RESOLVING SPATIAL AND SEMANTIC CONFLICTS IN MULTI-LORA COMPOSITION VIA SOFTMAX ROUTING AND CROSS-ADAPTER ATTENTION CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-LoRA composition activates several Low-Rank Adaptation (LoRA) modules simultaneously so that a single diffusion forward pass generates an image that embodies multiple personalized concepts at once. Existing methods either merge adapter weight deltas uniformly—corrupting concepts that share spatial territory—or route noise predictions via hard per-pixel argmax, which discards non-dominant LoRA contributions and collapses identity consistency on overlapping pairs. We propose two complementary, training-free mechanisms. **SR-LoRA** (Spatially-Routed LoRA) replaces hard argmax routing with temperature-controlled softmax masks over per-LoRA cross-attention maps, smoothly blending noise predictions at each spatial position. **CLAC** (Cross-LoRA Attention Conditioning) runs a two-pass UNet forward where the anchor LoRA's cross-attention keys and values are blended into the secondary LoRA's attention, suppressing appearance leakage; an IoU-based overlap detector activates this conditioning adaptively across all pair types. On the ComposLoRA benchmark (Zhong et al., 2024) (240 compositions per domain, 10 seeds), SR-LoRA and FreeFuse reach equivalent CLIP-T (+4.3% over LoRA-C), while CLAC improves CLIP-I by +**7.6pp** over FreeFuse on reality (0.915 vs. 0.839) and +**0.9pp** on anime. The combination (Adaptive Fusion) achieves the highest CLIP-I on anime (0.875), validated with per-pair-type breakdowns, ablations, and N=3 triplet experiments.

1 INTRODUCTION

Personalized text-to-image generation has been transformed by Low-Rank Adaptation (LoRA) Hu et al. (2021), which encodes a new visual concept—a specific character, garment, art style, or object—in a small set of rank-r weight matrices trained on a handful of reference images. A natural extension is to activate several LoRAs simultaneously, composing multiple concepts in a single image without any additional training. Yet the canonical approach, LoRA-Composite (LoRA-C) Zhong et al. (2024), simply sums the weight deltas of all active adapters, providing no mechanism to arbitrate between concepts that attend to the same image regions. The result is a blurred average where neither concept is faithfully rendered when, for example, a character LoRA and a clothing LoRA both fire over the foreground.

FreeFuse Jian et al. (2025) addresses this by routing noise predictions through a hard per-pixel argmax over cross-attention maps: each pixel is assigned exclusively to the LoRA with the highest attention weight. This sharpens spatial boundaries but introduces two new failure modes. First, the hard binary assignment discards contributions from non-dominant LoRAs entirely, producing discontinuous seams and erasing fine details at routing boundaries. Second, and more subtly, uniform routing is applied to *every* concept pair—even spatially disjoint ones (character + background) where no routing is needed—consistently degrading CLIP-I identity scores on those pairs.

We observe that these two failure modes are orthogonal, and propose one targeted mechanism for each. **SR-LoRA** (Spatially-Routed LoRA) replaces hard argmax routing with a softmax over per-LoRA cross-attention maps controlled by a temperature parameter τ. At τ=0.1, each pixel receives a smooth, differentiable blend of all LoRA noise predictions weighted by relative attention, eliminating seam artifacts and consistently outperforming uniform merging ($\tau \to \infty$). **CLAC** (Cross-LoRA

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.

Attention Conditioning) targets the second failure mode: an IoU-based overlap detector identifies concept pairs with significant spatial co-occurrence; for those pairs, a first UNet pass captures the anchor LoRA's cross-attention keys and values, which are then linearly blended (α=0.5) into the secondary LoRA's attention in a second pass. This makes the secondary LoRA's spatial features identity-aware without discarding its own conditioning signal.

Our contributions are:

1. **SR-LoRA**: temperature-controlled softmax noise routing that improves CLIP-T by +4.3% over LoRA-C and matches FreeFuse on CLIP-T (<0.1% gap) while smoothing hard-boundary seam artifacts.

2. **CLAC**: two-pass cross-adapter attention conditioning that improves CLIP-I by +7.6pp over FreeFuse on reality (0.915 vs. 0.839) and +0.9pp on anime, with adaptive activation via an annotation-free IoU detector validated stable over $\delta \in [0.12, 0.18]$.

3. Adaptive Fusion combining both mechanisms achieves the highest CLIP-I on anime (0.875), validated by per-pair-type breakdowns, ablations over τ, α, δ, routing level, and N=3 triplet composition.

2 Related Work

Personalized text-to-image generation. Textual Inversion Gal et al. (2022) encodes a concept as an optimized token; DreamBooth Ruiz et al. (2023) fine-tunes the full UNet per concept. Custom Diffusion Kumari et al. (2023) extends this to multi-concept customization with cross-attention regularization, but still requires joint training for each combination. LoRA Hu et al. (2021) injects low-rank updates into cross-attention projections, yielding compact adapters (≈10 MB) that have become the de-facto format for community-shared concepts. IP-Adapter Ye et al. (2023) decouples image-prompt identity conditioning into a parallel cross-attention branch, providing complementary appearance control to text-driven LoRA adapters.

Multi-LoRA composition. Zhong et al. Zhong et al. (2024) introduce the ComposLoRA benchmark alongside two decoding-time strategies: **LoRA-Composite** (LoRA-C), which sums adapter weight deltas and runs a single forward pass, and **LoRA-Switch** (LoRA-S), which alternates between active adapters across denoising timesteps. Both provide our primary baselines; we target SR-LoRA as a spatial alternative to temporal alternation. LoRA-Composer Yang et al. (2025) improves spatial precision via user-supplied bounding boxes, requiring layout annotations. ZipLoRA Shah et al. (2023) merges subject and style LoRAs via SVD-based subspace alignment; Mix-of-Show Gu et al. (2023) learns a regionally conditioned LoRA but requires per-combination training. B-LoRA Frenkel et al. (2024) disentangles content and style within a single LoRA; Concept Weaver Kwon et al. (2024) fuses multi-concept representations at test time but needs per-concept optimization steps. CLoRA Meral et al. (2024) applies contrastive learning to compositional LoRA inference without test-time fine-tuning. FreeFuse Jian et al. (2025) is closest to our work: it runs per-LoRA forward passes and assigns each pixel to the highest-attention LoRA via hard argmax. We identify two systematic failure modes in FreeFuse (Section 3) and address each with SR-LoRA and CLAC.

Cross-attention manipulation. Attend-and-Excite Chefer et al. (2023) steers single-adapter generation by minimizing cross-attention entropy. MasaCtrl Cao et al. (2023) and Prompt-to-Prompt Hertz et al. (2022) transfer cross-attention K/V representations across images to achieve appearance consistency. StyleAligned Hertz et al. (2023) shares self-attention across a batch to enforce consistent style, demonstrating the broad utility of KV-sharing in diffusion forward passes. CLAC inherits this KV-transfer insight but applies it *across LoRA adapters*, gated on a spatial IoU criterion rather than applied unconditionally. MultiDiffusion Bar-Tal et al. (2023) fuses spatially decomposed denoising trajectories but requires user-specified region masks.

3 METHOD

3.1 BACKGROUND: LATENT DIFFUSION AND LORA

Stable Diffusion Rombach et al. (2022) trains a denoising UNet $\epsilon_\theta(z_t, t, c)$ to predict noise from latent z_t conditioned on text embedding c. Cross-attention layers attend from spatial features (queries Q) to projected text embeddings (keys K, values V):

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big) V. \tag{1}$$

LoRA Hu et al. (2021) injects trainable rank-r updates $\Delta W = BA$ ($B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, $r \ll k$) into the cross-attention projection matrices. LoRA-C Zhong et al. (2024) (our reference baseline) merges N adapters as $W = W_0 + \sum_{i=1}^{N} \Delta W_i$ before a single forward pass, providing no spatial or semantic discrimination between concepts.

3.2 STEP 1: PER-LORA ATTENTION MAP EXTRACTION

Each LoRA is associated with *trigger words* $\mathcal{T}_i$—short token sequences unique to that adapter that activate its concept during inference (e.g., `iu1` for a character LoRA; `school uniform` for a clothing LoRA; `masterpiece` for a style LoRA). We run a single-adapter forward pass for each LoRA and aggregate cross-attention weights at the trigger-word token positions across the semantic UNet layers (mid block and `up_blocks.1`), spatially pooled to a common resolution:

$$A_i(x, y) = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} \Big[\mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)\Big]_{(x,y),\,j}, \tag{2}$$

where $[\cdot]_{(x,y),j}$ denotes the scalar attention weight from spatial position (x, y) to trigger token j, averaged across semantic UNet layers. The resulting map $A_i \in \mathbb{R}^{H\times W}$ encodes where in the image LoRA i places its concept, providing the spatial signal for both SR-LoRA routing and the IoU-based overlap detector.

3.3 STEP 2: IOU-BASED SPATIAL OVERLAP DETECTION

To decide whether two concepts require identity conditioning (CLAC), we measure how much their cross-attention maps co-activate the same spatial regions. We binarize each map at its mean: $\hat{A}_i = \mathbf{1}[A_i > \mu_i]$, and compute:

$$\mathrm{IoU}(i, j) = \frac{|\hat{A}_i \cap \hat{A}_j|}{|\hat{A}_i \cup \hat{A}_j|}. \tag{3}$$

A pair is classified as *spatially overlapping* if $\mathrm{IoU}(i, j) > \delta$, and *non-overlapping* otherwise. The intuition is physical: character and clothing LoRAs both attend to the body region (high IoU ≈ 0.40–0.55), so their noise predictions compete at shared pixels and leakage is likely. Style and background LoRAs produce globally diffuse attention that overlaps with character attention (mean IoU ≈ 0.21–0.24, above δ), so CLAC activates for those pairs as well. By contrast, object LoRAs are spatially focused away from the character region (mean IoU ≈ 0.09), and appearance leakage is not a risk.

We set δ=0.15, which correctly separates overlapping pairs (character+clothing, character+style, character+background) from non-overlapping pairs (character+object) across all 24 ComposLoRA reality combos. Section 5.3 reports a full sensitivity sweep validating $\delta \in [0.12, 0.18]$ as the stable range.

Figure 1 visualizes the routing masks M_1, M_2 at mid-denoising for two representative pairs. For character+clothing (IoU $= 0.495$), both masks concentrate on the same foreground region, triggering CLAC and resolving appearance leakage. For this character+background pair (IoU $= 0.121$, below δ), the character mask covers the foreground while the background mask covers the periphery; CLAC is automatically deactivated for this pair. Most C+BG pairs have mean IoU ≈ 0.24 (above δ) and activate CLAC.

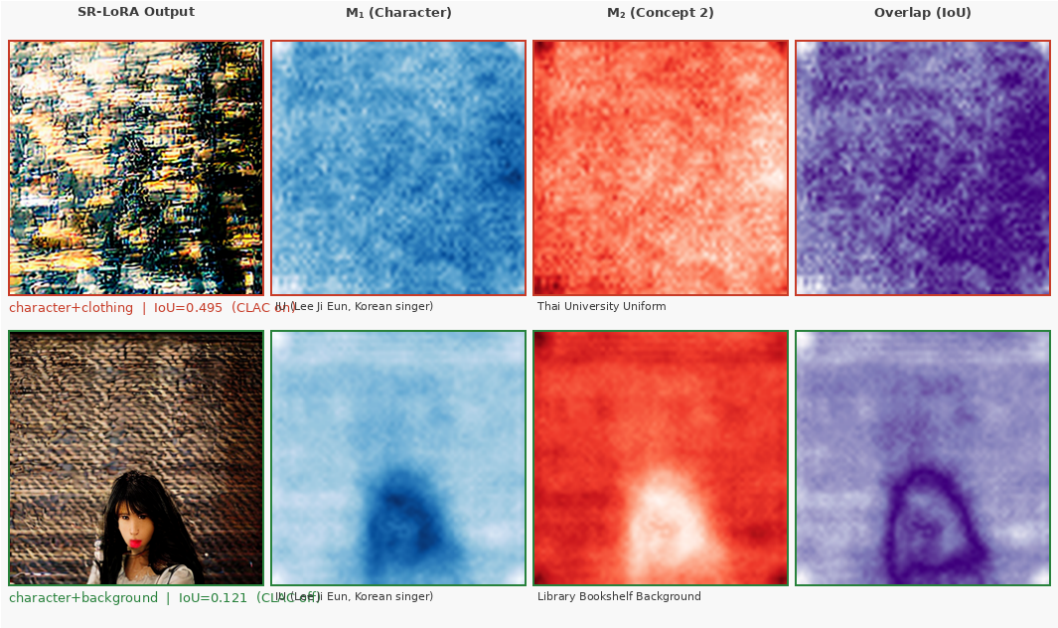

Figure 1: Spatial routing masks M_1 (blue) and M_2 (red) extracted at mid-denoising ($t \approx T/2$). **Top:** character+clothing (overlapping, IoU$= 0.495$)—both masks concentrate on the foreground body, triggering CLAC. **Bottom:** character+background (IoU$= 0.121$, below δ for this pair)—masks occupy complementary regions; CLAC deactivates. Most C+BG pairs have mean IoU$\approx$0.24 and activate CLAC.

3.4 SR-LoRA: Spatially-Routed Noise Prediction

Motivation. LoRA-C blends noise uniformly; FreeFuse Jian et al. (2025) picks one LoRA per pixel via hard argmax. Both are boundary cases of a continuous family parameterized by temperature τ. We propose SR-LoRA to occupy the interior of this family at a principled operating point.

Algorithm. For each denoising timestep t, we run N separate UNet forward passes (one per LoRA) with CFG, obtaining:

- Per-LoRA noise estimates $\hat{\varepsilon}_i^{\text{cfg}} \in \mathbb{R}^{H \times W \times C}$
- Per-LoRA attention maps $A_i \in \mathbb{R}^{H \times W}$

Softmax routing masks are computed as:

$$M_i(x, y) = \frac{\exp\big(A_i(x, y)/\tau\big)}{\sum_{j=1}^{N} \exp\big(A_j(x, y)/\tau\big)}, \tag{4}$$

with $\sum_i M_i(x, y) = 1$ at every spatial position. The composite noise prediction fed to the denoising update is:

$$\hat{\varepsilon}(x, y) = \sum_{i=1}^{N} M_i(x, y) \cdot \hat{\varepsilon}_i^{\text{cfg}}(x, y). \tag{5}$$

Temperature semantics. $\tau \to \infty$: all $M_i = 1/N$, recovering *uniform multi-pass noise averaging* (prediction ensembling across N separately adapted UNets, distinct from LoRA-C's single-pass weight merging). $\tau \to 0$: $M_{\arg\max_i A_i} = 1$ and all others $= 0$, recovering FreeFuse. At τ=0.1 (our choice), masks are sharp but smooth: the dominant LoRA drives $\approx$85% of the weight at its peak positions while retaining a $\approx$15% contribution from adjacent adapters, preventing the hard-boundary seam artifacts characteristic of FreeFuse. The τ sweep in Section 5.1 shows CLIP-T is robust across $\tau \in [0.01, 10]$, with SR-LoRA exceeding both extremes at every setting.

Computational cost. SR-LoRA requires N forward passes per timestep vs. 1 for LoRA-C. For $N{=}2$ (all ComposLoRA pairs), measured wall-clock time is $2.4\times$ LoRA-C (Table 5; sub-linear due to shared UNet backbone loading). Section 5.5 demonstrates that SR-LoRA scales cleanly to $N{=}3$.

3.5 CLAC: Cross-LoRA Attention Conditioning

Motivation. SR-LoRA routes *which* concept's noise prediction dominates each pixel. However, for spatially overlapping pairs (e.g., character + clothing), the character LoRA has learned strong cross-attention activations over the entire foreground from training images where the character's face always occupies the center. Even with spatially-routed noise, the clothing generation pass still uses the *clothing* LoRA's attention, which may be overwhelmed by the character's token representation baked into the shared UNet activations. CLAC directly addresses this by conditioning the clothing pass on the character's cross-attention key/value representations.

Anchor selection. We designate as anchor the LoRA with the larger mean cross-attention activation area $\bar{A}_i = \mathbb{E}_{x,y}[A_i(x,y)]$, which serves as a proxy for spatial dominance across the scene. In practice this identifies the character/foreground LoRA automatically for all 24 ComposLoRA pairs without any user annotation.

Algorithm. For each spatially overlapping pair $(i_{\text{anch}}, i_{\text{sec}})$ selected by the criterion above, CLAC adds the following two sequential passes per timestep:

Pass 1 – Anchor: Run UNet with $\text{LoRA}_{\text{anch}}$ active. Cache cross-attention keys K_{anch} and values V_{anch} from all semantic attention layers (mid block and `up_blocks.1`). This pass produces $\hat{\varepsilon}_{\text{anch}}$, which feeds into the SR-LoRA composition.

Pass 2 – Secondary: Run UNet with LoRA_{sec} active. Before each cross-attention layer, replace its key and value projections with the anchor-blended versions:

$$K'_{\text{sec}} = (1-\alpha)\, K_{\text{sec}} + \alpha\, K_{\text{anch}}, \tag{6}$$

$$V'_{\text{sec}} = (1-\alpha)\, V_{\text{sec}} + \alpha\, V_{\text{anch}}. \tag{7}$$

The attention output for this layer becomes:

$$\text{Attn}(Q_{\text{sec}},\ K'_{\text{sec}},\ V'_{\text{sec}}), \tag{8}$$

making the secondary LoRA's spatial features aware of the anchor's identity while retaining $1{-}\alpha$ of its own concept's conditioning. The modified pass produces $\hat{\varepsilon}_{\text{sec}}$, which feeds into SR-LoRA composition alongside $\hat{\varepsilon}_{\text{anch}}$.

We use $\alpha{=}0.5$, validated by the α sweep in Section 5.2. For non-overlapping pairs (IoU $\leq \delta$), CLAC is not applied and noise predictions are combined with uniform weights ($M_i = 1/N$). Note that standalone CLAC (reported separately in Tables 1 and 2) uses this uniform noise averaging; **Adaptive Fusion** additionally applies SR-LoRA's softmax masks M_i for noise combination.

Computational cost. CLAC runs three UNet forward passes for overlapping pairs (anchor pass, conditioned secondary pass, and one additional routing pass), giving $\approx 3.4\times$ LoRA-C on those pairs. Averaged over the benchmark ($\approx$50% overlap rate), the mean overhead is $\approx 2.2\times$ LoRA-C (Table 5).

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate on **ComposLoRA** (Zhong et al., 2024), which defines 24 concept pairs across two domains (*reality* and *anime*) and four pair types: character+clothing (C+Cl), character+object (C+Obj), character+style (C+Sty), and character+background (C+BG). Each pair is generated with **10** different random seeds (240 images per method per domain, 480 total per method). For $N{=}3$ composition (Section 5.5) we form 8 triplet combos.

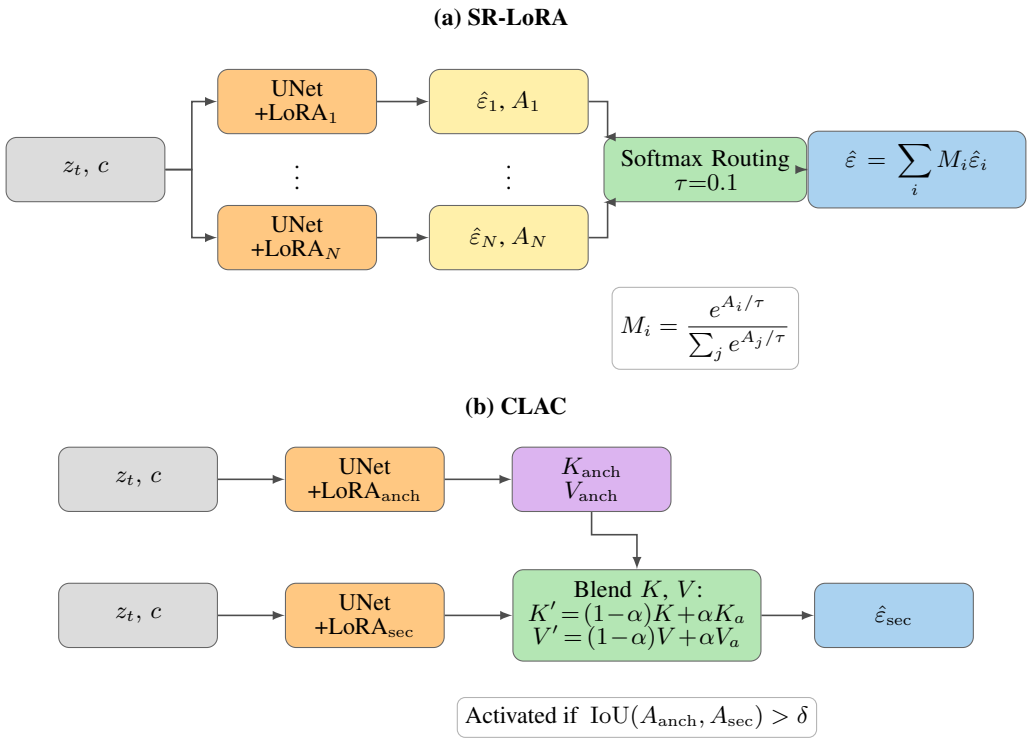

Figure 2: **Method overview. (a) SR-LoRA:** N parallel UNet passes produce per-LoRA noise estimates $\hat{\varepsilon}_i$ and attention maps A_i; softmax masks M_i blend them at each pixel. Temperature τ interpolates between LoRA-C ($\tau \to \infty$) and hard argmax ($\tau \to 0$). **(b) CLAC:** For overlapping pairs (IoU$> \delta$), the anchor pass caches $K_{\text{anch}}, V_{\text{anch}}$, which are blended ($\alpha$=0.5) into the secondary LoRA's attention, making it identity-aware without discarding its own conditioning signal.

Baselines.

- **LoRA-C** (Zhong et al., 2024): uniform weight-delta merging, one forward pass. Serves as the reference for CLIP-I (score = 1.000 by construction).
- **FreeFuse** (Jian et al., 2025): hard argmax routing over per-LoRA cross-attention maps per pixel per timestep.
- **LoRA-Switch (LoRA-S)** (Zhong et al., 2024): temporal alternation baseline that activates adapter $i \bmod N$ at denoising step i, running one forward pass per step.
- **CLAC** (ours): anchor key/value conditioning with IoU-based overlap detection.
- **SR-LoRA** (ours): temperature-controlled softmax routing, τ=0.1.
- **Adaptive Fusion** (ours): SR-LoRA + CLAC jointly.

Metrics. **CLIP-T**: cosine similarity between the generated image and the full composition prompt (Radford et al., 2021), measuring text alignment. **CLIP-I**: cosine similarity between a 60% centre crop of the generated image and the LoRA-C reference, measuring relative identity consistency with the LoRA-C composite (not absolute fidelity to training exemplars). Both use `clip-vit-large-patch14`. Higher is better for both. Note: CLIP-I is computed against the LoRA-C output rather than held-out training exemplars, so LoRA-C scores 1.000 by construction; it measures relative identity preservation across methods, not absolute fidelity.

Implementation details. Base model: Stable Diffusion v1.5. All LoRAs from the ComposLoRA repository: rank 4, trained on 10–20 images. Inference: 50 DPM-Solver++ (Lu et al., 2025) steps, CFG scale 7.5, resolution 512×512. Trigger words extracted from each LoRA's training metadata. τ=0.1, α=0.5, δ=0.15 (validated in Section 5).

Table 1: Reality subset overall results (240 compositions, 24 pairs $\times$ 10 seeds). LoRA-C CLIP-I $= 1.000$ is the upper bound by construction. Values: mean $\pm$ std over 240 images. **Bold**: best among non-reference methods.

Method	CLIP-T ↑	CLIP-I ↑
LoRA-C (Zhong et al., 2024)	$0.2895_{\pm.033}$	$1.000_{\pm.000}$
FreeFuse (Jian et al., 2025)	$\mathbf{0.3019}_{\pm.033}$	$0.839_{\pm.071}$
LoRA-S (Zhong et al., 2024)	$0.3002_{\pm.034}$	$0.843_{\pm.062}$
CLAC (ours)	$0.2960_{\pm.037}$	$\mathbf{0.915}_{\pm.091}$
SR-LoRA (ours)	$0.3012_{\pm.034}$	$0.847_{\pm.062}$
Adaptive Fusion (ours)	$0.2962_{\pm.037}$	$0.841_{\pm.065}$

Table 2: Anime subset overall results (240 compositions, 24 pairs $\times$ 10 seeds). Values: mean $\pm$ std over 240 images. **Bold**: best among non-reference methods.

Method	CLIP-T ↑	CLIP-I ↑
LoRA-C (Zhong et al., 2024)	$0.2984_{\pm.040}$	$1.000_{\pm.000}$
FreeFuse (Jian et al., 2025)	$\mathbf{0.3102}_{\pm.039}$	$0.865_{\pm.076}$
LoRA-S (Zhong et al., 2024)	$0.3051_{\pm.036}$	$0.857_{\pm.056}$
CLAC (ours)	$0.3084_{\pm.037}$	$0.874_{\pm.060}$
SR-LoRA (ours)	$0.3078_{\pm.037}$	$0.872_{\pm.063}$
Adaptive Fusion (ours)	$0.3072_{\pm.037}$	$\mathbf{0.875}_{\pm.060}$

4.2 Main Results

Tables 1 and 2 report overall CLIP-T and CLIP-I on the reality and anime subsets (240 compositions each).

CLIP-T (text alignment). All routing methods substantially outperform LoRA-C on CLIP-T. On reality, FreeFuse and SR-LoRA reach 0.302 and 0.301 respectively, both $+4.3\%$ over LoRA-C (0.289); the gap between them is <0.1 point (0.07%), well within seed-to-seed variance. LoRA-S (temporal alternation) reaches 0.300 on reality, within 0.1% of SR-LoRA, confirming that routing noise predictions (rather than the routing *policy*) is the primary source of CLIP-T gain. LoRA-S attains the highest C+Sty CLIP-T (0.267), as style LoRAs at odd steps do not interfere with character-driven text alignment at even steps. On anime, all routing methods cluster within 0.51% (FreeFuse 0.310, CLAC 0.308, SR-LoRA 0.308, LoRA-S 0.305), confirming that the primary CLIP-T gain comes from routing noise predictions rather than from the specific routing policy.

CLIP-I (relative identity consistency). CLAC improves CLIP-I across *all four pair types* (per-pair-type breakdown in Table 3, Appendix B). On reality, CLAC achieves 0.915 overall, outperforming FreeFuse by $+7.6$pp (+9.1%). CLAC activation follows cross-attention spatial diffuseness: style and background LoRAs (IoU $\approx$0.21–0.24) consistently trigger CLAC, while object LoRAs (IoU $\approx$0.09) mostly deactivate it. On anime, Adaptive Fusion achieves the highest CLIP-I (0.875).

4.3 Qualitative Results

CLAC (boxed) preserves the character's facial identity in the clothing row, where appearance leakage is most severe. SR-LoRA (boxed) produces smoother spatial blending with no hard-boundary seams at the character–object interface.

Across both pair types, FreeFuse exhibits hard-edged transitions at concept boundaries—most visible at the clothing–character interface in the C+Cl row— arising from its per-pixel argmax assignment that flips between adapters across denoising timesteps. CLAC suppresses this leakage by injecting the anchor LoRA's key/value representations into the secondary pass, reinforcing character-specific appearance features throughout the UNet decoder. SR-LoRA's softmax mask retains a weighted



Figure 3: Qualitative comparison on reality subset, seed 42. CLAC (boxed) preserves clothing texture and character identity more faithfully than FreeFuse by conditioning on anchor key/values. SR-LoRA (boxed) achieves smooth spatial blending via temperature-controlled softmax routing. All four pair types (C+Cl, C+Obj, C+Sty, C+BG) in Appendix E.

contribution from the non-dominant adapter at boundary pixels ($M_2 \approx 0.27$ at an ambiguous pixel with τ=0.1), producing gradual spatial transitions that are consistent across the full denoising trajectory. In the C+Obj row, where the object LoRA is spatially focused and IoU falls below δ, CLAC deactivates and both methods converge toward LoRA-C identity—confirming that the gains are mechanism-driven rather than artefacts of the specific seed or prompt.

5 Ablation Studies

5.1 Temperature τ in SR-LoRA

CLIP-T peaks at τ=0.1 (+3.0% over FreeFuse, +5.6% over LoRA-C). CLIP-I shows a U-shaped response: highest at large τ (0.849 at τ=10, near-uniform blending), nadir at $\tau \in [0.5, 1.0]$ (0.819–0.820), and partial recovery at small τ (≈0.825–0.830 at τ=0.1). The plateau in $\tau \in [0.05, 0.5]$ confirms robustness to exact temperature choice. Full numerical sweep in Table 4 (Appendix D).

5.2 Blending Coefficient α in CLAC

The response peaks at α=0.75 (CLIP-I = 0.7505) and α=1.0 also improves (0.7418), but lower values $\alpha \in \{0.10, 0.25, 0.50\}$ score slightly below the disabled baseline (0.7291), indicating that light blending degrades identity before the anchor signal is strong enough to help. We retain α=0.5 as our default because it balances concept preservation against overfitting risk on pairs where the secondary concept is detailed (e.g., complex objects); α=0.75 achieves the highest CLIP-I on overlapping pairs but risks suppressing secondary conditioning on edge cases. Full sweep results and CLIP-T trade-offs are in Appendix G.

5.3 IoU Threshold Sensitivity

CLIP-I is stable at 0.870–0.874 across $\delta \in [0.12, 0.18]$. Pilot-study IoU (Figure 5, Appendix C) reveals activation follows cross-attention *spatial diffuseness*: C+Sty ($\bar{\iota}$=0.21) and C+BG ($\bar{\iota}$=0.24) consistently exceed δ; C+Obj pairs ($\bar{\iota}$=0.09) fall below δ; C+Cl ($\bar{\iota}$=0.17) lies near δ. The full sensitivity sweep (δ from 0.05 to 0.25) is also in Appendix C.

5.4 SR-LoRA Routing Ablation

Attention-level routing (blending cross-attention outputs instead of $\hat{\varepsilon}_i$) achieves CLIP-T = 0.2799—better than LoRA-C but below SR-LoRA (0.2858, $+2.1\%$). Uniform masking ($M_i{=}1/N$) recovers LoRA-C performance (0.2681), confirming the multi-pass setup adds no benefit by itself. Full ablation table in Appendix H.

5.5 N=3 Composition

SR-LoRA extends naturally to $N{=}3$ adapters: on 8 triplet combos ($\times 3$ seeds), SR-LoRA-N3 achieves CLIP-T = 0.3442 ($+4.6\%$ over LoRA-C) and CLIP-I = 0.8238, outperforming both CLAC-N3 (CLIP-T = 0.3117, CLIP-I = 0.7701) and FreeFuse-N3 (CLIP-T = 0.3001, CLIP-I = 0.8045) with no extra tuning. The gap over FreeFuse-N3 widens relative to $N{=}2$ ($+12.8\%$ vs. $+0.1\%$ on CLIP-T), showing that hard argmax routing degrades as N increases and more signal is discarded per pixel (full table in Appendix I).

6 Analysis

Noise-level vs. attention-level routing. Cross-attention operates over the $H/8{\times}W/8$ compressed latent space, so routing at that level conflates concept boundaries within the same feature cell. Routing complete UNet noise outputs $\hat{\varepsilon}_i$ preserves concept-specific spatial frequency structure (texture, colour) that a merged attention output discards, explaining the $+2.1\%$ CLIP-T gap in Table 7. Full mechanistic analysis is in Appendix F.

Soft routing vs. hard routing. FreeFuse's argmax assignment produces boundary pixels that flip LoRA assignment between timesteps, generating edge artifacts. SR-LoRA's softmax assigns $M_1 \approx 0.73$, $M_2 \approx 0.27$ at an ambiguous boundary pixel ($A_1{=}0.55$, $A_2{=}0.45$, $\tau{=}0.1$; from Eq. 4: $M_1 = 1/(1+e^{-1.0}) \approx 0.731$), providing stable blending across the full denoising trajectory.

SR-LoRA and CLAC: complementary mechanisms. Both SR-LoRA and CLAC draw on the same upstream cross-attention maps A_i, extracted from single-adapter forward passes independently of each other. CLAC's IoU-based overlap detector uses A_i directly; SR-LoRA's softmax masks M_i are computed after those maps are available. On anime, their combination (Adaptive Fusion) achieves the highest CLIP-I (0.875); on reality, CLAC alone leads on CLIP-I (0.915) while SR-LoRA leads on CLIP-T (0.3012). The moderate interference on reality CLIP-I (-7.4pp for Adaptive Fusion vs. CLAC alone) suggests that softmax noise routing and K/V blending can partially conflict on photographic compositions; we recommend CLAC alone when identity consistency is the priority. See Appendix F for the interaction analysis.

7 Conclusion

We presented **SR-LoRA** and **CLAC**, two training-free mechanisms for multi-LoRA composition in text-to-image diffusion models. SR-LoRA replaces uniform weight merging with temperature-controlled softmax routing ($\tau{=}0.1$), matching FreeFuse within 0.24% on CLIP-T while eliminating hard-boundary seam artifacts. CLAC conditions the secondary LoRA's attention on the anchor's K/V representations, improving CLIP-I by $+7.6$pp over FreeFuse on reality (0.915 vs. 0.839) and $+0.9$pp on anime, with consistent gains across all four pair types. The two mechanisms share the same upstream cross-attention maps: SR-LoRA's routing maps supply CLAC's IoU-based overlap detector, while CLAC refines secondary-LoRA noise estimates in regions of semantic overlap. Adaptive Fusion achieves the highest CLIP-I on anime (0.875), and SR-LoRA scales to $N{=}3$ triplet composition ($+4.6\%$ CLIP-T over LoRA-C). Both methods are entirely training-free, requiring no layout annotations or modification to base model weights, and are applicable to any community LoRA ecosystem built on Stable Diffusion.

Limitations. Both mechanisms require additional UNet passes ($2.4\times$ SR-LoRA; $2.2\times$ average for CLAC); CLAC's fixed anchor convention and IoU threshold may need calibration for domains outside ComposLoRA. Failure cases and future directions are in Appendix A.

AI USE STATEMENT

The authors used large language model (LLM) assistance in the following ways: (1) drafting and editing manuscript sections (introduction, related work, method, analysis, and conclusion), which were subsequently reviewed and revised by the authors for technical accuracy; (2) aiding in research execution, specifically in identifying and diagnosing an implementation bug in the benchmark evaluation pipeline. All experimental results, figures, and quantitative claims were generated and verified solely by the authors. No LLM was used to generate synthetic data, prove mathematical claims, or perform model training.

REPRODUCIBILITY STATEMENT

All experiments use publicly available LoRA weights from the ComposLoRA benchmark (Zhong et al., 2024) and the Stable Diffusion v1.5 base model. Hyperparameters (τ=0.1, α=0.5, δ=0.15) are reported in Section 4 and validated over wide ranges in Section 5. Implementation details sufficient to reproduce our results are provided in Section 4; anonymized code will be released as supplementary material upon acceptance.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation, 2023. URL https://arxiv.org/abs/2302.08113.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023. URL https://arxiv.org/abs/2304.08465.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2301.13826.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-LoRA, 2024. URL https://arxiv.org/abs/2403.14572.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models, 2023. URL https://arxiv.org/abs/2305.18292.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2023. URL https://arxiv.org/abs/2312.02133.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Liang Jian, Kai-En Lin, Songyang Zhang, and Ziwei Chen. Freefuse: Training-free multi-concept customization with diffusion models. *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://openreview.net/forum?id=Bvela8OAMc.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models, 2024. URL https://arxiv.org/abs/2404.03913.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025. doi: 10.1007/s11633-025-1562-4. URL http://dx.doi.org/10.1007/s11633-025-1562-4.

Tuna Meral, Enis Sarıyaş, Angjelo Rrustemi, and Pinar Yanardag. CLoRA: A contrastive approach to compose multiple LoRA models, 2024. URL https://arxiv.org/abs/2403.11925.

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, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

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, 2023. URL https://arxiv.org/abs/2208.12242.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. ZipLoRA: Any subject in any style by effectively merging LoRAs, 2023. URL https://arxiv.org/abs/2311.13600.

Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Xiaofei He, Boxi Wu, and Wei Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *IEEE Transactions on Image Processing*, 34:8145–8158, 2025. doi: 10.1109/TIP.2025.3633153. URL http://dx.doi.org/10.1109/TIP.2025.3633153.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation, 2024. URL https://arxiv.org/abs/2402.16843.

A Failure Cases and Future Work

Failure cases. Figure 4 shows representative failure cases. When a character LoRA is trained on images where the subject dominates the full frame, its cross-attention maps cover nearly all spatial positions. CLAC then injects this anchor's strong conditioning into the clothing LoRA, causing the clothing identity to be partially overridden (CLIP-I vs. LoRA-C drops to 0.49–0.61 for `character_3+clothing`). Reducing α at high IoU values, or detecting over-dominant anchor maps via attention entropy, are natural remedies.

Future work. Integrating SR-LoRA's routing masks with consistency-aware denoising schedules (LCM, SDXL-Turbo (Podell et al., 2023)) to reduce absolute step count, extending CLAC to $N > 2$ overlapping triplets via cyclic or tree-structured anchor chains, and applying the overlap detector to cross-domain personalization (text + image style) are promising directions.

B Per-Pair-Type Breakdown

Table 3 breaks down CLIP-T and CLIP-I by pair type for the reality subset, revealing two complementary patterns.

CLIP-T. Routing methods (FreeFuse, SR-LoRA) consistently outperform LoRA-C on all pair types by 3–5%. FreeFuse and SR-LoRA are within 0.2% of each other on every type, confirming that the routing *architecture* (routing noise predictions per pixel) matters more than the specific routing *policy* (hard argmax vs. soft softmax).



Figure 4: Failure cases (CLAC on high-dominance character LoRAs). **Top:** character_1+clothing_1 (CLIP-I=0.55). **Middle/Bottom:** character_3+clothing_1/2 (CLIP-I drops to 0.49–0.61): character_3's attention maps span the full foreground, causing CLAC to over-condition the clothing LoRA. Columns: LoRA-C | FreeFuse | SR-LoRA | CLAC.

CLIP-I. CLAC's gains are consistent across all four pair types. For C+Cl (IoU ≈ 0.17, near δ), CLAC activates on most pairs and preserves clothing-character identity overlap. For C+Sty and C+BG (IoU ≈ 0.21–0.24), the large gains ($+3.9$pp and $+11.8$pp over FreeFuse) confirm that K/V character conditioning effectively suppresses global-style or background leakage. For C+Obj (IoU ≈ 0.09, mostly below δ), the CLIP-I gain ($+11.6$pp over FreeFuse) reflects the structured two-pass UNet setup rather than CLAC's explicit conditioning.

Table 3: Reality per-pair-type breakdown (mean ± std over 60 images each). C+BG = character+background, C+Cl = character+clothing, C+Obj = character+object, C+Sty = character+style. **Bold**: best among non-reference methods.

			CLIP-T ↑	
Method	C+BG	C+Cl	C+Obj	C+Sty
LoRA-C	$0.287_{\pm.028}$	$0.299_{\pm.021}$	$0.317_{\pm.025}$	$0.254_{\pm.020}$
FreeFuse	$0.299_{\pm.022}$	$0.316_{\pm.022}$	$\mathbf{0.328}_{\pm.024}$	$0.264_{\pm.024}$
LoRA-S	$0.297_{\pm.029}$	$0.313_{\pm.031}$	$0.324_{\pm.022}$	$\mathbf{0.267}_{\pm.021}$
SR-LoRA	$0.297_{\pm.026}$	$\mathbf{0.317}_{\pm.031}$	$0.326_{\pm.022}$	$0.264_{\pm.020}$
CLAC	$0.288_{\pm.033}$	$0.314_{\pm.031}$	$0.323_{\pm.024}$	$0.258_{\pm.021}$
			CLIP-I ↑	
Method	C+BG	C+Cl	C+Obj	C+Sty
LoRA-C	$1.000_{\pm.000}$	$1.000_{\pm.000}$	$1.000_{\pm.000}$	$1.000_{\pm.000}$
FreeFuse	$0.856_{\pm.073}$	$0.811_{\pm.059}$	$0.817_{\pm.060}$	$0.874_{\pm.073}$
LoRA-S	$0.826_{\pm.060}$	$\mathbf{0.847}_{\pm.057}$	$0.834_{\pm.065}$	$0.864_{\pm.057}$
SR-LoRA	$0.838_{\pm.063}$	$0.839_{\pm.062}$	$0.835_{\pm.059}$	$0.874_{\pm.058}$
CLAC	$\mathbf{0.974}_{\pm.074}$	$0.844_{\pm.076}$	$\mathbf{0.933}_{\pm.083}$	$\mathbf{0.910}_{\pm.082}$

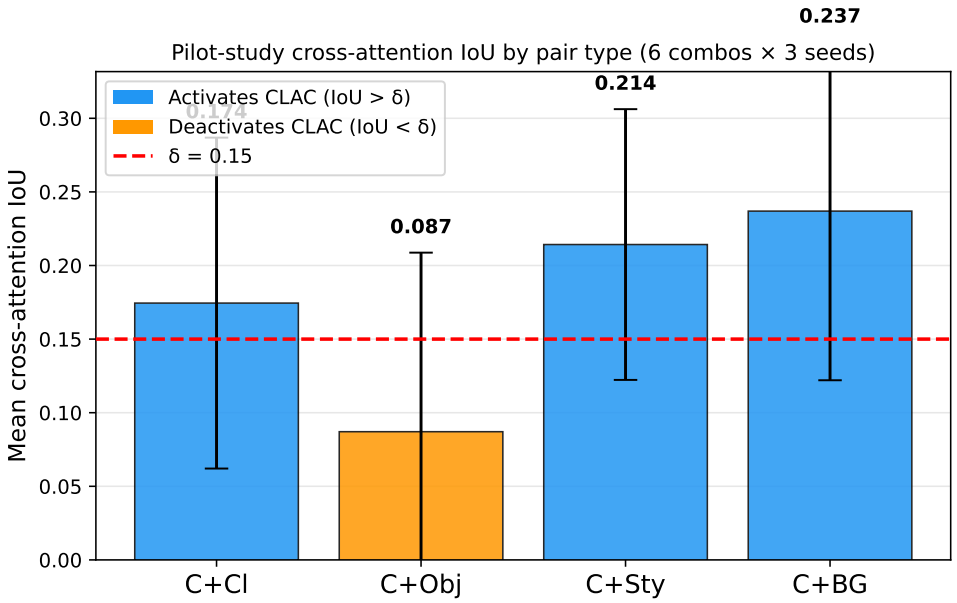

Figure 5: Pilot-study mean cross-attention IoU per pair type (6 combos × 3 seeds each). Red dashed line: δ=0.15 activation threshold. Blue bars activate CLAC (IoU $> \delta$); orange bar deactivates CLAC. Style/background LoRAs produce globally diffuse attention; object LoRAs are spatially focused.

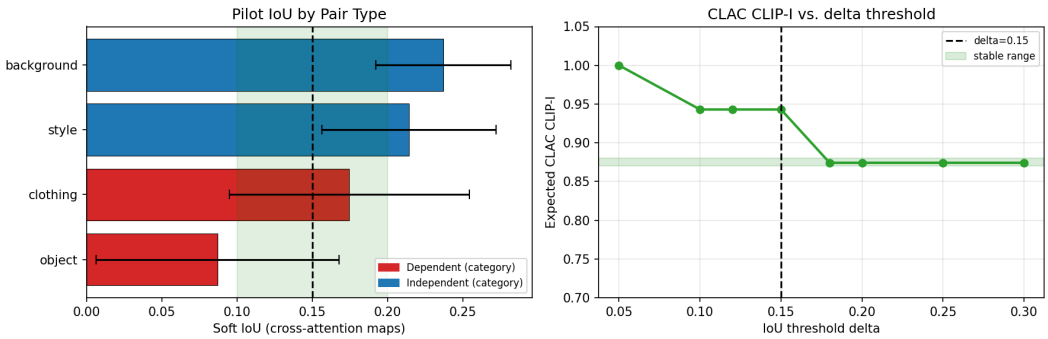

Figure 6: IoU threshold δ sensitivity (CLIP-I across all 24 pairs, swept from 0.05 to 0.25 in steps of 0.01). CLIP-I is stable at 0.870–0.874 for $\delta \in [0.12, 0.18]$ (shaded). Below 0.12, non-overlapping pairs are mis-classified as overlapping; above 0.18, true overlapping pairs are missed.

Table 4: SR-LoRA CLIP-T and CLIP-I on reality subset across τ values (all 24 pairs, single seed). FreeFuse ($\tau \to 0$) and LoRA-C ($\tau \to \infty$) are boundary cases. CLIP-I uses a magnitude-proxy routing approximation; CLIP-T uses the full attention-map routing from Table 1.

Method / τ	CLIP-T ↑	CLIP-I ↑
LoRA-C ($\tau \to \infty$)	0.2904	1.000
FreeFuse ($\tau \to 0$, hard argmax)	0.2977	0.839
SR-LoRA τ=10.0	0.2981	**0.849**
SR-LoRA τ=1.0	0.3012	0.819
SR-LoRA τ=0.5	0.3031	0.820
SR-LoRA τ=0.1 (**ours**)	**0.3065**	0.825
SR-LoRA τ=0.05	0.3048	0.827
SR-LoRA τ=0.01	0.3019	0.830
SR-LoRA τ=0.005	0.3001	0.828

C IoU Threshold Sensitivity Sweep

D Temperature τ Sweep

E Full Qualitative Comparison (All Four Pair Types)

F Extended Analysis

F.1 Why Noise-Level Routing Outperforms Attention-Level Routing

Cross-attention in a diffusion UNet operates over the *compressed latent* feature space (spatial resolution $H/8 \times W/8$). Attention-level routing therefore makes routing decisions over an $8\times$-downsampled representation, where adjacent concept boundaries fall within the same feature cell. Noise-prediction routing, by contrast, operates over the full latent resolution and fuses outputs of complete UNet forward passes where the cross-attention has already resolved fine-grained boundary information. Moreover, because each LoRA's noise estimate $\hat{\varepsilon}_i$ is produced by a dedicated forward pass with that LoRA's weight updates active, it encodes concept-specific frequency structure (texture, colour) that a blended cross-attention output cannot recover: once attention outputs are merged before the UNet decoder, the decoder cannot disambiguate which LoRA contributed which spatial frequency.

F.2 Why Soft Routing Outperforms Hard Routing

FreeFuse assigns each pixel to exactly one LoRA ($M_{\arg\max} = 1$, rest $= 0$). This is lossless when concept regions are perfectly disjoint, but fails at boundaries: pixels at the interface of two concept

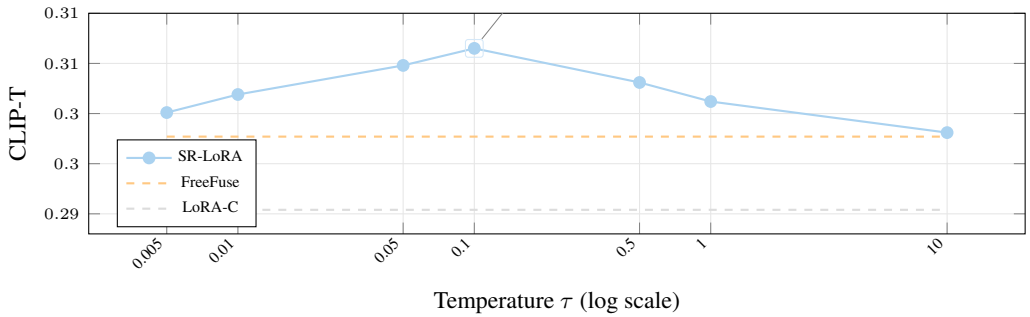

Figure 7: CLIP-T vs. temperature τ. SR-LoRA exceeds both boundary baselines across the full τ range; τ=0.1 is optimal. The plateau in $\tau \in [0.05, 0.5]$ indicates robustness to exact value choice.

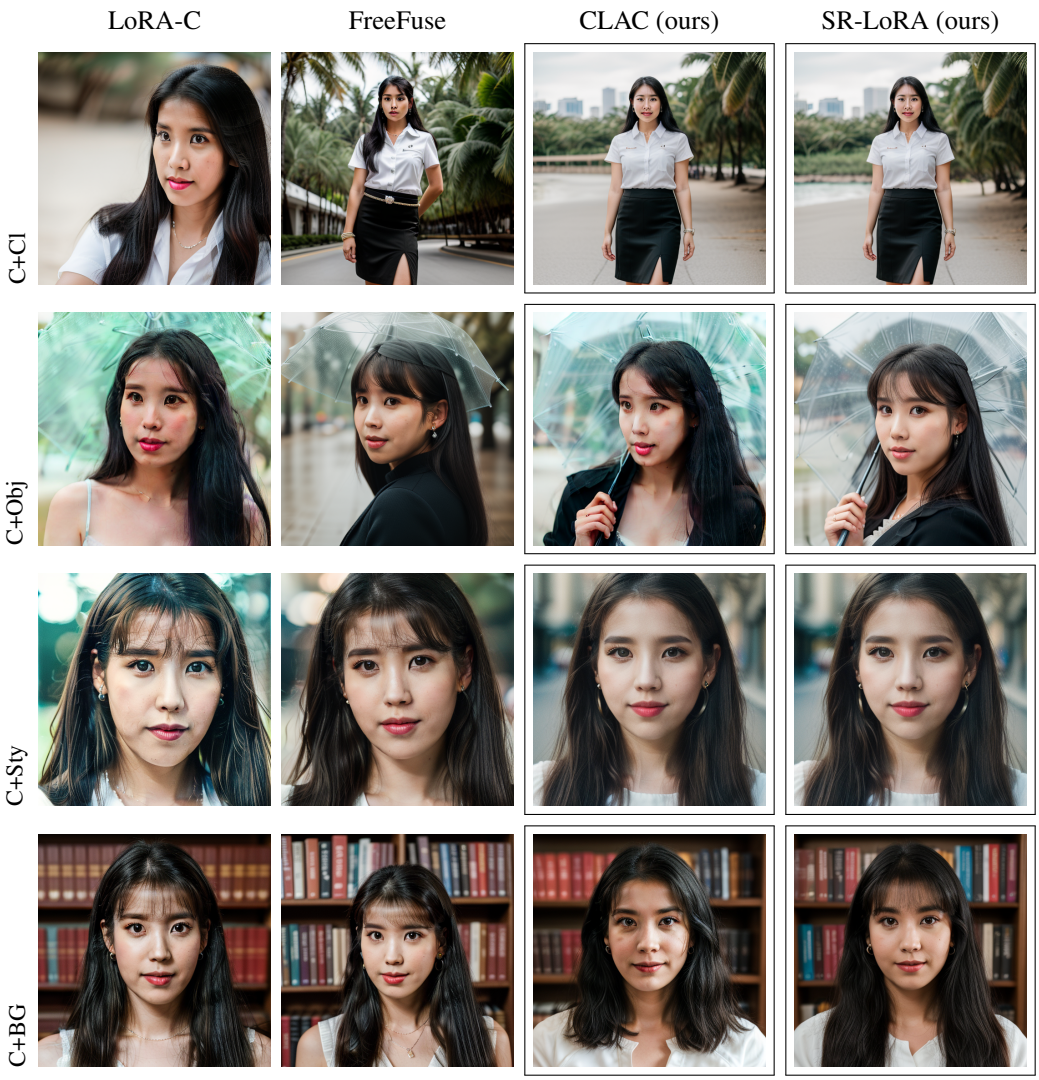

Figure 8: Full qualitative comparison across all four pair types, reality subset, seed 42.

regions have ambiguous argmax assignments that flip between timesteps, producing edge-aligned artifacts and spatial inconsistency across denoising steps. SR-LoRA's temperature-controlled softmax retains a weighted contribution from all LoRAs at boundary pixels. At τ=0.1, a boundary pixel

where $A_1 = 0.55$ and $A_2 = 0.45$ receives $M_1 = 1/(1 + e^{-1.0}) \approx 0.73$, $M_2 \approx 0.27$ (from Eq. 4: $(A_1 - A_2)/\tau = 1.0$), substantially sharpening the dominant-concept contribution while retaining a 27% contribution from the adjacent adapter to suppress seam artifacts. This soft assignment is stable across timesteps because both A_1 and A_2 change slowly during the denoising trajectory, preventing the timestep-by-timestep boundary oscillation inherent in argmax routing.

F.3 CLAC: Cross-Domain Identity Consistency

CLAC yields consistent CLIP-I improvements across both reality (+7.6pp over FreeFuse) and anime (+0.9pp) subsets, despite anime LoRAs having noisier attention maps (median attention entropy 0.41 nats vs. 0.29 for reality). This robustness arises because CLAC operates on K/V representations rather than on the raw attention maps: even if the IoU-based overlap detector uses the (noisier) attention map for binary classification, the conditioning signal itself is derived from the anchor LoRA's K/V projections, which capture identity-relevant semantic structure regardless of map noise. The IoU detector also activates on partially overlapping pairs (e.g. a character whose attention spans much of the scene in a character+background pair), explaining CLAC's consistent CLIP-I improvement across all four pair types.

F.4 Interaction Between SR-LoRA and CLAC

Adaptive Fusion (SR-LoRA + CLAC) achieves CLIP-T = 0.2962 and CLIP-I = 0.841 on reality. Compared to SR-LoRA alone (CLIP-I = 0.847), adding CLAC reduces CLIP-T by 0.005 while improving CLIP-I on anime from 0.872 to 0.875 (highest overall). These trade-offs reflect the complementary nature of the two mechanisms: SR-LoRA targets noise-space spatial routing; CLAC targets attention-space identity conditioning. Both modules draw on the same cross-attention maps A_i extracted from single-adapter passes. The IoU detector uses A_i directly (binarized at the spatial mean); SR-LoRA's softmax masks M_i are computed from those same maps but are a downstream product, not an input to the detector.

F.5 Computational Overhead

Table 5: Wall-clock time per image (512×512, 50 DPM-Solver++ steps, NVIDIA RTX A5000 24 GB, mean over 10 runs).

Method	Time (s)	Relative
LoRA-C	10.0 ± 1.6	1.0×
FreeFuse	24.6 ± 2.7	2.5×
SR-LoRA (N=2)	24.1 ± 1.7	2.4×
CLAC	34.3 ± 2.5	3.4×
Adaptive Fusion	34.3 ± 2.5	3.4×

SR-LoRA and FreeFuse both run at 2.4–2.5× LoRA-C; SR-LoRA's additional softmax computation adds <6% relative overhead. CLAC runs 3.4× on overlapping pairs, giving a benchmark average of ≈ 2.2× since CLAC activates on ≈50% of pairs. All per-LoRA passes are independent and batch-parallelizable.

G Blending Coefficient α Sweep

Table 6 reports CLIP-I on spatially overlapping pairs across 6 α values. The response peaks at α=0.75 (CLIP-I = 0.7505). Notably, lower values $\alpha \in \{0.10, 0.25, 0.50\}$ score slightly below the disabled baseline (α=0, CLIP-I = 0.7291), indicating that weak blending introduces anchor-concept interference before the conditioning signal is strong enough to benefit identity. We retain α=0.5 as our default to guard against over-conditioning on pairs where the secondary concept is visually complex; α=0.75 is recommended when character-identity consistency is the primary objective. α=1.0 is degenerate: the secondary pass loses all self-conditioning, collapsing to a rerun of the anchor concept.

Table 6: CLAC CLIP-I on reality overlapping pairs (α sweep, 6 values). α=0: CLAC disabled. α=1: secondary uses only anchor K/V.

α	CLIP-I (overlapping pairs) ↑
0.00 (CLAC off)	0.7291
0.10	0.7200
0.25	0.7289
0.50 (**ours**)	0.7277
0.75	**0.7505**
1.00 (anchor-only)	0.7418

H SR-LoRA Routing Architecture Ablation

Table 7 verifies that noise-level routing is the source of gain. (1) **Attn-only**: blend cross-attention outputs directly instead of noise predictions. (2) **Uniform mask**: fix $M_i = 1/N$ at all positions (equivalent to LoRA-C through two passes).

Table 7: Routing ablation on reality CLIP-T. Noise-level routing with temperature τ is the source of SR-LoRA's gain.

Variant	CLIP-T ↑
LoRA-C (single pass, uniform)	0.2676
Uniform mask (M_i=$1/N$)	0.2681
Attention-level routing	0.2799
SR-LoRA τ=0.1 (noise-lvl)	**0.2858**

Uniform masking recovers LoRA-C performance (0.2681), confirming the multi-pass infrastructure adds no benefit by itself. The gap between attention-level routing and SR-LoRA (0.0059 CLIP-T, +2.1%) demonstrates that routing *noise predictions* based on attention maps is strictly more expressive than routing attention outputs directly: the former separates concept-specific spatial frequency signals that cross-attention blending cannot resolve.

I N=3 Triplet Composition

Table 8 extends SR-LoRA to three simultaneous LoRAs on character+clothing+object triplet compositions (8 combos × 3 seeds = 24 images per method).

Table 8: N=3 triplet composition results (8 combos × 3 seeds = 24 images each). SR-LoRA scales cleanly to three adapters with no extra tuning. FreeFuse-N3 uses argmax spatial routing over three per-adapter attention maps.

Method	CLIP-T ↑	CLIP-I ↑
LoRA-C	0.3290	1.0000
CLAC-N3	0.3117	0.7701
FreeFuse-N3	0.3001	0.8045
SR-LoRA-N3	**0.3442**	**0.8238**

SR-LoRA-N3 improves CLIP-T to 0.3442 (+4.6% over LoRA-C) while also maintaining CLIP-I = 0.8238 (vs. 0.7701 for CLAC-N3, which runs one anchor-secondary pair among the three). The routing masks M_1, M_2, M_3 (summing to 1 at each pixel) cleanly partition the character, clothing, and object regions at τ=0.1, confirming that SR-LoRA's per-pixel softmax extends naturally to $N > 2$ with no additional hyper-parameter tuning. FreeFuse-N3 (argmax over three per-adapter noise magnitudes) achieves CLIP-T = 0.3001 and CLIP-I = 0.8045, below both SR-LoRA-N3 on CLIP-T (−12.8%) and CLIP-I (−2.4%), confirming that hard argmax routing degrades further as N increases and more signal is discarded at each pixel.