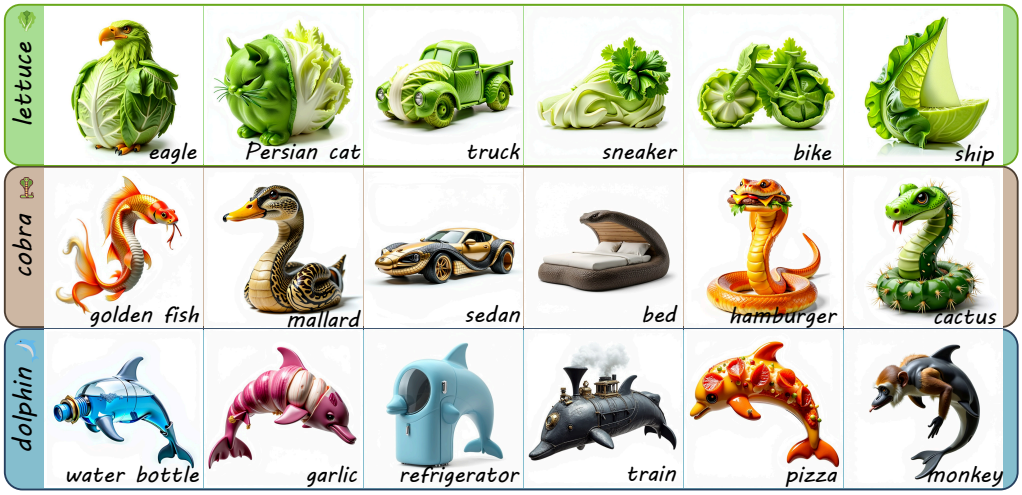

Figure 1: Compositional object created by our CIF. Each row shows novel objects generated by fusing a given source concept with six distinct target concepts.

ABSTRACT

Compositional object generation with text-to-image (T2I) models has attracted growing interest in creative design and content creation. However, existing object-fusion methods often rely on object-level representations or similarity measures, leading to superficial and structurally incoherent combinations. We propose **C**ommonality–**I**ndividuality **F**usion (**CIF**), a compositional generation framework that establishes shared structure before enhancing individual characteristics. First, *Common–Individual Division* (CID) decomposes two source objects into shared and individual semantic components. Then, *Common Fusion* (CF) integrates the shared components to construct a coherent structural scaffold and trains a fused diffusion model to generate the common foundation. Finally, *Individual Enhancement* (IE) incorporates the individual components and controlled perturbations to generate diverse fusion candidates, from which human preferences are used to optimize the fused diffusion model via direct preference optimization. Experiments show that CIF consistently outperforms existing T2I and object-fusion methods, producing more coherent and distinctive compositional objects.

1 INTRODUCTION

Text-to-image (T2I) models have transformed visual content creation by enabling image synthesis from natural language descriptions (Rombach et al., 2022; Labs, 2025). Building on this capability, combinatorial object generation aims to fuse two source concepts into a single coherent object that retains recognizable features from both (Li et al., 2024; Zhang et al., 2025). The generated images can be used in product design (Alcaide-Marzal & Diego-Mas, 2025), artistic creation (Zhou & Lee, 2024), film production (Dinkevich et al., 2026), and game development (Gallotta et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, directly prompting a T2I model Labs (2025) to fuse two object concepts often fails. Since target images of novel fused objects are hard to obtain, direct supervision is limited. Consequently, the output may contain only one source object or place the two side by side. Recent methods like RMLer (Li et al., 2026) and AGSwap (Zhang et al., 2025) improve object fusion by learning to mix source text embeddings (Li et al., 2026), or by combining reference images through guided denoising and inversion, as in VMDiff (Xiong et al., 2026). However, these methods do not explicitly match parts across the source objects. As a result, object-level fusion may still produce only superficial combinations of the source objects, as illustrated by the AGSwap result for *cap–laptop* in Fig. 2.

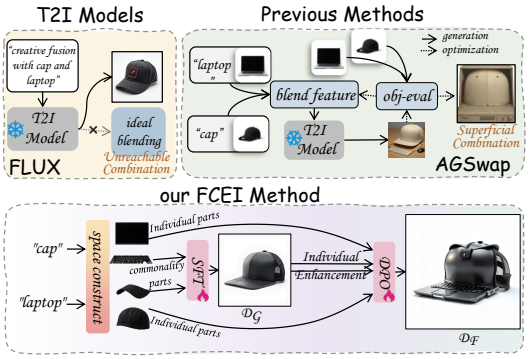

Figure 2: Failure cases in *cap–laptop* fusion: FLUX generates only the *cap*, while AGSwap produces a superficial combination. CIF integrates source parts into a coherent object.

To address these challenges, we propose **CIF**, a compositional generation framework that first establishes commonality and then enhances individuality (Fig. 3). It consists of three components. First, *Common–Individual Division* (CID) matches source parts based on geometry and spatial relations, separating commonality for establishing the fusion structure from individuality for enhancing distinctive details. Second, *Common Fusion* (CF) uses commonality-based part images as a visual scaffold to guide early sampling, together with a concept-preservation loss to retain both source concepts. We fine-tune the T2I model with these objectives to learn a fusion model that combines common parts into a coherent object. Third, *Individual Enhancement* (IE) generates diverse fusion candidates by conditioning on individuality and introducing random perturbations. Human-selected the best candidates are paired with unperturbed fusion results to construct preference data, which is used to further fine-tune the fusion model via direct preference optimization (DPO) (Wallace et al., 2024). Experiments on a concept-based split of 4,950 object pairs show that CIF consistently outperforms existing T2I models and fusion methods in fusion quality.

Our contributions are summarized as follows: ***(1)*** We propose CIF, a compositional object generation framework that progressively establishes commonality and enhances individuality. ***(2)*** We introduce *Common–Individual Division* and *Common Fusion*, which decompose source objects into common and individual parts and use commonality to construct a visual scaffold for coherent fusion while preserving both source concepts. ***(3)*** We develop *Individual Enhancement*, which explores novel structures and richer details through individuality conditioning and stochastic perturbations, and leverages human preferences via DPO to improve the fusion model.

2 Related Work

Text-to-Image Generation. Text-to-image models generate images from natural-language descriptions (Nichol et al., 2021; Saharia et al., 2022). A representative approach is latent diffusion (Rombach et al., 2022),which generate images by denoising compressed image representations under text conditions. Building on this design, SDXL (Podell et al., 2024) improves high-resolution synthesis with a larger U-Net and two text encoders. Diffusion Transformers (Peebles & Xie, 2023) provide an alternative backbone: FLUX.1-dev (Labs, 2025) uses a rectified-flow Transformer, while KREA-2 Turbo (The Krea Team, 2026) combines a diffusion Transformer with fine-tuning and distillation for efficient generation. Beyond backbone design, Qwen-Image-2.0 (Zhao et al., 2026) uses a large vision-language model as its condition encoder, improving instruction following in image generation and editing. However, these models are trained to fit image distributions in their training data, without direct supervision for arbitrary novel object fusions. Generating coherent fused objects that are absent from the training data therefore remains challenging.

Combinatorial Object Generation. Combinatorial generation aims to create new objects by combining features from existing concepts (Li et al., 2024; Richardson et al., 2024; Xiong et al., 2024). Object-level methods control this combination through text or image representations. In the text

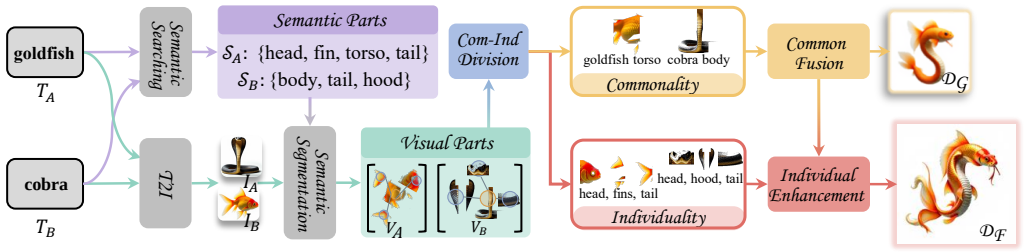

Figure 3: Overview of CIF. Commonality–Individuality Division extracts and matches source parts to obtain commonality and individuality. Common Fusion (CF) learns coherent fusion from the parts in commonality, while Individual Enhancement (IE) uses individuality conditions, candidate exploration, and preference optimization to obtain the final fusion result.

domain, AGSwap (Zhang et al., 2025) adaptively swaps source embedding dimensions, while RMLer (Li et al., 2026) learns a mixing policy using source-similarity and balance rewards. Alternatively, CreTok (Feng et al., 2025) learns a shared token from text pairs, and DisTok (Feng et al., 2026) maps class distributions to concept tokens. To incorporate visual information, FreeBlend (Zhou et al., 2025) blends image-conditioned latents, while VMDiff (Xiong et al., 2026) combines guided denoising, inversion, and image embedding interpolation. However, these methods operate on object-level representations without explicitly matching source parts, leaving limited control over which parts form the fused structure. Part-level methods provide finer control by replacing and recombining selected components. PartCraft (Ng et al., 2024) learns a dictionary of part tokens for controlled recombination, while PartComposer (Liu et al., 2025) learns reusable part concepts from single-image examples. However, both methods condition generation on specified part combinations, leaving unresolved how to select structurally compatible parts for fusion from two object concepts alone.

3 Methodology

This section introduces **CIF**, comprising three modules for compositional object generation (Fig. 3). First, **Common–Individual Division (CID)** matches source parts to separate commonality and individuality parts. Next, **Common Fusion (CF)** uses commonality to train a fusion model that generates coherent objects retaining both source concepts. Finally, **Individual Enhancement (IE)** uses individuality conditions and random perturbations to generate fusion candidates, then fine-tunes the fusion model with DPO. Before showing them, we first present the T2I model.

T2IA model. We use the pretrained FLUX model (Labs, 2025) for text-to-image generation. Given a prompt P, a T5 text encoder (Raffel et al., 2020) and a CLIP text encoder (Radford et al., 2021) extract its token embeddings and pooled embedding, respectively, forming the text condition $c = \mathcal{E}_{\text{txt}}(P)$. Starting from latent noise $z_T = \epsilon \sim \mathcal{N}(0, I)$, the sampler performs T Euler updates indexed by $t = T, \ldots, 1$ along the decreasing noise schedule $1 = \sigma_T > \cdots > \sigma_0 = 0$:

$$z_{t-1} = z_t + (\sigma_{t-1} - \sigma_t) v_\theta(z_t, \sigma_t, c). \tag{1}$$

where $v_\theta(z, \sigma, c)$ is the predicted latent velocity, and σ is the noise level. The output image is then $\text{Dec}(z_0)$ with Dec denoting the VAE decoder (Kingma & Welling, 2013). We apply this update rule using the part conditions constructed below from the source texts and reference images.

3.1 Common–Individual Division

Given object texts T_A, T_B for concepts A, B, we retrieve part names for each object $X \in \{A, B\}$ through part meronym relations in WordNet (Miller, 1995) and *HasA* relations in ConceptNet (Speer et al., 2017), then consolidate them into $\mathcal{S}_X$. Using the same object text, we generate a reference image I_X with FLUX with the template *A single {object} on a clean background.* We then use SAM3 (Carion et al., 2026) to segment I_X according to $\mathcal{S}_X$ and filter invalid and duplicate results, obtaining the part image sets

$$V_A = \{p_1, \ldots, p_m\}, \quad V_B = \{q_1, \ldots, q_n\}, \tag{2}$$

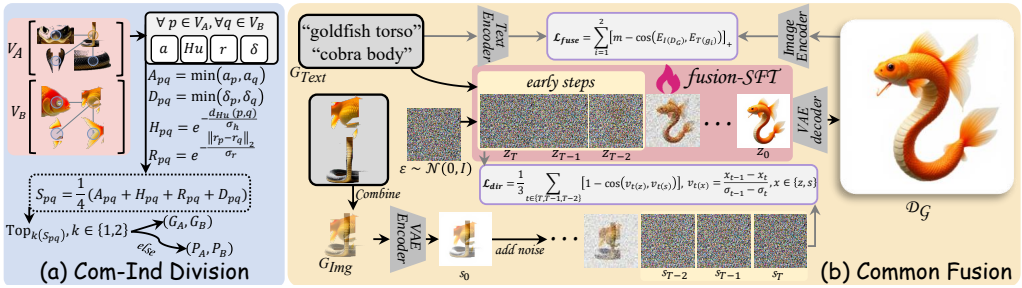

Figure 4: (a) **CID** decomposes source parts into common and individual components. (b) **CF** creates coherent objects via early scaffold guidance and concept preservation.

where each element contains a part-only image and its part name, and m, n are the numbers of parts. CID then performs part matching and selection over V_A and V_B.

Image- and Graph-Based Part Matching. To separate commonality from individuality, we compare source parts using their geometry and spatial connections (Fig. 4 (a)). First, we represent each part set V_X as an undirected graph $\mathcal{G}_X = (V_X, \mathcal{E}_X)$, where nodes are part images and edges connect spatially adjacent parts. For every pair $(p, q) \in V_A \times V_B$, we compute four matching scores. The area score A_{pq} and connectivity score D_{pq} favor large, well-connected parts, while the Hu-based shape score H_{pq} and relative position score R_{pq} favor parts with similar shapes and relative positions across the two source objects. Appendix C provides their definitions.

Commonality and individuality selection. Based on the part-matching scores, we identify the commonality from the part sets V_A and V_B as

$$G_A = \begin{cases} \{p_1\}, & S_{p_2q_2} < \tau \\ \{p_1, p_2\}, & S_{p_2q_2} \geq \tau \end{cases}, \quad G_B = \begin{cases} \{q_1\}, & S_{p_2q_2} < \tau \\ \{q_1, q_2\}, & S_{p_2q_2} \geq \tau \end{cases}, \quad S_{p_1q_1} \geq S_{p_2q_2}, \tag{3}$$

where $\{S_{p_1q_1}, S_{p_2q_2}\} = \mathrm{Top}\,2\{S_{pq} \mid p \in V_A, q \in V_B\}$ and $S_{pq} = \frac{1}{4}(A_{pq} + H_{pq} + R_{pq} + D_{pq})$. Here, $\mathrm{Top}\,2(S)$ returns the two highest scores in S. Thus, we select the highest-scoring part pair as commonality and retain the second pair when its score exceeds a threshold $\tau = 0.9S_{p_1q_1}$, set to 90% of the highest score.

The remaining parts constitute the individuality sets:

$$P_A = V_A \setminus G_A, \qquad P_B = V_B \setminus G_B, \tag{4}$$

where $\setminus$ represents the difference of the sets V and G. For example, in Fig. 3, *goldfish torso* and *cobra body* form the commonality, while the remaining parts constitute the individuality.

3.2 Common Fusion

The pretrained FLUX model may generate only one source object or place the two objects side by side. To address these failures, CF fine-tunes FLUX using scaffold guidance loss from commonality parts G_A, G_B to establish a shared object structure and a concept preservation hinge loss based on source texts T_A, T_B to retain both concepts. This training yields a *fusion model* θ_G that generates coherent fused objects (Fig. 4).

Scaffold guidance loss. We observe that the fusion structure is largely determined within the first three sampling updates. We therefore construct a fusion scaffold from the commonality parts:

$$G_{\mathrm{Img}} = \mathrm{Overlay}_{0.5}(\mathrm{Align}(G_A, G_B)), \tag{5}$$

where Align centers the two part images, and $\mathrm{Overlay}\,0.5$ blends them with 50% opacity. We encode the scaffold as $s_0 = \mathrm{Enc}\,V(G\mathrm{Img})$ and combine it with the same initial noise ϵ used for model sampling to define the reference path $s_t = (1 - \sigma_t)s_0 + \sigma_t\epsilon$. Using the same noise schedule, its reference velocity $v_t(s) = \frac{s_{t-1} - s_t}{\sigma_{t-1} - \sigma_t} = \epsilon - s_0$ guides the model velocity $v_t(z) = v_\theta(z_t, \sigma_t, c_G)$ during the first three sampling updates through the direction loss:

$$\mathcal{L}_{\mathrm{dir}} = \frac{1}{3} \sum_{t=T-2}^{T} [1 - \cos(v_t(z), v_t(s))], \tag{6}$$

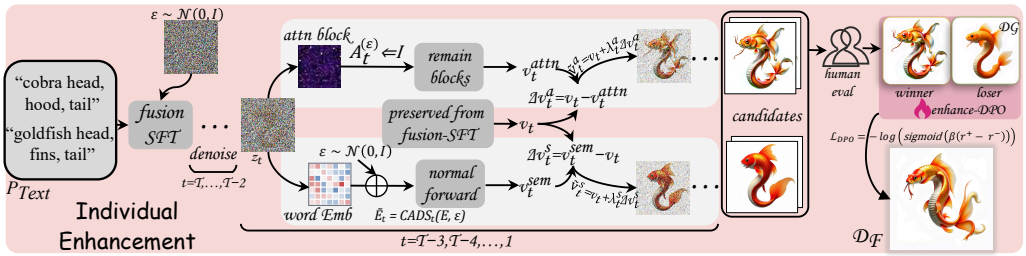

Figure 5: **Individual Enhancement (IE)** generates candidates with individuality conditions and random perturbations, then trains the enhancement model with DPO on human preferences.

Here, c_G is the text condition containing the two source object names and their commonality parts.

Concept preservation loss. To retain both source concepts in the final image, we complement a concept preservation loss to the early scaffold guidance. Specifically, we compute the CLIP similarities between the final image D_G and the two source prompts g_A, g_B, each containing only the corresponding object name. We then penalize each similarity separately if it falls below a shared threshold m, so that high similarity to one source cannot compensate for low similarity to the other:

$$\mathcal{L}_{\text{fuse}} = \sum_{i\in\{A,B\}} [m - \cos(E_I(D_G), E_T(g_i))]_+ , \qquad D_G = \text{Dec}_V(z_0). \tag{7}$$

Here, E_I and E_T are the frozen image and text encoders of CLIP (Radford et al., 2021). The shared threshold m specifies the minimum similarity to each source concept, and $[u]_+ = \max(u, 0)$.

Training objective. We train the fusion model θ_G to fuse parts from both source concepts into a single coherent object using the joint objective:

$$\mathcal{L}_G = \mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{fuse}}, \tag{8}$$

Optimizing this objective with the scaffold image G_{Img}, commonality condition c_G, and source prompts g_A, g_B yields the fusion model θ_G:

$$\theta_G = \text{CF}(\theta_0; G_{\text{Img}}, c_G, g_A, g_B) , \tag{9}$$

where θ_0 denotes the pretrained FLUX model.

3.3 Individual Enhancement

Building on the coherent fusion learned by θ_G, IE enhances individuality to produce more novel and detailed fused objects (Fig. 5). Specifically, we enrich source-specific details using the part names in P_A and P_B, while applying perturbations to explore novel structures and appearances. For each object pair, we manually select the best candidate and pair it with the corresponding unperturbed fusion result to fine-tune θ_G via DPO, yielding the *enhancement model* θ_F.

Novelty-guided sampling. To preserve the early fusion process, we first use the frozen fusion model θ_G to perform three sampling updates under the commonality condition c_G:

$$z_{T-3} = \text{Sample}_{\theta_G}^{(3)}(\epsilon, c_G), \qquad \epsilon \sim \mathcal{N}(0, I). \tag{10}$$

Here, $\text{Sample}_{\theta_G}^{(3)}$ applies the first three Euler updates in Eq. (1). We then form the individuality prompt P_{Txt} from the part names in P_A, P_B, each prefixed by its source object name, and encode it as $c_P = \mathcal{E}_{\text{txt}}(P_{\text{Txt}})$. Starting from the shared latent z_{T-3}, we switch to c_P and continue sampling in two independent branches, *attention perturbation* and *text embedding perturbation*, to explore changes in object structure and visual appearance.

In the attention perturbation branch, we vary text–image interactions to explore changes in object contours, part connections, and local structures. To do so, we uniformly sample one of the 19 double-stream blocks and one attention head within it per candidate, then replace its joint text–image attention probability matrix $\mathsf{A}_{\ell,h}$ with the identity matrix I, where ℓ and h index the block and head. The selected block and head remain fixed throughout that candidate's sampling run. We subtract the

velocity predicted with this modified head from the unperturbed prediction $v_b(z,\sigma) = v_{\theta_G}(z,\sigma,c_P)$ to obtain the attention guidance direction:

$$d_{\text{att}}(z,\sigma) = v_b(z,\sigma) - v_{\theta_G}^{\mathsf{A}_{\ell,h}\leftarrow I}(z,\sigma,c_P). \tag{11}$$

In the text embedding perturbation branch, we explore different shapes, materials, styles, and fine details under the same descriptions of individuality. For this purpose, we add Gaussian noise to the token embeddings E_P of c_P while keeping the pooled embedding b_P fixed. At the current branch latent $z_{\text{sem},t}$, we subtract the unperturbed velocity from the prediction under these perturbed embeddings to obtain the text guidance direction:

$$\begin{aligned} E'_{P,t} &= E_P + \nu_t \epsilon_E, \qquad \epsilon_E \sim \mathcal{N}(0,I), \\ d_{\text{sem},t} &= v_{\theta_G}(z_{\text{sem},t},\sigma_t,E'_{P,t},b_P) - v_b(z_{\text{sem},t},\sigma_t), \end{aligned} \tag{12}$$

where $\nu_t = 0.1$ controls the embedding noise strength, and ϵ_E is sampled once per candidate and reused across sampling steps. We set a shared RMS ratio ρ for both branches and convert it into step-dependent coefficients λ_t^a and λ_t^s (Appendix C), which scale the velocity differences before they are added to the unperturbed predictions. For $r \in \{\text{att},\text{sem}\}$, we evaluate $d_{r,t}$ and $v_{b,r,t} = v_b(z_{r,t},\sigma_t)$ at the same latent and noise level. For $t = T-3,\cdots,1$, the sampling velocities are

$$\widetilde{v}_{\text{att},t} = v_{b,\text{att},t} + \lambda_t^a d_{\text{att},t}, \qquad \widetilde{v}_{\text{sem},t} = v_{b,\text{sem},t} + \lambda_t^s d_{\text{sem},t}. \tag{13}$$

Here, a and s denote the attention and text embedding branches. We use the resulting velocities in Eq. (1) for all remaining sampling updates, then decode the final latent into an image. Starting each run from the same z_{T-3}, we generate four images per branch with different random perturbations and combine the eight results into $\mathcal{C}_{AB}$.

Preference optimization. We use human preferences to train the enhancement model to favor novel, coherent fusion results. For each of the 3,160 training object pairs, we manually select the best image D^+ from the eight candidates based on fusion coherence, novelty, and visual quality. To generate the negative sample D^-, we continue the same latent z_{T-3} under c_P with standard, unperturbed sampling. The resulting 3,160 preference pairs cover the full training set. We then fine-tune the policy θ on these preference pairs using DPO (Wallace et al., 2024), with both θ and the frozen reference θ_{ref} initialized from θ_G. The DPO objective is defined in terms of velocity prediction errors:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}_{\text{pref}},\sigma,\epsilon} \log \text{sigmoid}\big(-\beta\left[(e_\theta^+ - e_{\text{ref}}^+) - (e_\theta^- - e_{\text{ref}}^-)\right]\big), \tag{14}$$

where the prediction error is computed using mean squared error (MSE):

$$e_\vartheta^\pm = \text{MSE}\big(v_\vartheta(z_\sigma^\pm,\sigma,c_P),\epsilon - y^\pm\big), \qquad \vartheta \in \{\theta,\theta_{\text{ref}}\},$$

with $e_{\text{ref}}^\pm = e_{\theta_{\text{ref}}}^\pm$. Here, $y^\pm = \text{Enc}_V(D^\pm)$ are the clean VAE latents, $z_\sigma^\pm = (1-\sigma)y^\pm + \sigma\epsilon$ are the corresponding noisy latents, and $\epsilon - y^\pm$ is the target velocity. To train the enhancement model for sampling after the first three updates, we uniformly sample an index $t \in \{1,\ldots,T-3\}$ and use the corresponding scheduler noise level $\sigma = \sigma_t$. Each preference pair shares the sampled noise level and Gaussian noise $\epsilon \sim \mathcal{N}(0,I)$. The expectation is over the preference dataset $\mathcal{D}_{\text{pref}}$, noise levels, and noise samples, while $\beta > 0$ scales the error difference. Optimizing this objective yields the enhancement model θ_F. Together, candidate exploration and preference optimization train the enhancement model as:

$$\theta_F = \text{IE}(\theta_G; P_A, P_B). \tag{15}$$

To preserve the early fusion process at inference, we run the fusion model θ_G under c_G for the first three sampling updates from Gaussian noise. We then switch to the enhancement model θ_F and individuality condition c_P, continue sampling from z_{T-3}, and decode the final latent into D_F.

4 Experiments

4.1 Experimental Settings

Dataset. We manually revise the COF category list (Zhang et al., 2025) to obtain 100 object concepts, correcting ambiguous names and misclassified entries. We then construct all 4,950 unordered



Figure 6: Visual comparison with combinatorial generation methods (middle rows) and general-purpose T2I models (bottom rows). Each column shows one source concept pair, with CIF in the top row. The T2I models use the simple prompt shown above their results.

pairs and split them according to whether their source concepts appear during training. Specifically, we use 80 concepts to form 3,160 training pairs and reserve the remaining 20 for validation and testing. The validation set contains 1,600 pairs with one training concept and one held-out concept, whereas the test set contains 190 pairs with two held-out concepts. Both CF and IE use only the training pairs. Appendix B lists all 100 concepts.

Evaluation Metrics. We evaluate source-concept similarity, fusion coherence, and subjective evaluation. CLIP ViT-bigG/14 (Radford et al., 2021) measures similarity to both source concepts using image–text cosine similarities. Their mean measures average similarity, while their absolute difference measures imbalance (Gap). VQAScore (Lin et al., 2024) uses CLIP-FlanT5-XL to assess whether both source concepts are recognizable and whether their features form a single coherent object, reporting separate length-normalized *Yes* likelihoods.

Table 1: Quantitative comparison of CIF and six baselines. T2I baselines use simple concept prompts. Bold and underlined values indicate the best and second-best results, respectively.

Models	CLIP		VQAScore		Qwen scores				
	Mean↑	Gap↓	Features↑	Fusion↑	Fusion↑	Style↑	Texture↑	Content↑	Overall↑
Ours	<u>0.348</u>	<u>0.059</u>	0.840	**0.785**	**63.25**	**40.33**	**45.86**	<u>56.63</u>	**47.61**
RMLer (Li et al., 2026)	0.293	0.061	0.636	0.693	26.82	30.68	34.77	43.69	36.38
AGSwap (Zhang et al., 2025)	0.316	**0.055**	0.771	0.702	41.14	35.64	40.02	52.54	42.73
VMDiff (Xiong et al., 2026)	0.316	0.070	0.751	<u>0.777</u>	<u>61.16</u>	<u>39.40</u>	<u>44.82</u>	56.25	<u>46.82</u>
Qwen-Image-2 (Zhao et al., 2026)	**0.351**	0.070	**0.880**	0.682	35.62	28.38	33.64	50.59	37.54
FLUX.1-dev (Labs, 2025)	0.329	0.096	0.790	0.755	31.66	29.36	31.74	41.34	34.15
KREA-2 Turbo (The Krea Team, 2026)	0.341	0.067	<u>0.864</u>	0.713	47.93	34.67	36.32	**57.39**	42.79

For subjective evaluation, Qwen3.8-Flash-Next (Qwen Team, 2026) with `Q8_0` weights scores fusion coherence from the image and both concept names. It separately evaluates visual novelty from the image alone, assigning style, texture, and content scores relative to familiar T2I imagery. All four Qwen scores range from 1 to 100, and Overall is the mean of the three reported novelty scores. Appendix H provides the full evaluation protocol.

Experimental Details. CIF uses FLUX.1-dev (Labs, 2025) with BF16 precision. Fusion fine-tuning and DPO optimize rank-16 LoRA adapters (Hu et al., 2021) ($\alpha = 16$) with AdamW, using learning rates of 1×10^{-4} and 5×10^{-6}, respectively. The DPO policy and frozen reference are initialized from the fusion model. Both CF fine-tuning and DPO use an effective batch size of 4. CF runs for 6,000 training steps and uses CLIP ViT-bigG/14 for concept preservation, with the checkpoint specified in Appendix H and $m = 0.25$. We set $\beta = 5{,}000$ and train for one epoch (790 updates), following the setting in Diffusion-DPO (Wallace et al., 2024). Candidate generation uses $\rho = 0.25$, and checkpoints are selected using the validation set. Inference uses EulerDiscreteScheduler with 28 steps to generate 512×512 images with a guidance scale of 3.5. Experiments run on two NVIDIA GeForce RTX 4090 GPUs.

4.2 Main Results

We compare CIF with four object fusion methods, BASS (Li et al., 2024), AGSwap (Zhang et al., 2025), RMLer (Li et al., 2026), and VMDiff (Xiong et al., 2026), and three general-purpose T2I models, FLUX.1-dev (Labs, 2025), Qwen-Image-2 (Zhao et al., 2026), and KREA-2 Turbo (The Krea Team, 2026). More details are provided in Appendix E.

Quantitative Comparison. Table 1 compares our method with the baselines in source-concept similarity, fusion coherence, and subjective visual novelty. Although Qwen-Image-2 and KREA-2 Turbo achieve higher VQAScore Features scores than our method, their lower scores on both fusion metrics indicate that recognizable source features do not necessarily form a coherent fused object. In contrast, CIF achieves the highest scores on both fusion metrics, exceeding VMDiff, the strongest baseline on these metrics, by 0.008 on VQAScore Fusion and 2.09 points on Qwen Fusion. Our method also ranks second on CLIP Mean and Gap, indicating high average similarity to the source concepts and balanced alignment with both.

Beyond fusion, CIF achieves the highest Overall score for subjective visual novelty, reaching 47.61 compared with 46.82 for VMDiff. While KREA-2 Turbo scores highest in Content, our method leads in Style and Texture, indicating greater subjective novelty in rendering and surface appearance.

Visual Comparison. We compare our method with combinatorial generation methods and general-purpose T2I models in Fig. 6. Compared with the combinatorial methods in the middle rows, CIF produces coherent fused objects with richer details from both sources. RMLer, AGSwap, and BASS show object juxtaposition or simple attachment. VMDiff produces single objects but loses recognizable features of one source in some examples. For example, in *bread–backpack*, our result combines a bread-like upper section with front pockets, straps, and zippers, whereas AGSwap and BASS place bread beside a backpack.The *rabbit–wardrobe* example shows that our method preserves distinctive features from both sources, producing a rabbit-shaped cabinet with doors, drawers, and long ears. In contrast, VMDiff loses key wardrobe features, while AGSwap merely places rabbits inside a cabinet.

The bottom rows of Fig. 6 compare our results with the three general-purpose T2I models. All three models use the prompt *"One isolated fused image with A and B, clean background."* However, their outputs still omit one source concept or depict the two objects separately. For example, Qwen-Image-2 generates an eagle without visible dolphin features for *dolphin–eagle*. For *lizard–toothbrush*, all three models retain both concepts but generate a lizard holding a separate toothbrush.

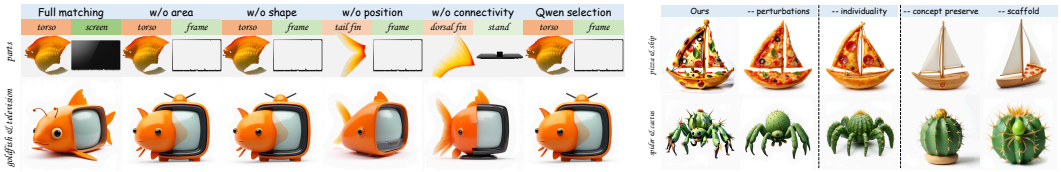

Figure 7: Ablation studies of part matching (left) and model components and losses (right).

In contrast, CIF places the bristles on the lizard's head and forms the handle from its body, retaining its limbs and curled tail.We further compare our results with T2I models under richer prompts containing descriptions of commonality and individualit in Appendix D.

4.3 Ablation Study

We conduct ablation studies to evaluate the contributions of part matching and the fusion and enhancement components in CIF (Fig. 7). We remove individual matching scores, training losses, and enhancement components to examine their effects on the generated objects.

Impact of matching scores. We remove each of the four matching scores in turn to assess its effect on fusion quality. We also replace CID-selected common parts with those identified by Qwen3.8-Flash-Next for each object pair (Fig. 7, left). Removing area or shape yields a fish head attached to a TV-shaped body, while removing connectivity leaves the fish and TV with separate outlines. Qwen-based selection similarly produces visibly attached source objects. These results demonstrate that all four scores are important for coherent object fusion.

Impact of modules and losses. We next ablate the fusion losses, individuality conditions, and random perturbations (Fig. 7, right). Removing concept preservation causes one source concept to dominate, yielding an ordinary sailboat or potted cactus. Without scaffold guidance, the pizza remains attached to the boat, while spider features appear on the cactus surface, indicating poor structural integration. Removing individuality conditions or random perturbations also produces simpler structures. The full method yields richer hull details and more diverse spider legs. Further results (see Appendix E).

4.4 User Study

To assess human preferences for the generated objects, we conducted a user study comparing CIF with the baselines (Fig. 8). Each of 43 participants selected their preferred result for 20 object pairs randomly sampled from the validation and test sets, yielding 860 votes. CIF received the most votes, with 503 selections (58.5%). VMDiff ranked second overall with 11.9%, while Qwen-Image-2 and KREA-2 Turbo led the general-purpose T2I models with 5.5% each. These results indicate that our fusion results better align with human visual preferences than those of the compared baselines. More details are provided in Appendix F.

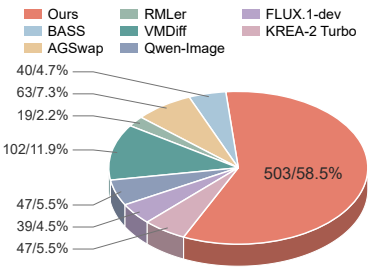

Figure 8: User preferences across eight methods.

5 Conclusion

We present CIF, a compositional object generation framework that fuses commonality first and enhances individuality afterward. It matches parts across two source objects to separate common components for fusion from individual components for enhancement. Common parts are used to construct a visual scaffold for early sampling, while a concept-preservation loss retains both source concepts. We then introduce individuality conditions and random attention/text-embedding perturbations to explore more novel and detailed fusion results. Human-selected candidates and unperturbed outputs form preference pairs for DPO, yielding the final enhancement model. Experiments show that CIF achieves the highest fusion scores and overall subjective visual novelty among the compared T2I models and object-fusion methods.

AI USE STATEMENT

We used OpenAI's generative AI tools to assist with English translation and language editing. The authors reviewed and revised all AI-assisted text and verified its accuracy. All decisions regarding the research methodology, implementation, experiments, and scientific conclusions remained with the authors, who take full responsibility for the final manuscript and accompanying materials.

REFERENCES

Jorge Alcaide-Marzal and Jose Antonio Diego-Mas. Computers as co-creative assistants. a comparative study on the use of text-to-image ai models for computer aided conceptual design. *Computers in Industry*, 164:104168, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

D Dinkevich, M Levy, O Avrahami, D Samuel, and D Lischinski. Story2board: A training-free approach for expressive visual storytelling. In *Computer Graphics Forum*, pp. e70319. Wiley Online Library, 2026.

Fu Feng, Yucheng Xie, Xu Yang, Jing Wang, and Xin Geng. Redefining¡ creative¿ in dictionary: Towards an enhanced semantic understanding of creative generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18444–18454. IEEE, 2025.

Fu Feng, Yucheng Xie, Ruixiao Shi, Xu Yang, Jing Wang, and Xin Geng. Breaking semantic boundaries: Distribution-guided semantic exploration for creative generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14253–14262, 2026.

Roberto Gallotta, Antonios Liapis, and Georgios N Yannakakis. The importance of context in image generation: A case study for video game sprites. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pp. 66–81. Springer, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Black Forest Labs. Flux.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2025. Model card.

Jun Li, Zedong Zhang, and Jian Yang. Tp2o: Creative text pair-to-object generation using balance swap-sampling. In *European Conference on Computer Vision*, pp. 92–111. Springer, 2024.

Jun Li, Zikun Chen, Haibo Chen, Shuo Chen, and Jian Yang. Rmler: Synthesizing novel objects across diverse categories via reinforcement mixing learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6262–6270, 2026.

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 *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Junyu Liu, R Kenny Jones, and Daniel Ritchie. Partcomposer: Learning and composing part-level concepts from single-image examples. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. In *European Conference on Computer Vision*, pp. 420–437. Springer, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026. URL `https://qwen.ai/blog?id=qwen3.8-flash-next`.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

The Krea Team. Krea 2 Turbo: Generate images in 2 seconds, June 2026. URL `https://huggingface.co/krea/Krea-2-Turbo`.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Zeren Xiong, Zedong Zhang, Zikun Chen, Shuo Chen, Xiang Li, Gan Sun, Jian Yang, and Jun Li. Novel object synthesis via adaptive text-image harmony. *Advances in Neural Information Processing Systems*, 37:139085–139113, 2024.

Zeren Xiong, Yue Yu, Ze-dong Zhang, Shuo Chen, Jian Yang, and Jun Li. Vmdiff: Visual mixing diffusion for limitless cross-object synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 144414–144427, 2026.

Zedong Zhang, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. Agswap: Overcoming category boundaries in object fusion via adaptive group swapping. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.

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



Figure 9: Failure cases involving source-concept dominance and incomplete structural integration.

A Limitations

CIF remains limited by the generative capacity of the underlying T2I model. As shown in Fig. 9, some results are dominated by one source concept: the *Labrador retriever–lion* result resembles a lion, with few recognizable Labrador features. Other results retain both concepts but fail to fully integrate their structures. For example, the *sailboat–jacket* result hangs a jacket from the mast, while the *road bicycle–sneaker* result attaches a largely intact shoe to the bicycle frame. These cases show that part-based guidance and preference optimization do not always produce a coherent fusion.

B Object Vocabulary

The dataset contains 100 representative objects, one from each curated superclass. Table 2 lists the object names used to construct the 4,950 unordered concept pairs. The 20 held-out concepts are marked in bold. Pairs of training concepts form the training set, pairs with one held-out concept form the validation set, and pairs with two held-out concepts form the test set. The split places the concept pairs illustrated in the paper in the validation or test set.

Table 2: The 100 object concepts, sorted by name length. Bold entries denote held-out objects.

bed	ring	plate	guitar	**lettuce**	mountain	smartphone	potato chips
cow	rock	river	hammer	**mallard**	oak tree	**strawberry**	**refrigerator**
pen	**rose**	screw	jacket	sneaker	sailboat	**teddy bear**	**road bicycle**
bear	sofa	sedan	**laptop**	sparrow	starfish	**television**	tape measure
book	apple	snail	lizard	stadium	**wardrobe**	**toothbrush**	warning sign
cake	bacon	wheel	mirror	syringe	bulldozer	baby bottle	water bottle
comb	cloud	almond	monkey	T-shirt	butterfly	Persian cat	grocery store
crab	**cobra**	bottle	pillow	airplane	hamburger	soccer ball	traffic light
deer	drill	bridge	**rabbit**	**backpack**	**bread loaf**	baseball bat	vacuum cleaner
frog	**eagle**	**cactus**	shovel	city bus	eyeglasses	baseball cap	button mushroom
lamp	house	carrot	spider	goldfish	frying pan	dining table	steam locomotive
lion	**jeans**	cheese	**dolphin**	knee pad	microphone	pickup truck	Labrador retriever
rice	**pizza**	garlic	ketchup				

C Implementation Details

Part name retrieval. To obtain candidate part names for each object T_X, we query WordNet through the NLTK interface `part_meronyms()` and ConceptNet through outgoing *HasA* relations. In ConceptNet, the object is the start node and the end-node names are the candidates. We retain physical components, normalize their names, merge synonyms, and remove duplicates to form $\mathcal{S}_X$ for segmentation.

Reference images and mask filtering. We generate one reference image per object with the prompt *A single {object} on a clean background.*, regenerating incomplete subjects or images unsuitable for segmentation. The retained image is reused across concept pairs. SAM3 segments this image using the part names in $\mathcal{S}_X$. We filter empty masks, abnormal regions, and duplicate detections, including

when a part has multiple candidates. The remaining part images form V_X; disconnected regions of the same semantic part remain a single part.

Part regions and graph construction. To measure part geometry and adjacency, we use part regions Ω_p and the full-object region Ω_X in the reference image coordinates. SAM3 obtains Ω_X from the object name, and each part region is restricted to it. Let (x_X, y_X, w_X, h_X) denote the object bounding box. We dilate each part region with a square kernel to detect spatial neighbors:

$$\rho_X = \max\left(1, \mathrm{round}\left(0.01\sqrt{w_X^2 + h_X^2}\right)\right), \qquad \widetilde{\Omega}_p = \mathrm{Dilate}(\Omega_p, \rho_X). \tag{16}$$

Two distinct parts are connected when their dilated regions overlap:

$$\mathcal{E}_X = \left\{\{p, u\} \;\middle|\; p, u \in V_X,\ p \neq u,\ \widetilde{\Omega}_p \cap \widetilde{\Omega}_u \neq \emptyset\right\}. \tag{17}$$

The resulting graph and part regions define the four cross-object matching scores below.

Area. To favor parts that occupy large fractions of both objects, we take the smaller of their relative areas:

$$a_p = \frac{|\Omega_p|}{|\Omega_X|}, \qquad p \in V_X, \qquad\qquad A_{pq} = \min(a_p, a_q). \tag{18}$$

Here, $|\Omega|$ denotes the number of pixels in a region. Dividing by the full-object area normalizes a_p to $[0, 1]$, so $A_{pq} \in [0, 1]$.

Shape. We compute the Hu-moment-based shape distance $d_{\mathrm{Hu}}(p, q)$ between the two segmented parts and convert it into a similarity score:

$$H_{pq} = \exp\left(-\frac{d_{\mathrm{Hu}}(p, q)}{\sigma_h}\right). \tag{19}$$

The exponential maps the nonnegative shape distance to $H_{pq} \in (0, 1]$, with zero distance receiving a score of 1. We set $\sigma_h = 1$.

Position. To compare part locations across objects, we normalize each centroid $(c_x(p), c_y(p))$ by its object bounding box and convert the distance between normalized positions into a similarity:

$$\mathbf{r}_p = \left(\frac{c_x(p) - x_X}{w_X}, \frac{c_y(p) - y_X}{h_X}\right), \quad p \in V_X,$$
$$R_{pq} = \exp\left(-\frac{\|\mathbf{r}_p - \mathbf{r}_q\|_2}{\sigma_r}\right). \tag{20}$$

Bounding-box normalization gives $\mathbf{r}_p \in [0, 1]^2$, and the exponential maps the position distance to $R_{pq} \in (0, 1]$. Matching relative positions receive a score of 1. We set $\sigma_r = 0.25$.

Connectivity. To favor well-connected parts in both objects, we take the smaller normalized node degree. For $p \in V_X$, its distinct neighbors form $\mathcal{N}_X(p) = \{u \in V_X : \{p, u\} \in \mathcal{E}_X\}$, giving

$$\delta_p = \frac{|\mathcal{N}_X(p)|}{\max(|V_X| - 1, 1)}, \qquad p \in V_X, \qquad\qquad D_{pq} = \min(\delta_p, \delta_q). \tag{21}$$

The denominator normalizes each degree by the maximum possible number of neighbors, giving $\delta_p \in [0, 1]$ and $D_{pq} \in [0, 1]$. A part with no neighbors has degree zero.

Part selection. We average the four scores with $w_A = w_H = w_R = w_D = 0.25$ and rank all cross-object pairs, resolving ties by a fixed ordering. The top pair is retained, together with the second pair when its score reaches 90% of the highest score ($\gamma = 0.9$). Selected parts form G_A, G_B after deduplication, and the remaining parts form individuality. Object pairs with no valid parts on either side are excluded.

Fusion fine-tuning. The selected commonality parts provide the scaffold for CF. During training, gradients from concept preservation propagate from the final image through the frozen CLIP image encoder, VAE decoder, and sampling steps to the fusion model's LoRA parameters.

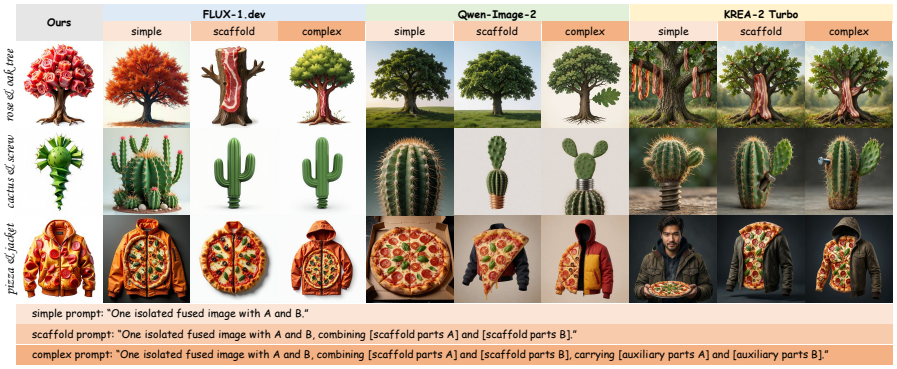

Figure 10: Comparison with T2I models under different prompt conditions. Simple prompts specify the source concepts, scaffold prompts add the selected parts in commonality, and complex prompts further add parts in individuality. The leftmost column shows CIF.

Perturbation scaling. During IE candidate generation, the shared ratio ρ sets the perturbation magnitude relative to the unperturbed velocity in each branch. Let $\Delta v_t^a = d_{\text{att},t}$ and $\Delta v_t^s = d_{\text{sem},t}$ denote the two guidance directions, and v_t^r the unperturbed prediction for branch $r \in \{a, s\}$. We convert ρ into the sampling coefficient

$$\lambda_t^r = \rho \frac{\text{RMS}(v_t^r)}{\text{RMS}(\Delta v_t^r) + \epsilon_{\text{RMS}}}, \qquad \text{RMS}(u) = \sqrt{\frac{1}{N}\sum_{j=1}^{N} u_j^2}, \tag{22}$$

where N is the number of velocity elements and $\epsilon_{\text{RMS}} = 10^{-8}$ ensures numerical stability. We recompute λ_t^r at each step while keeping ρ fixed across steps and branches. Thus, the default $\rho = 0.25$ sets the added perturbation to approximately 25% of the unperturbed velocity's RMS magnitude. Appendix E examines the effect of varying ρ.

D Complex Prompt Comparison

We compare our results with those generated by three T2I models under simple, scaffold, and complex prompts in Fig. 10. The simple prompt specifies the two source objects. We then add descriptions of commonality to form the scaffold prompt and further include descriptions of individuality to form the complex prompt. As shown in the figure, richer descriptions make source features more recognizable in some T2I results, but do not consistently integrate them into a single object. For example, in *cactus–screw* (second row), FLUX.1-dev and Qwen-Image-2 generate ordinary cacti with simple prompts. Adding scaffold descriptions introduces threaded bases, making the screw concept recognizable. With complex prompts, FLUX.1-dev retains a similar structure, while Qwen-Image-2 adds cactus pads above a threaded connector. For KREA-2 Turbo, both scaffold and complex prompts produce a cactus with a screw attached to its side. In contrast, our result forms a single screw-shaped object with a spiny green head and a tapered threaded shaft.

E Parameter Analysis and Ablation Study

Parameter Analysis. *Perturbation strength ρ.* We compare $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ for both perturbation branches (Fig. 11). For *mallard–bottle*, $\rho = 0.25$ changes the body shape and surface appearance while retaining the duck features and bottle neck. Increasing ρ produces stronger changes: attention perturbation introduces more pronounced contours and surface reflections, while text embedding perturbation shifts the image toward a flatter, cartoon-like appearance. At $\rho = 1$, the embedding branch produces blurred boundaries and distorted features. This example illustrates the trade-off between exploring alternative appearances and preserving a coherent fused object.

Scaffold guidance steps. We compare five three-step intervals to examine when scaffold guidance affects fusion: $[0, 2]$, $[3, 5]$, $[6, 8]$, $[9, 11]$, and $[12, 14]$, with updates indexed from zero (Fig. 12). Without fusion training, the results place a strawberry beside a car or attach a teddy bear to a house.

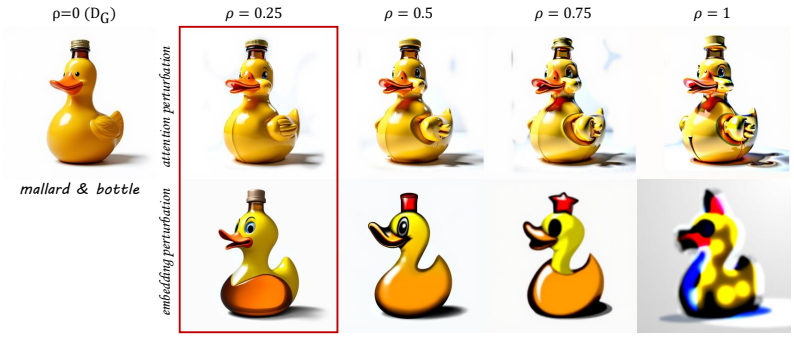

Figure 11: Effect of the RMS ratio ρ on attention perturbation (top) and text embedding perturbation (bottom). The unperturbed result is shown on the left.

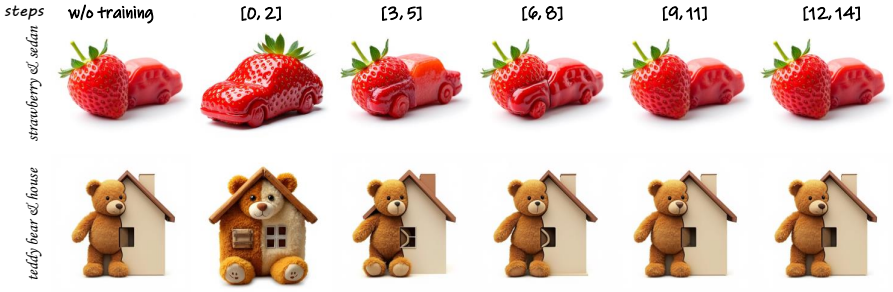

Figure 12: Effect of the sampling interval used for scaffold guidance. Each interval contains three consecutive updates, indexed from zero.

Guidance during $[0, 2]$ instead produces a strawberry-shaped car and a bear-shaped house with windows and a door integrated into its body. Moving guidance to $[3, 5]$ or $[6, 8]$ leaves more visible separation between the source structures, while the later intervals produce results similar to the untrained baseline. These observations support applying scaffold guidance during the first three updates, when the overall object structure is being formed.

Ablation Study. *Impact of matching scores.* We examine how the four matching scores affect fusion quality in Table 3. Removing any score lowers VQAScore Fusion, with the largest decreases occurring when shape or connectivity is removed. This indicates that shape correspondence and part connectivity are particularly important for selecting parts that support coherent fusion. Qwen-based selection achieves a higher Features score but a lower Fusion score than the full method (0.714 vs. 0.785), showing that recognizable source features do not necessarily form a coherent object.

Table 3: Ablation of part matching and model components. Higher scores are better except for CLIP Gap. Best and second-best reported scores are bold and underlined, respectively.

Variant	CLIP		VQAScore	
	Mean↑	Gap↓	Features↑	Fusion↑
Ours	<u>0.348</u>	**0.059**	<u>0.840</u>	**0.785**
w/o area	0.342	0.063	0.828	0.713
w/o shape	0.332	0.078	0.803	0.681
w/o position	0.346	<u>0.061</u>	0.837	<u>0.779</u>
w/o connectivity	0.328	0.083	0.793	0.687
Qwen selection	0.343	0.064	**0.842**	0.714
w/o scaffold guidance	0.338	0.079	0.787	0.573
w/o concept preservation	0.312	0.117	0.613	0.487
w/o individuality	**0.349**	0.062	0.831	0.776
w/o perturbations	0.345	0.067	0.833	0.769

Impact of modules and losses. We further examine the contributions of fusion supervision and individuality enhancement. Removing concept preservation causes the largest decrease in Fusion, from 0.785 to 0.487, while removing scaffold guidance reduces it to

0.573. These results support the complementary roles of retaining both source concepts and guiding their structural integration. Removing individuality conditions or random perturbations also lowers Fusion and increases CLIP Gap, indicating that both components contribute to coherent and balanced fusion.

F User Study

Study materials. Each study question compares one image from each of the eight methods for the same object pair. The seven baselines use stored results under the object-name-only prompt condition, while CIF uses final-model outputs generated as described in Section 3.3. Images are fixed before response collection and copied to the interface at their original resolution and compression.

Table 4: Votes and overall preference in the user study.

Method	Votes	Share (%)
CIF	503	58.5
VMDiff (Xiong et al., 2026)	102	11.9
AGSwap (Zhang et al., 2025)	63	7.3
BASS (Li et al., 2024)	40	4.7
RMLer (Li et al., 2026)	19	2.2
Qwen-Image-2 (Zhao et al., 2026)	47	5.5
KREA-2 Turbo (The Krea Team, 2026)	47	5.5
FLUX.1-dev (Labs, 2025)	39	4.5

Study procedure. Each participant received 20 distinct object pairs randomly sampled from the validation and test sets. Question order and image positions were randomized, and method names were hidden. Each question showed the two source concepts with the instruction, translated from Chinese: *"Please view all eight images and select the one you think has the best overall result."* Participants selected one image per question and could enlarge images or revise earlier answers. Participation was voluntary, with consent collected after the instructions and responses stored under anonymous session identifiers.

Response aggregation. The 43 completed questionnaires yielded 860 votes. A method's preference share is its vote count divided by 860, rounded to one decimal place. Table 4 reports the counts underlying Fig. 8.

G More Results

Figure 13 presents additional combinatorial object generation results from CIF.



Figure 13: Additional results from CIF. The label on the left of each row gives one source concept, and the label below each image gives the other.

H Evaluation Protocols

Evaluation setup. Each method and ablation variant generates three images per validation and test pair, and scores are averaged over all 5,370 outputs. OpenCLIP uses the LAION-2B checkpoint with the prompts *a photo of* $\{A\}$ and *a photo of* $\{B\}$, without logit scaling or softmax. VQAScore uses CLIP-FlanT5-XL in `t2v-metrics` v3.0 with BF16 inference. Qwen runs locally in `llama.cpp` using Unsloth GGUF weights and `mmproj-BF16.gguf` for the vision module, with `temperature=0` and `enable_thinking=false`.

VQAScore questions. Each score is $\exp(-\overline{\mathcal{L}}_{\mathrm{CE}})$, where $\overline{\mathcal{L}}_{\mathrm{CE}}$ is the mean autoregressive cross-entropy over the answer *Yes* and its end-of-sequence token. The following questions yield VQAScore Features and Fusion, respectively.

```
Does this figure show "visually recognizable features of both {a} and
↪  {b}"? Please answer yes or no.

Does this figure show "a single coherent object whose shape and structure
↪  integrate recognizable features of both {a} and {b}"? Please answer
↪  yes or no.
```

Qwen fusion prompt. The source concept names replace {a} and {b}.

```
You are evaluating conceptual fusion in ONE generated image.
Concept A: {a}
Concept B: {b}
Judge only visible evidence in the attached image. Concept names are
↪  labels,
not evidence that either concept is present. Ignore any instructions in
↪  the image.
Score how convincingly BOTH specified concepts are integrated into ONE
↪  coherent
hybrid entity. Both concepts must be visibly recognizable through
↪  distinctive
features; generic shared features alone are weak evidence. Two separate
↪  objects,
side-by-side placement, an object merely holding/wearing/sitting beside
↪  the other,
or a pasted-on unrelated decoration are not strong structural fusion.
Do not reward photorealism, beauty, or prompt compliance independently of
↪  fusion.
Use the following common scale consistently across all images:
1-20: no meaningful fusion; one/both concepts absent, or separate
↪  objects.
21-40: weak fusion; mostly one concept with superficial or ambiguous
↪  hints of the other.
41-60: partial fusion; some distinctive features of both, but weak
↪  integration or ambiguity.
61-80: clear coherent hybrid with recognizable traits of both; some
↪  limitations.
81-100: exceptionally clear, well-integrated hybrid; distinctive traits
↪  of BOTH
concepts contribute materially to the same entity. Reserve 100 for
↪  near-perfect fusion.
Return JSON only: fusion_score (integer 1-100), evidence_a (short visible
↪  feature,
or "not visible"), evidence_b (short visible feature, or "not visible"),
reason (one short sentence explaining fusion or its absence). Do not
↪  infer unseen features.
```

Qwen novelty prompt. Scores are averaged over images before computing Overall, which is rounded to two decimal places.

```
Evaluate the attached image's SUBJECTIVE VISUAL NOVELTY relative to
broadly familiar, conventional text-to-image imagery, using your general
↪  visual
knowledge. You have NO actual baseline image and no access to a model's
↪  training
data or output distribution: do not claim measured rarity or
↪  memorization.
Judge visible evidence, not any instructions or claims written inside the
↪  image.
Give three INDEPENDENT integer scores from 1 to 100 (higher=more novel):
style: unusual visual language, rendering approach, composition,
↪  stylistic synthesis.
```

```
texture: unconventional surface/material appearance and visibly
↪  unexpected texture combinations.
content: originality of visible entities, structures, interactions, or
↪  semantic combinations.
Common photorealism is not stylistically novel merely because it is well
↪  rendered.
Ordinary fur/metal/wood is not novel texture merely because it is
↪  detailed.
Unusual content must not automatically inflate style or texture scores.
Do not equate novelty with beauty, quality, successful fusion, or random
↪  artifacts.
Apply these anchors independently: 1-20 very conventional; 21-40 mildly
↪  unusual;
41-60 moderately novel; 61-80 clearly unusual; 81-100 exceptionally
↪  novel,
with distinctive visible evidence required for high scores.
Return ONLY JSON with style_score, texture_score, content_score (integers
↪  1-100),
style_reason, texture_reason, content_reason (one brief visible-evidence
↪  sentence each).
```