BIND-AND-ROUTE: EXPLICIT REFERENCE BINDING AND ATTENTION ROUTING FOR MULTI-SUBJECT PERSONALIZED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-subject personalized generation aims to synthesize multiple customized subjects in a single image while preserving both textual controllability and reference appearance fidelity. However, we observe that existing personalization methods degrade substantially as the number of reference subjects increases, leading to reduced text alignment and weakened appearance consistency. By analyzing multi-modal attention patterns, we find that such degradation is closely related to *Text-Reference Misalignment* and *Subject Confusion*: the text of a subject and its corresponding reference condition may attend to different generation regions, while the target generation regions of a subject may attend to mismatched reference conditions. To address these issues, we propose **Bind-and-Route**, a framework for robust multi-subject personalization. First, we establish explicit text-reference associations through placeholder-reference contrastive learning and bidirectional binding resamplers, which mutually inject the associated textual and reference information to facilitate binding in multi-modal attention. Second, we introduce a lightweight attention router, which uses image-text attention as a localization prior to route image tokens toward matched reference conditions. Experiments demonstrate that our method can achieve strong text alignment and reference appearance fidelity, offering an effective solution under unified multi-modal conditioning.

1 INTRODUCTION

Personalized image generation aims to synthesize images of user-specified subjects while following diverse text prompts. Early methods mainly adapt text-to-image models to a single subject through textual inversion Mokady et al. (2023) or model fine-tuning Ruiz et al. (2023), while later zero-shot methods improve efficiency and flexibility by injecting reference features without per-subject optimization or fine-tuning Wei et al. (2023); Shi et al. (2023); Xiao et al. (2025). Meanwhile, multi-subject personalized generation has attracted increasing attention. Recent methods Tan et al. (2025); Garibi et al. (2025); Mou et al. (2025); Chen et al. (2026a); Wu et al. (2025a); Labs et al. (2025) indicate a broader trend toward unified multi-modal conditioning.

Diffusion transformers (DiTs) provide a natural and elegant formulation for integrating multiple conditions. In these architectures, especially those based on multi-modal attention, text, noisy image, and references can interact within a unified attention space. This design makes it convenient to introduce multiple references as additional condition tokens. However, as shown in Fig. 1 (a), this seemingly straightforward extension becomes unreliable in multi-subject scenarios. As the number of reference subjects increases, existing methods often suffer from degraded text alignment and weakened appearance consistency. Generated images may fail to follow the target prompt precisely, and each subject may fail to faithfully preserve the appearance of its corresponding reference image, even exhibiting severe copy-paste artifacts, where the reference pose or appearance is overly replicated against the intended text semantics.

To better understand this degradation, we analyze the multi-modal attention patterns in multi-subject personalized generation and observe two issues. (i) *Text-Reference Misalignment*: the text-token span of a subject and its corresponding reference condition attend to different image regions. (ii)

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.

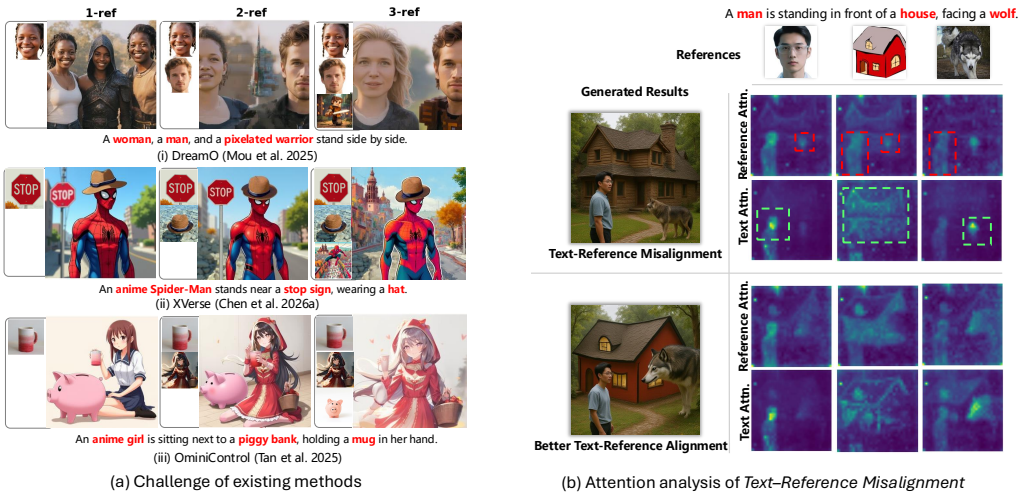


Figure 1: Challenge and analysis. (a) increasing the number of references tends to introduce copy-paste artifacts, subject confusion, and missing subjects in existing methods. (b) text–reference misalignment, where a subject's textual span (green box) and its reference image (red box) attend to different regions, underlies the challenges above; the bottom row shows the improved alignment between the reference and text attention maps.

Subject Confusion: the generated region of one subject may attend to multiple visually or semantically similar reference subjects. We attribute the former to the decoupling between text and reference images in multi-modal attention (Fig.1 (b)); when the model fails to establish a reliable binding relationship, the pre-trained semantic alignment can drift. This issue may be further exacerbated by collage-style training data, leading to prominent copy-paste artifacts (Fig. 1 (a)). The latter often causes appearance inconsistency and, in severe cases, leads to subject fusion and missing target subjects in the final generation.

To address these challenges, we propose **Bind-and-Route**, a framework that improves multi-subject personalized generation through **Explicit Reference Binding** and **Attention Routing**. Some results of our method are illustrated in Fig. 2. **Explicit Reference Binding** aims to establish explicit associations between each reference condition and its corresponding textual description to mitigate *Text–Reference Misalignment*. Inserting subject-specific placeholders into the prompt is a common practice for personalized generation. However, in multi-subject scenarios, learnable placeholders alone are insufficient to establish associations with their corresponding reference conditions. To address this issue, we explicitly inject association information into both placeholder and reference representations. Specifically, we use reference intermediate features as queries to resample the corresponding text features, yielding semantic-aware biases that strengthen the association between the text span and its matched reference condition. Similarly, we inject reference-aware cues into placeholder features to make the placeholders more tightly grounded to their corresponding references. In addition, we use contrastive objectives to encourage matched placeholder-reference pairs to be more correlated than mismatched pairs.

For the *Subject Confusion* issue, we introduce **Attention Routing**. Our key observation is that the text associated with each reference condition often provides more reliable localization cues for the target generation region in multi-modal attention, as shown in Fig.1 (b). We design a lightweight attention router that uses noise-text attentions as localization priors. The router guides noisy image tokens within each target subject region to attend to the correct reference condition, while penalizing attention to mismatched conditions. Specifically, it predicts attention-logit biases based on noise-to-text attentions, thereby reallocating attention over reference images. To avoid simply amplifying the overall conditioning strength, we further apply an offset mechanism to the routing bias, ensuring that the total attention mass from the noisy image to all condition branches remains unchanged.

Overall, our contributions are summarized as follows.

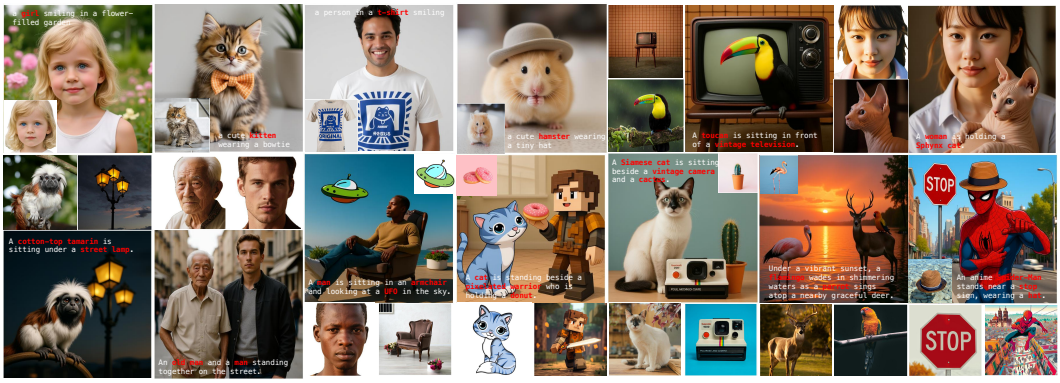


Figure 2: Bind-and-Route improves text alignment and appearance consistency with fewer perceptible copy-paste artifacts.

- We propose **Explicit Reference Binding** to build stable associations between corresponding text spans and reference conditions. It injects bidirectional associations between placeholder and reference representations via resamplers, along with contrastive objectives that distinguish matched placeholder-reference pairs from mismatched ones.
- We introduce **Attention Routing**, which uses noise-to-text attention as localization priors to guide noisy image tokens toward their corresponding reference conditions in the multi-modal attention. It reallocates noise-to-reference attention to emphasize target references without requiring extra masks at inference time.
- Extensive experiments demonstrate that **Bind-and-Route** improves both text alignment and reference appearance consistency in multi-subject personalized generation, validating the effectiveness of explicit binding and attention routing under the multi-modal attention paradigm.

2 Related Work

Personalized image generation. Personalized image generation aims to synthesize user-specified subjects in new contexts. Early diffusion-based methods typically rely on per-subject optimization or model fine-tuning Mokady et al. (2023); Ruiz et al. (2023). Another line explores zero-shot personalization, where reference images are encoded into conditioning embeddings and injected into the diffusion model without per-subject training or fine-tuning Wei et al. (2023); Shi et al. (2023); Ye et al. (2023); Wang et al. (2024b); Guo et al. (2024); Xiao et al. (2025). To improve controllability in multi-subject generation, many works impose explicit spatial or attention constraints, such as masks, layouts, or bounding boxes, to localize different subjects and reduce feature leakage Wang et al. (2025b); Zhou et al. (2024); Dahary et al. (2024). Routing-based methods further train reference encoders and routers to adaptively assign references to latent regions He et al. (2025a;b); Wang et al. (2024a). However, these methods typically rely on separately trained encoders or projectors to inject reference information into text embeddings or text-noise attention, making cross-modal alignment indirect and less intrinsic to the generation model.

Multi-modal personalization. Recent works formulate personalization as multi-modal in-context conditioning, where text, reference images, and target image tokens are jointly modeled by unified architectures such as DiTs Tan et al. (2025); Wang et al. (2025a); Wu et al. (2025c); Mou et al. (2025); Chen et al. (2026a); She et al. (2025). Another line leverages richer multi-modal conditions or video-generation priors to enhance controllability Chen et al. (2026b); Xu et al. (2025); Chen et al. (2025); Cao et al. (2026). However, existing multi-modal architectures often rely on implicit associations between textual subject spans and their corresponding references, which can lead to text-reference decoupling or subject confusion in multi-subject scenarios. In contrast, our method explicitly binds text and reference by using text-noise attention priors, which are naturally aligned with the reference modality, to let references compete for text-derived regions.

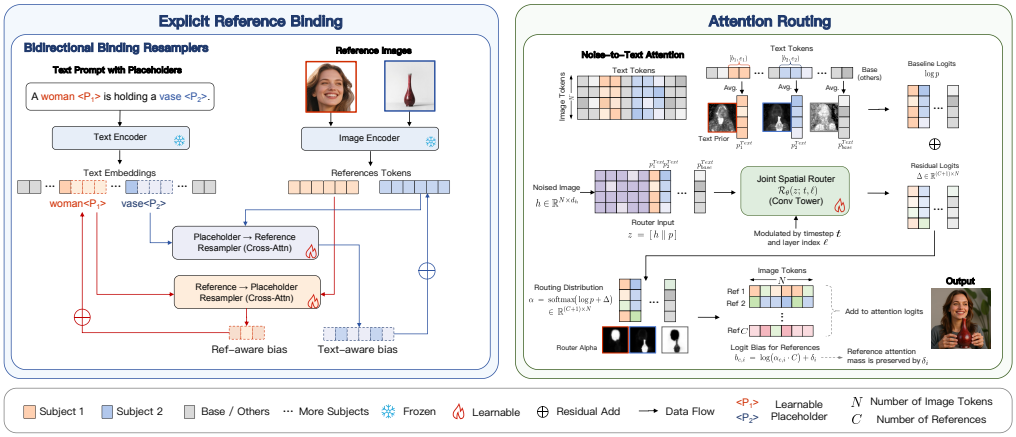


Figure 3: The pipeline of *Explicit Reference Binding* and *Attention Routing*. The bidirectional binding resamplers refine placeholder and reference embeddings before attention layers. For clarity, we illustrate the data flow of a single text-reference pair in each direction. The joint spatial router then controls attention allocation within the attention layers. The trained router can accurately extract target generation regions from text priors, as shown in the visualized router alphas. Detailed architectures of the resampler and router are shown in the Appendix.

3 METHOD

3.1 EXPLICIT REFERENCE BINDING

A common practice in personalized generation is to insert placeholders into the text prompt so that reference images can be introduced at their corresponding textual positions. However, in multi-subject personalized generation, we find that learnable placeholder embeddings alone are insufficient to reliably bind each placeholder to its corresponding reference image, especially when the model does not employ a CLIP-like vision-language aligned encoder. We therefore propose an explicit bidirectional binding mechanism: reference information is injected into the placeholder representations, while condition-associated textual information is injected into the reference intermediate features through resamplers. Compared with directly learning standalone placeholders, this condition-aware resampling provides stronger reference-text associations. We further introduce contrastive objectives that pull matched placeholder-reference pairs together and penalize mismatched pairs in multi-subject generation. The pipeline is shown in Fig.3.

Bidirectional binding resamplers. The multi-modal attention naturally provides a shared space where different conditioning modalities can interact with each other. However, simply inserting learnable placeholders into the text prompt, or adding learnable idx embeddings Mou et al. (2025) to reference condition tokens, does not guarantee a stable binding between them. To address this issue, we introduce bidirectional binding resamplers that inject target-associated information into both placeholder and reference representations before they enter multi-modal attention. As shown in Fig. 3 (red lines), in the reference-to-placeholder direction, we use the learnable placeholder embedding as a query to attend to its corresponding reference condition features. The resampled output serves as a reference-aware residual bias, which is added to the placeholder representation, allowing the placeholder to carry visual cues from its matched reference. In the text-to-reference direction (blue lines in Fig. 3), we use the reference condition representation as a query to attend to the associated text embedding slice. The resulting text-aware residual bias is injected into the reference representation, enabling the reference tokens to encode their semantic role in the prompt. This design preserves the original textual or visual semantics while introducing explicit association cues. As a result, each placeholder and its corresponding reference condition are equipped with paired binding biases, making them more likely to attend to each other in subsequent multi-modal attention layers.

Bidirectional contrastive objectives. In addition to bidirectional resamplers, we further introduce contrastive objectives tailored for multi-subject personalized generation. For the placeholder-to-reference direction, we define a *Placeholder-to-Reference Contrastive Loss*, denoted as $\mathcal{L}_{\mathrm{P2R}}$. Given the placeholder span of subject i, denoted as T_i^p, and the reference condition tokens of subject j, denoted as C_j, we compute their attention score by averaging the attention mass from T_i^p to C_j:

$$s(T_i^p, C_j) = \frac{1}{|T_i^p||C_j|} \sum_{t \in T_i^p} \sum_{c \in C_j} A_{t,c}^{\mathrm{P2R}} \tag{1}$$

Here, $A_{t,c}^{\mathrm{P2R}}$ denotes the attention probability from placeholder token t to reference token c. For each placeholder T_i^p, all reference conditions are treated as candidates, and the matched reference C_i is used as the positive target. The contrastive loss is formulated as:

$$\mathcal{L}_{\mathrm{P2R}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{s(T_i^p, C_i)}{\sum_{j=1}^{K} s(T_i^p, C_j)} \tag{2}$$

where K denotes the number of reference condition candidates for each placeholder, and N denotes the number of placeholders.

Similarly, for the reference-to-placeholder direction, we define a *Reference-to-Placeholder Contrastive Loss*, denoted as $\mathcal{L}_{\mathrm{R2P}}$. Given the reference condition tokens of subject i, denoted as C_i, and the placeholder span of subject j, denoted as T_j^p, we compute their attention score by averaging the attention mass from C_i to T_j^p:

$$s(C_i, T_j^p) = \frac{1}{|C_i||T_j^p|} \sum_{c \in C_i} \sum_{t \in T_j^p} A_{c,t}^{\mathrm{R2P}} \tag{3}$$

$A_{c,t}^{\mathrm{R2P}}$ denotes the attention probability from reference token c to placeholder token t. For each reference condition C_i, all placeholders are treated as candidates, and the matched placeholder T_i^p is used as the positive target. The contrastive loss is formulated as:

$$\mathcal{L}_{\mathrm{R2P}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{s(C_i, T_i^p)}{\sum_{j=1}^{K} s(C_i, T_j^p)} \tag{4}$$

where K denotes the number of placeholder candidates for each reference condition, and N denotes the number of reference conditions. The overall bidirectional contrastive objective is defined as:

$$\mathcal{L}_{bind} = \mathcal{L}_{\mathrm{P2R}} + \mathcal{L}_{\mathrm{R2P}} \tag{5}$$

Together with the bidirectional resamplers, $\mathcal{L}_{bind}$ explicitly strengthens placeholder-reference correspondence, leading to more reliable reference binding in multi-subject personalized generation.

3.2 Attention Routing

In the multi-modal attention, modalities jointly compete for attention, and as the number of reference conditions increases, the attention allocated to the target reference declines. We observe that the noise-to-text attention map corresponding to each subject's describing text tokens, which we term the *text prior*, provides a reliable indication of that subject's generation region. Using this cue to guide reference attention requires no external masks at inference time; moreover, since the cue is read from within the model itself, it naturally preserves the text–image alignment of the pretrained model.

Joint spatial router. We design a convolutional *Joint Spatial Router* that takes the noised image query together with the text prior of every reference as input, and predicts per-token attention biases that redistribute attention across references. The pipeline is shown in Fig. 3. Concretely, for the c-th

reference whose describing text tokens occupy span $[b_c, e_c)$, we define its *text prior* as the averaged noise-to-text attention probability,

$$p_c^{\text{text}} \;=\; \frac{1}{e_c - b_c} \sum_{i=b_c}^{e_c-1} A_{:,\,i}^{\text{noise}\to\text{text}} \tag{6}$$

where $A^{\text{noise}\to\text{text}} \in \mathbb{R}^{N\times L}$ is the head-averaged noise-to-text attention map, N is the number of noised image tokens and L is the number of text tokens. Analogously, we define a *base text prior* $p_{\text{base}}^{\text{text}} \in \mathbb{R}^N$ from the text tokens not covered by any reference span, which represents background regions.

Let h denote the projected query feature of the noised image at the current attention layer. We concatenate h with the stacked text priors $p = [p_1^{\text{text}}, \ldots, p_C^{\text{text}}, p_{\text{base}}^{\text{text}}]$ along the channel dimension to form the router input,

$$z \;=\; [\, h \,\|\, p \,] \tag{7}$$

The features z are then reshaped to the 2D latent grid and processed by a lightweight convolutional tower , which yields per-token residual logits

$$\Delta \;=\; \mathcal{R}_\theta(z;\, t, \ell) \tag{8}$$

where $\mathcal{R}_\theta$ denotes the router, and t and ℓ denote the denoising timestep and the attention-layer index used to modulate the tower, respectively. The final per-token routing distribution combines the residual logits with the log-prior and normalizes across branches,

$$\alpha \;=\; \mathrm{softmax}\big(\log p + \Delta\big) \tag{9}$$

We apply the routing distribution α to the multi-modal attention as a per-token, per-reference logit bias rather than a hard gate, so that the model retains the full unified attention computation and only reshapes how attention is allocated among reference branches. For reference c at token i, we add the bias

$$b_{c,i} \;=\; \log\big(\alpha_{c,i} \cdot C\big) \tag{10}$$

to every attention logit whose key belongs to reference c. A positive bias ($\alpha_{c,i} > 1/C$) strengthens attention toward reference c in its target region, while a negative bias suppresses the non-target references elsewhere.

Directly adding $\{b_{c,i}\}$ can, however, shift the total attention mass on reference branches relative to the text and main-image branches, which risks destabilizing text alignment. To prevent this, we restore the aggregate reference mass with a *per-token common offset* applied uniformly to all reference logits. Let

$$M_i^{\text{before}} \;=\; \sum_{m\in\mathcal{K}_{\text{ref}}} A_{i,m}^{\text{before}}, \quad M_i^{\text{mid}} \;=\; \sum_{m\in\mathcal{K}_{\text{ref}}} A_{i,m}^{\text{mid}} \tag{11}$$

denote the total softmax probability on all reference keys $\mathcal{K}_{\text{ref}}$ before and after adding $\{b_{c,i}\}$. We compute the common offset

$$\begin{aligned} \delta_i &= \mathrm{logit}\big(M_i^{\text{before}}\big) - \mathrm{logit}\big(M_i^{\text{mid}}\big), \\ \text{where} \quad \mathrm{logit}(p) &= \log\big(p/(1-p)\big) \end{aligned} \tag{12}$$

and form the final bias added to every reference attention logit as

$$\tilde{b}_{c,i} \;=\; b_{c,i} + \delta_i \tag{13}$$

Finally, $\tilde{b}_{c,i}$ redistributes attention among references while preserving the overall balance of the pretrained multi-modal attention.

Mask supervision objectives. We supervise the router using ground-truth foreground masks $m_c \in \{0,1\}^N$. Tokens are classified into three categories: foreground ($\max_c m_{c,i} \geq \tau_{\text{fg}}$), background ($\sum_c m_{c,i} \leq \tau_{\text{bg}}$), and ambiguous (including overlap regions). Ambiguous tokens are excluded from loss computation. We define three complementary losses.

Cross-entropy loss. Treating the router allocation as token-level classification with target y_i, which is the dominant condition for foreground tokens, or the base branch for background tokens, we have

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\mathcal{V}|} \sum_{i\in\mathcal{V}} \log \alpha_{y_i,\,i} \tag{14}$$



Figure 4: Qualitative comparison between our Bind-and-Route and other methods.

where $\mathcal{V}$ is the set of non-ambiguous tokens.

Negative loss. Directly penalizing the residual allocation of the target condition on foreground tokens $\mathcal{F}$:

$$\mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \left(1 - \alpha_{y_i,\, i}\right) \tag{15}$$

Margin loss. Enforcing a margin m between the target and the strongest competitor:

$$\mathcal{L}_{\text{margin}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \max\Big(0,\ \max_{c \neq y_i} \alpha_{c,\, i} - \alpha_{y_i,\, i} + m\Big) \tag{16}$$

The total router loss is

$$\mathcal{L}_{\text{router}} = \lambda_{\text{ce}}\, \mathcal{L}_{\text{ce}} + \lambda_{\text{neg}}\, \mathcal{L}_{\text{neg}} + \lambda_{\text{margin}}\, \mathcal{L}_{\text{margin}} \tag{17}$$

4 Experiment

4.1 Implementation Details

We use Flux.1-dev as our pretrained model and train our method based on the OminiControl Tan et al. (2025) baseline. We train LoRA, bidirectional binding resamplers, joint spatial router, and 4-token placeholders, for a total of 285M trainable parameters. At inference time, generating one image with 28 steps on an NVIDIA H800 takes about 10 seconds. Additional details including hyperparameters and datasets are provided in the Appendix.

4.2 Qualitative Comparison

We compare our method with state-of-the-art approaches, including OminiControl Tan et al. (2025), UNO Wu et al. (2025c), DreamO Mou et al. (2025), MOSAIC She et al. (2025), XVerse Chen et al. (2026a), MS-Diffusion Wang et al. (2025b), OmniGen2 Wu et al. (2025b), and Qwen-Image-Edit Wu et al. (2025a). Among them, MS-Diffusion is built on SDXL, while the rest adopt DiT



Figure 5: When using the same references under different prompts, our method better follows the text and generates consistent appearance with minimal copy-paste artifacts.

Table 1: PSRBench results under 3- and 4-reference settings. Bold and underline denote the best and second-best results.

Method	3-Reference				4-Reference			
	Subject	Semantic	Aesthetic	Average	Subject	Semantic	Aesthetic	Average
OminiControl	0.5190	0.8638	<u>0.9469</u>	<u>0.7766</u>	0.4026	<u>0.8543</u>	0.9462	0.7344
XVerse	0.5117	0.8109	0.8643	0.7290	0.4468	0.7778	0.8344	0.6863
OmniGen2	0.5133	0.8553	0.7983	0.7223	**0.4779**	0.7739	0.8228	0.6915
Qwen-Image-Edit	0.5209	**0.8936**	0.7875	0.7340	0.4461	0.7326	0.6259	0.6015
UNO	0.5055	0.8426	0.9322	0.7601	<u>0.4740</u>	0.7783	<u>0.9534</u>	<u>0.7352</u>
DreamO	**0.5280**	0.8553	0.9363	0.7732	0.4513	0.8109	0.9327	0.7316
MOSAIC	0.4568	<u>0.8830</u>	0.9005	0.7468	0.3710	0.8413	0.8765	0.6963
MS-Diffusion	0.4437	0.7745	0.7790	0.6657	0.4375	0.7261	0.7835	0.6490
Ours	<u>0.5237</u>	0.8587	**0.9621**	**0.7815**	0.4692	**0.8630**	**0.9922**	**0.7748**

backbones. Except for Qwen-Image-Edit and OmniGen2, all methods use Flux.1 as the pretrained model. For inference, MS-Diffusion (SDXL-based) uses 50 denoising steps, while the other methods use 28 steps. Qualitative comparisons are shown in Fig. 4.

With *Explicit Reference Binding*, our personalization training preserves the pretrained text-image alignment as the number of reference images increases. In contrast, when conditioning on multiple references, some methods (e.g., DreamO and XVerse) exhibit misalignment, such as missing subjects mentioned in the prompt or incorrect interactions between subjects. In some cases, XVerse and Qwen-Image-Edit directly copy and paste reference images into the generated results. We further illustrate these issues in Fig. 5 by using the same set of references with different prompts.

As the number of references grows, prior methods often suffer from degraded appearance consistency in generated images. In severe cases, some methods show obvious fusion of subject attributes or missing target subjects. For example, when generating three human subjects (row 4) or simultaneously generating camera and computer (row 5), similar categories exacerbate the problem. XVerse maintains relatively good appearance consistency, but heavy copy–paste artifacts reduce overall visual quality. Benefiting from *Attention Routing*, which encourages text-related regions to attend to the correct reference, our method effectively mitigates subject confusion and subject missing.

4.3 Quantitative Comparison

We further conduct quantitative evaluations on PSRBench Wang et al. (2026). PSRBench consists of 50 cases each for the 3- and 4-reference settings. We report four metrics: subject consistency (*Subject*), semantic alignment (*Semantic*), aesthetic preference (*Aesthetic*), and the average of the three metrics (*Average*). Higher scores indicate better performance for all metrics. The results are reported in Tab. 1.

Our method delivers strong results across all metrics, especially in *Aesthetic*, scored by a notable human-preference reward model HPSv3 Ma et al. (2025), suggesting better perceived visual quality with fewer perceptible copy-paste artifacts. With competitive *Subject* and *Semantic* scores, it further achieves the best *Average* among all methods. Notably, OminiControl, which is the training

Table 2: Ablation study under 3- and 4-reference settings. Each row cumulatively adds one component to the previous row; bold marks our full method.

Method	3-Reference				4-Reference			
	Subject	Semantic	Aesthetic	Average	Subject	Semantic	Aesthetic	Average
baseline	0.5190	0.8638	0.9469	0.7766	0.4026	0.8543	0.9462	0.7344
+ ph/idx embedding	0.5202	0.8649	0.9431	0.7761	0.4120	0.8611	0.9369	0.7367
+ contrastive loss	0.5199	0.8654	0.9533	0.7795	0.4174	0.8625	0.9558	0.7452
+ resampler	0.5214	**0.8657**	0.9550	0.7807	0.4378	0.8624	0.9634	0.7545
+ router (**Ours**)	**0.5237**	0.8587	**0.9621**	**0.7815**	**0.4692**	**0.8630**	**0.9922**	**0.7748**



Figure 6: Qualitative examples of the ablation study. Our method improves both text-image alignment and appearance consistency.

codebase we build upon, uses the same training dataset as ours, ruling out data differences as a confounding factor. For a more comprehensive human-centric comparison, we conducted a user study, with detailed results in the Appendix.

4.4 Ablation Study

We conduct ablation studies on 3- and 4-reference subsets, as shown in Tab. 2. Our baseline is OminiControl. We first introduce placeholder embeddings in the text embeddings and add idx embeddings to the reference embeddings (+ *ph/idx embedding*). Our placeholder embedding is added directly to the encoded prompt embeddings (rather than inserted into the text), so it does not affect contextual semantics during encoding. Results show that this improves image–text alignment, yielding a better *Semantic* score.

Next, we introduce *Explicit Reference Binding*. Bidirectional contrastive objectives (+ *contrastive loss*) can be used with the *ph/idx embeddings* to strengthen text–reference binding and improve appearance consistency. Bidirectional binding resamplers (+ *resampler*) further inject associated reference and textual information into the placeholder/reference embeddings, making the binding easier to establish.

Finally, we add *Attention Routing* (*router*). Based on the localization prior from the pretrained image–text alignment, the joint spatial router reallocates attention to the reference conditions, bringing a significant improvement in the *Subject* score. Qualitative examples are shown in Fig. 6.

5 Conclusion and Discussion

We present **Bind-and-Route** for multi-subject personalized generation. It uses *Bidirectional Binding Resamplers* with contrastive objectives to explicitly associate textual descriptions with reference conditions, and introduces a lightweight *Joint Spatial Router* to guide noisy image tokens toward the correct references. By mitigating *Text-Reference Misalignment* and *Subject Confusion*, our method improves both text alignment and appearance consistency in multi-subject personalized generation.

Our method still has limitations, such as tonal shifts, which we mainly attribute to the training data. Our multi-subject data are entirely derived from the open-source Echo-4o dataset Ye et al. (2025), which is generated by Nano-banana and thus exhibits a style gap from real images.

In future work, we plan to address the shortage of high-quality multi-subject personalized generation data, establish more diverse evaluation metrics, and explore tighter text-reference binding on newer model architectures for improved multi-subject personalization.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure: generating synthetic datasets (all of which are derived from open-source datasets), helping develop theoretical models and conceptual frameworks, assisting in the writing of proofs, designing and providing feedback on research methodology and experiments, implementing methods, assisting with translation, and supporting qualitative and thematic data analysis. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, cleaning and reformatting datasets, or interpreting results; these tasks are not applicable to this work. Additionally, we used generative AI tools for the following tasks with recommended disclosure: creating or modifying scientific figures and images, creating or editing software code, summarizing and analyzing existing literature, brainstorming, sourcing and searching for information, editing the paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. Specifically, AI-generated or -modified figures were manually checked by the authors for correctness and clarity; LLM-suggested code and text edits were verified and tested for correctness by the authors; and AI-assisted literature summaries were cross-checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Hengyuan Cao, Yutong Feng, Biao Gong, Yijing Tian, Yunhong Lu, Chuang Liu, and Bin Wang. Dimension-reduction attack! video generative models are experts on controllable image synthesis. *Advances in Neural Information Processing Systems*, 38:81329–81358, 2026.

Bowen Chen, Haomiao Sun, Li Chen, Xu Wang, Daniel Du, Xinglong Wu, et al. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *Advances in Neural Information Processing Systems*, 38:7437–7459, 2026a.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025.

Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3174–3182, 2026b.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 432–448, 2024. doi: 10.1007/978-3-031-72630-9_25.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025.

Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. 37:36777–36804, 2024.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14399–14408, 2025a.

Junjie He, Yuxiang Tuo, Binghui Chen, Chongyang Zhong, Yifeng Geng, and Liefeng Bo. Anystory: Towards unified single and multiple subject personalization in text-to-image generation. *arXiv preprint arXiv:2501.09503*, 2025b.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18325–18334, 2025a.

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2026.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025b.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15943–15953, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025c.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025. doi: 10.1007/s11263-024-02227-z.

Ruihang Xu, Dewei Zhou, Fan Ma, and Yi Yang. Contextgen: Contextual layout anchoring for identity-consistent multi-instance generation. *arXiv preprint arXiv:2510.11000*, 2025.

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

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024.

A Implementation Details

Hyperparameters. The bidirectional binding resamplers are applied to the reference condition and text hidden states before the first block at each timestep. Placeholders are added after text encoding and thus do not disturb contextual semantics, as reflected by the higher *Semantic* scores in Tab. 1. The joint spatial router uses block range $[11, 40] \in [1, 57]$ and timestep range $[0.3, 0.7] \in [0, 1]$, chosen based on text-prior localization ability and training performance. α and the mask objectives are computed layer-wise and averaged. In Eq. 17, we use $\lambda_{\text{ce}} = 0.5$, $\lambda_{\text{neg}} = 0.1$, and $\lambda_{\text{margin}} = 0.1$ with margin $m = 0.1$. The text prior is detached, so these losses only affect the router. Detailed architectures of the resampler and router are shown in Fig. 7. We set LoRA rank to 16, AdamW optimizer, and batch size to 1. Training is staged, with 100k total steps. Inference is mask-free. Our method supports any number of references in the training dataset while preserving the pretrained model's standard generation ability.

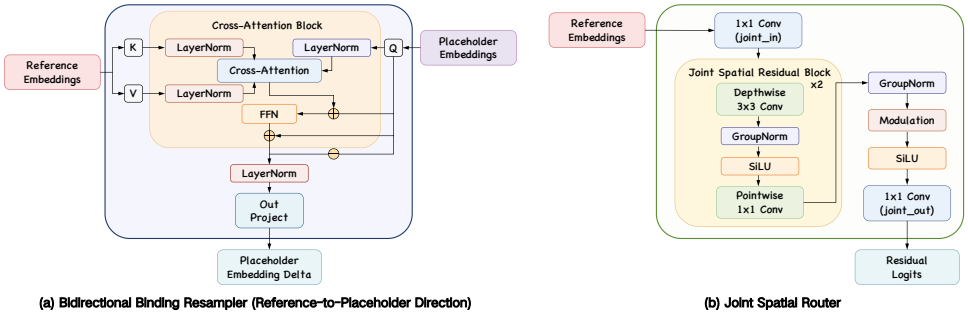


Figure 7: Detailed architectures of the bidirectional binding resampler and joint spatial router. For the bidirectional binding resampler, we use the reference-to-placeholder direction as an example; the placeholder-to-reference direction is the same, except with the input and output reversed.

Datasets. All data comes from open-source datasets. We use SemAlign-MS-Subjects200K She et al. (2025) and Nano-consistent-150k as single-subject datasets, and a 15k subset from Echo-4o Ye et al. (2025) for multi-subject training. We clean the data and re-annotate prompts, subject labels, and masks. Masks are extracted using GroundingDINO and matched to references using the Hungarian algorithm.

B User Study

We conducted a user study to compare human-perceived quality across methods in terms of subject consistency (*Subject*), semantic alignment (*Semantic*), aesthetic preference (*Aesthetic*), and copy-paste artifacts (*Copy-Paste*), for which no suitable automatic metric is available. For each criterion, participants rank the results of different methods on five test cases. Methods ranked 1st–2nd, 3rd–5th, and 6th–9th receive 3, 2, and 1 point, respectively. Scores are averaged over all test cases for each method and criterion. We collected 10 questionnaires, with results reported in Tab. 3.

Table 3: User study results across different criteria.

Method	Subject	Semantic	Aesthetic	Copy-Paste
OminiControl	2.0	1.8	1.4	1.7
UNO	1.4	1.4	2.2	1.6
DreamO	1.8	1.5	1.3	1.8
MOSAIC	1.2	2.2	1.8	2.0
XVerse	2.0	1.9	2.0	1.7
MS-Diffusion	1.2	1.4	1.2	1.4
OmniGen2	1.9	1.6	2.1	1.6
Qwen-Image-Edit	2.1	1.7	1.8	1.6
Ours	**2.4**	**2.5**	**2.2**	**2.6**

Table 4: XVerseBench results under 1-, 2-, and 3-reference settings.

	1-reference					2-reference					3-reference				
Method	DPG	AES	IP	ID	AVG	DPG	AES	IP	ID	AVG	DPG	AES	IP	ID	AVG
OminiControl	91.35	57.05	69.70	40.99	64.77	89.32	53.74	63.06	27.36	58.37	84.85	53.80	57.19	24.86	55.18
UNO	87.41	56.93	**72.96**	60.20	69.38	89.34	54.53	62.81	39.92	61.65	84.23	54.10	57.67	34.81	57.70
DreamO	94.09	55.61	70.17	76.97	74.21	91.09	52.08	**66.89**	50.30	65.09	80.62	52.00	56.57	38.01	56.80
MOSAIC	95.53	**58.05**	67.04	59.94	70.14	91.65	**55.54**	61.04	42.67	62.73	88.37	**55.49**	55.30	31.55	57.68
XVerse	93.26	57.08	67.53	**79.39**	**74.32**	89.49	53.97	66.15	**67.59**	**69.30**	**92.74**	51.80	**62.96**	**42.59**	**62.52**
MS-Diffusion	85.89	52.43	60.56	26.39	56.32	80.25	51.51	54.49	19.50	51.44	71.80	52.17	48.77	16.23	47.24
OmniGen2	93.84	53.76	70.08	51.59	67.32	92.16	52.95	63.50	47.97	64.15	87.07	52.38	57.61	34.58	57.91
Qwen-Image-Edit	**97.21**	53.74	67.59	51.74	67.57	**92.74**	51.80	62.96	42.59	62.52	87.72	53.22	58.13	35.56	58.65
Ours	91.81	57.72	70.47	57.51	69.38	90.05	54.82	61.94	40.60	61.85	88.03	54.54	57.89	36.01	59.12

As shown in Tab. 3, our method achieves the highest scores across all four criteria, demonstrating consistently superior perceptual quality over the baselines. Its pronounced advantage in *Copy-Paste* further indicates more effective suppression of perceptible copy-paste artifacts without compromising subject consistency or semantic alignment.

C METRICS

We evaluate our method on PSRBench Wang et al. (2026), a recent high-quality multi-subject personalization benchmark supporting up to four references. We also report results on XVerseBench Chen et al. (2026a), another widely used benchmark comprising 90 one-reference, 120 two-reference, and 90 three-reference cases. All metrics are higher-is-better: DPG measures text-image alignment, AES evaluates aesthetic quality, DINO similarity assesses object appearance consistency (IP), ArcFace similarity measures identity consistency (ID), and AVG denotes the overall average. Results are reported in Tab. 4.

Our method achieves high DPG and AES scores, while remaining robust as the number of references increases. On the 3-reference subset, it ranks among the top three across all metrics. XVerse, which introduced XVerseBench, achieves the highest ID and IP similarity scores. However, qualitative results in Fig. 4 and Fig. 5 reveal severe copy–paste artifacts, consistent with its lowest AES score among all methods on the 3-reference subset. Qwen-Image-Edit benefits from VLM-based modeling and therefore exhibits strong text–image alignment, as reflected by its DPG scores, but its ID/IP consistency remains below that of specialized Flux.1-based methods. Compared with OminiControl, which uses the same training data, our method delivers substantial improvements in DPG and ID/IP similarity, validating the effectiveness of the proposed approach.

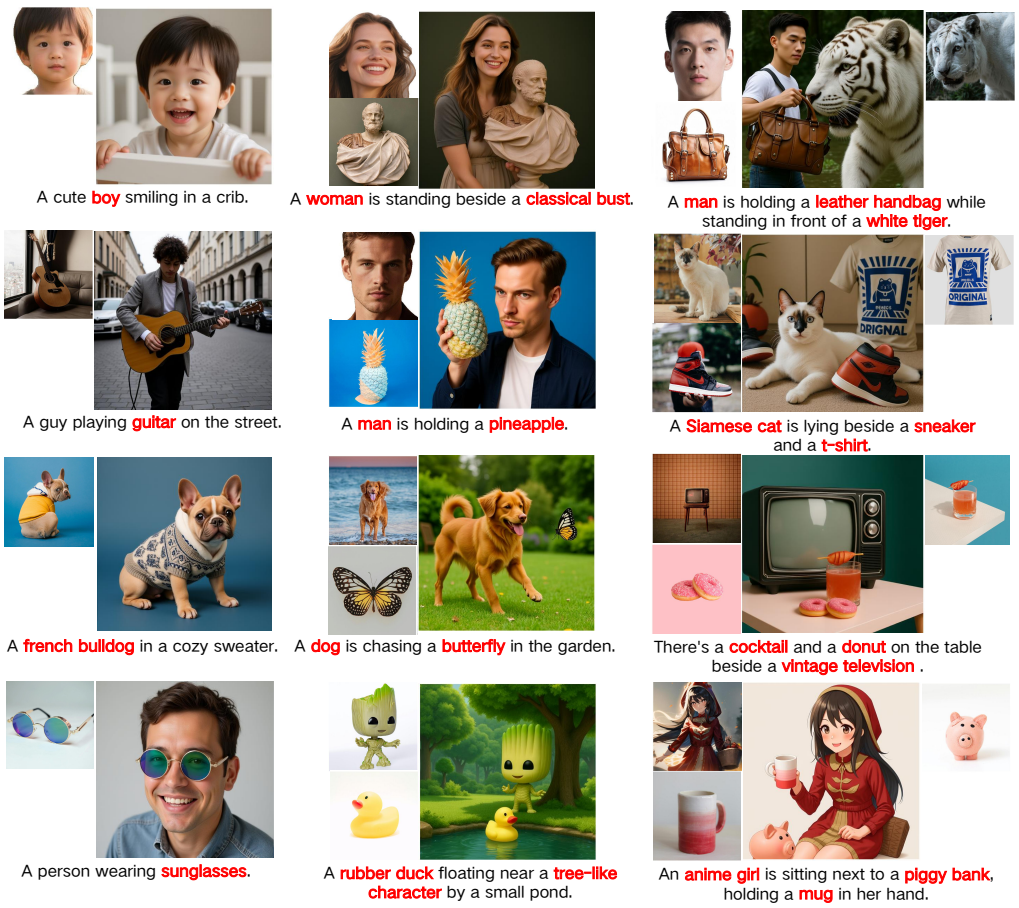


Figure 8: Results of our method.