IARouter: Identity–Attribute Routing for Compositional Text-to-Image Personalization

Anonymous authors
Paper under double-blind review



Figure 1: **Identity–attribute routing and compositional range.** Columns 3–5 compare direct Nested Attention (M3), our identity-only baseline B1 (M4), and IARouter (M6); the corresponding attribute prompts appear below each image group.

Abstract

Multi-reference personalization must preserve each identity and assign every requested attribute to the correct person. Shared attention often mixes these assignments. We introduce IARouter, a training-free method that routes reference identity information to the corresponding heads and textual attributes to the corresponding body regions. First, a reference-free pass locates the people and freezes their head and body ownership maps. Personalized generation first establishes a text-conditioned layout and then activates the two routes. The attribute route redistributes the existing attribute-attention budget across the person-specific body regions. Ordinary text conditioning, the frozen backbone, and self-attention remain unchanged. On our two-subject benchmark ($N = 144$ images per method), our variants achieve the best value in every reported metric. This includes the highest target-versus-decoy binding (0.663) and valid-face rate (0.997) among the evaluated baselines. Paired ablations show that attribute routing improves both binding and complete prompt response. These results support identity–attribute routing for improving two-subject compositional personalization.

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.

1 Introduction

Subject-driven personalization places referenced people in new scenes while following a text prompt. Encoder-based methods preserve identity without per-subject optimization (Xiao et al., 2023; Ye et al., 2023; Li et al., 2024; Wang et al., 2024; Patashnik et al., 2025), but multiple references introduce competing assignments. For "`<person1>` wears a striped jacket and `<person2>` wears a polka-dot jacket," the model must assign both the reference and clothing phrase to the correct person. It may preserve the faces while swapping their clothing. The numbered placeholders are the literal identity tokens used in our implementation.

Figure 1 uses the same ordered references, prompts, and seeds across methods and spans seven clothing, material, and scene settings; the supplementary material provides the exact prompts and seeds. This failure combines two routing problems. *Identity routing* sends each reference representation to one spatial region. *Attribute routing* sends each clothing phrase to the same region. A shared cross-attention map does not guarantee that these two assignments agree. Simply increasing reference strength is insufficient: early identity injection can change the number and layout of people, while strong global injection can reduce prompt editability.

We introduce IARouter, an inference-time router using frozen SDXL (Podell et al., 2024) and Nested Attention reference values (Patashnik et al., 2025). A reference-free pass localizes people and freezes their soft head and body ownership maps. Personalized generation restarts from the same noise, runs five text-only steps, then routes reference values to the corresponding heads and attribute attention to the corresponding bodies. This separates identity and clothing control while conserving the attribute budget up to numerical tolerance. Neither the backbone nor the reference encoder is trained, and the full U-Net feature map is not gated.

Our contributions are:

- We formulate compositional multi-subject personalization as two coupled assignment problems: routing each reference identity and its textual attributes to the same generated person.
- We introduce IARouter, a training-free router that combines delayed head-local identity injection with budget-conserving body-local attribute routing while keeping the backbone and reference encoder frozen.
- We demonstrate consistent gains over the compared baselines in paired two-subject experiments. Ablations isolate the effect of each route, while qualitative tests characterize broader compositional behavior and remaining failure modes.

2 Related Work

Personalized text-to-image generation. Optimization-based methods learn a token or update model parameters for each new concept (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023; Gu et al., 2023). Encoder-based methods instead extract subject features in one forward pass (Xiao et al., 2023; Ye et al., 2023; Li et al., 2024; Wang et al., 2024). FastComposer localizes subject conditioning and delays its injection during denoising (Xiao et al., 2023). MuDI uses segmented subject priors to separate multiple identities (Jang et al., 2024). We keep a frozen query-dependent reference encoder and focus on assigning each identity and its body-local text attribute to the same person.

Query-dependent reference values. Nested Attention uses the current U-Net query to read a value from reference-image tokens instead of using one static subject vector (Patashnik et al., 2025). This representation is expressive for one identity. For multiple identities, however, the original formulation does not specify which generated person should receive each independent reference or its nearby textual attributes. IARouter keeps the original encoder and value construction and changes only where and when these values are applied.

Attention control and semantic leakage. Cross-attention control can improve object presence and editing without retraining (Hertz et al., 2023; Chefer et al., 2023). DeLeaker dynamically reweights attention to reduce semantic leakage (Ventura et al., 2025). IARouter instead targets personalized composition. It uses matched ownership signals to route identity values and attribute phrases, and it preserves the total attention budget assigned to those phrases.

3 Method

3.1 Problem Setup

The input consists of a prompt P and R independently referenced people $\{I_r\}_{r=1}^{R}$. The prompt contains numbered identity tokens such as `<person1>` and `<person2>`. Before denoising, token r is assigned to reference I_r and person slot r. We denote this fixed binding by $\Pi : r \mapsto \big(I_r, \mathcal{A}_r(P)\big)$, where $\mathcal{A}_r(P)$ denotes the body-local token occurrences attached to token r. We use s_r^{id} for the sequence position of that token in the tokenized prompt. The goal is to place reference r in slot r and assign it the body-local phrase attached to the same token. Both the diffusion model and the Nested Attention reference encoder remain frozen. We validate $R = 2$ quantitatively and provide a separate qualitative check for $R = 3$ in Figure 6.

At denoising step n, U-Net layer ℓ, and attention head h, let $Q_q \in \mathbb{R}^{d_h}$ be the query at spatial position q. Standard text cross-attention is

$$z_{q,s} = \frac{Q_q (K_s^{\mathrm{text}})^\top}{\sqrt{d_h}}, \qquad a_{q,s} = \mathrm{softmax}_s(z_{q,s}), \qquad \phi_q^{\mathrm{base}} = \sum_s a_{q,s} V_s^{\mathrm{text}}. \tag{1}$$

Here s indexes token occurrences, including separate indices for repeated words; text denotes the text branch and $\top$ the transpose. Keys and values are d_h-dimensional. The base output ϕ_q^{base} sums text-token contributions and is distinct from the prefix latent $z^{(5)}$ and reference features. We omit n, ℓ, h where clear.

3.2 Frozen Reference Encoding and Nested Values

We align and resize each reference to 512×512, then process it with the frozen CLIP image processor, vision encoder, and Q-Former from Nested Attention. The non-CLS features produce M tokens $E_r = \{e_{r,m}\}_{m=1}^{M}$. At each U-Net cross-attention layer,

$$K_{r,m}^{\mathrm{ref}} = e_{r,m} W_K^{\mathrm{ref}}, \qquad V_{r,m}^{\mathrm{ref}} = e_{r,m} W_V^{\mathrm{ref}}. \tag{2}$$

The superscript ref labels the reference-image branch. The current image query reads each reference independently:

$$b_{q,r,m} = \mathrm{softmax}_m\left(\frac{Q_q (K_{r,m}^{\mathrm{ref}})^\top}{\sqrt{d_h}}\right), \qquad v_{q,r}^{*} = \sum_{m=1}^{M} b_{q,r,m} V_{r,m}^{\mathrm{ref}}. \tag{3a}$$

We then match the norm of the retrieved Value to that of the corresponding identity-token Value:

$$\widehat{v}_{q,r}^{*} = \frac{\left\| V_{s_r^{\mathrm{id}}}^{\mathrm{text}} \right\|_2}{\left\| v_{q,r}^{*} \right\|_2 + \epsilon} v_{q,r}^{*}. \tag{3b}$$

In Equation 3b, $V_{s_r^{\mathrm{id}}}^{\mathrm{text}} \in \mathbb{R}^{d_h}$ is the Value vector of identity token r. Its ℓ_2 norm provides the target magnitude for the retrieved reference vector $v_{q,r}^{*}$.

3.3 Anonymous Ownership from a Reference-Free Localization Pass

We compute ownership with a separate 30-step reference-free localization pass. This pass uses the same prompt, seed, and initial noise as personalized generation, but disables identity injection and attribute routing. At fixed checkpoints, we record cross-attention for

generic person tokens and predicted-x_0 images. After the pass, a frozen detector and segmenter locate candidate people. We associate candidates across checkpoints, retain the stable primary people, order them deterministically, and freeze their head and body masks. No reference image is used during localization, and the localization output is not reused as the personalized output.

The personalized trajectory then restarts from the same initial noise. Its first five denoising steps are run without reference injection or attribute routing:

$$z_T \equiv z^{(0)} \xrightarrow[\text{routing off}]{n=0,\dots,4} z^{(5)}. \tag{4}$$

Identical prompts, noise, seeds, schedulers, and routing-off computation reproduce the localization pass's first five updates, establishing a text-only layout. Projecting the frozen masks to each cross-attention resolution gives head ownership $u^{\text{head}}_{q,r} \in [0,1]$ and body ownership $u^{\text{body}}_{q,r} \in [0,1]$. Values approach 1 inside a region and 0 outside, with soft boundaries. These reference-independent maps control identity and body-local attributes, respectively.

For attributes, define

$$\begin{aligned} m^{\text{body}}_q &= \text{clip}\left(\sum_{r=1}^{R} u^{\text{body}}_{q,r}, 0, 1\right), \\ \rho^{\text{body}}_{q,r} &= \frac{u^{\text{body}}_{q,r}}{\sum_{j=1}^{R} u^{\text{body}}_{q,j} + \epsilon_\rho}. \end{aligned} \tag{5}$$

The support term m^{body}_q indicates whether q belongs to any primary person, while $\rho^{\text{body}}_{q,r}$ gives the relative ownership of person r. These terms control only the added conditioning; they do not multiply the latent, noise, or full U-Net features. If localization cannot identify a reliable person assignment, both routes remain off.

3.4 Delayed Identity Routing

The identity increment contributed by reference r is

$$\Delta\phi^{\text{id}}_{q,r} = a_{q,s^{\text{id}}_r}\left(\widehat{v}^{*}_{q,r} - V^{\text{text}}_{s^{\text{id}}_r}\right). \tag{6}$$

We open identity injection at step 5:

$$\gamma_{\text{id}}(n) = \mathbb{1}[n \geq 5], \qquad \phi^{\text{id}}_q = \phi^{\text{base}}_q + \gamma_{\text{id}}(n) \sum_{r=1}^{R} u^{\text{head}}_{q,r} \Delta\phi^{\text{id}}_{q,r}. \tag{7}$$

The indicator $\mathbb{1}[\cdot]$ is 1 when its condition is true and 0 otherwise. Therefore, $\gamma_{\text{id}}(n)$ keeps identity injection off for steps 0–4 and turns it on from step 5. Equation 7 keeps the base output and adds the reference residual only inside the corresponding head region. This schedule lets text establish the layout before identities are written. Self-attention remains unchanged.

3.5 Budget-Conserving Body-Local Attribute Routing

Let $\mathcal{A}_r(P)$ contain the token occurrences in the complete body-local phrase attached to person r. For example, "`<person1>` wearing a striped denim jacket" adds *wearing, striped, denim, jacket* to $\mathcal{A}_1(P)$. Repeated words remain separate when they occur in phrases for different people. We exclude ambiguous scene words and external objects.

At query q, the base model has already assigned a total attention budget

$$B^{\text{attr}}_q = \sum_{r=1}^{R} \sum_{s \in \mathcal{A}_r(P)} a_{q,s}. \tag{8}$$

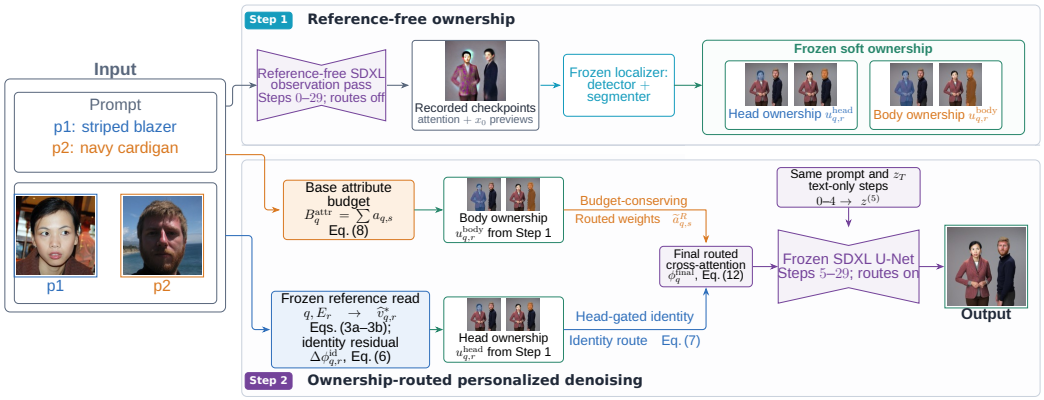

Figure 2: **Method overview.** Step 1 runs a complete 30-step reference-free observation pass, localizes the primary people from recorded checkpoints, and freezes their soft head and body ownership maps. Step 2 restarts from the same noise, repeats text-only steps 0–4 to recover $z^{(5)}$, and uses the frozen ownership maps for identity and attribute routing from step 5.

Within each phrase, the base logits define

$$\eta_{q,s|r} = \frac{\exp z_{q,s}}{\sum_{v\in\mathcal{A}_r(P)} \exp z_{q,v} + \epsilon_\eta}, \qquad \sum_{s\in\mathcal{A}_r(P)} \eta_{q,s|r} \le 1. \tag{9}$$

Here $\epsilon_\rho > 0$ and $\epsilon_\eta \ge 0$ denote denominator stabilizers. Our implementation uses $\epsilon_\rho = 10^{-8}$ and evaluates η with stable softmax, corresponding to $\epsilon_\eta = 0$. Starting at $n = 5$, attribute occurrences are rerouted as

$$\widetilde{a}^R_{q,s} = \left(1 - \lambda(n) m^{\text{body}}_q\right) a_{q,s} + \lambda(n) m^{\text{body}}_q B^{\text{attr}}_q \rho^{\text{body}}_{q,r} \eta_{q,s|r}, \quad s \in \mathcal{A}_r(P), \tag{10}$$

where $\lambda(n) = \mathbb{1}[n \ge 5]$ by default. Tokens outside the selected attribute phrases keep their original weights: $\widetilde{a}^R_{q,s} = a_{q,s}$ for $s \notin \bigcup_r \mathcal{A}_r(P)$. The router does not apply another token-wise softmax. With the stated stabilizer settings, the attribute budget is conserved up to numerical tolerance:

$$\sum_{r=1}^{R} \sum_{s\in\mathcal{A}_r(P)} \widetilde{a}^R_{q,s} \approx B^{\text{attr}}_q. \tag{11}$$

Thus, routing redistributes the attribute budget while approximately preserving its total. The stabilizers can introduce a small normalization deficit; we check Equation 11 with an absolute error tolerance of 10^{-6}. The final cross-attention output combines ordinary text, routed body-local attributes, and head-local identity information:

$$\begin{aligned} \phi^{\text{final}}_q = &\sum_{s\in\mathcal{S}_{\text{other}}} a_{q,s} V^{\text{text}}_s + \sum_{r=1}^{R} \sum_{s\in\mathcal{A}_r(P)} \widetilde{a}^R_{q,s} V^{\text{text}}_s \\ &+ \sum_{r=1}^{R} a_{q,s^{\text{id}}_r} \left[V^{\text{text}}_{s^{\text{id}}_r} + \gamma_{\text{id}}(n) u^{\text{head}}_{q,r} \left(\widehat{v}^*_{q,r} - V^{\text{text}}_{s^{\text{id}}_r} \right) \right]. \end{aligned} \tag{12}$$

Here $\mathcal{S}_{\text{other}}$ contains all tokens that are neither attributes nor identities. The three terms preserve ordinary text conditioning, route the existing body-local attribute budget, and write the query-dependent reference Value inside the corresponding head region. The unconditional classifier-free-guidance branch and all U-Net self-attention layers remain unchanged. Figure 2 summarizes the complete flow.

4 Experiments

4.1 Questions, Baselines, and Protocol

We ask three questions. (Q1) Does IARouter improve subject–attribute binding? (Q2) Does it preserve identity, structure, and prompt fidelity? (Q3) Which failures remain in difficult two-subject cases?

Backbone and references. We draw identity references from FFHQ-Wild-500 and process them with the official Nested Attention `align_face`$\rightarrow$ 512^2 $\rightarrow$CLIP pipeline. The references provide identity only; their original clothing is not used to choose prompt attributes. Twelve held-out images form six disjoint pairs. E1 evaluates M0–M4 and M6 over six pairs, 12 prompts, and four seeds, for 1,728 images. E2 compares M4–M6 on a focused subset of three pairs, six prompts, and two seeds, for 108 images; matching M4 and M6 cases are shared with E1. For the main comparison on prompts with distinct categories or materials (D) and stylized rendering (Y), every method is evaluated on the same six pairs, six prompts, and four seeds, giving 144 images per method. B2 uses this same 144-condition grid.

Compared methods. We compare text-only SDXL (M0), FastComposer (M1), MuDI (M2), direct multi-reference Nested Attention (M3), the delayed identity-only baseline B1 (M4), the reference-from-step-0 variant B2 (M5), and IARouter R (M6). Every method uses the same 144 D/Y conditions. B1 and M6 share $z^{(5)}$ and differ only in attribute routing. B2 instead enables reference control from step 0. FastComposer uses its released checkpoint. MuDI uses its official implementation and one 2,010-step LoRA for each fixed identity pair. Within each comparison, the prompts, reference order, and seeds are fixed across methods.

Metrics. We use ArcFace to evaluate each detected face (Deng et al., 2019). The identity metrics include similarity to the correct and incorrect references, their margin, and top-1 assignment. For D/Y prompts, **Bind** measures how often the target phrase scores above the same-image decoy phrase for each person. **PromptResponse** (PR) requires this ordering to be correct for every target person in an image. We measure structure with valid-face rate, extra-person rate, person-mask IoU, and router blocking. LPIPS measures change within paired images (Zhang et al., 2018). The fixed presence threshold is -1, so CrossMix and shared-attribute presence are constant and should not be interpreted as clothing-recognition accuracy.

Statistical protocol. Before generation, we fix the seeds, references, prompts, parsers, thresholds, and exclusion rules. Every generated condition is retained, including failed seeds, and invalid detections remain in the denominator. For Bind and PR, we report exact success counts and paired 95% bootstrap confidence intervals clustered by identity tuple (10,000 draws; seed 20260813). The supplementary material gives the execution matrix, reference order, exact prompts, seeds, and metric definitions.

4.2 Two-Subject Main Results

Table 1 reports 144 D/Y images for each method. B2 achieves the highest Bind (0.663) and PR (0.410), B1 achieves the highest identity top-1 accuracy (0.847), and M6 achieves the highest valid-face rate (0.997). Thus, our three variants contain the best result in every reported column. Starting reference control at step 0 improves attribute ranking but lowers the valid-face rate to 0.969. In the controlled B1/M6 comparison, Bind increases from 155/288 to 181/288 decisions, a gain of 0.0903 (95% cluster-bootstrap CI $[0.0729, 0.1076]$). PR increases from 44/144 to 51/144 images, a gain of 0.0486 (95% CI $[0.0139, 0.0833]$). Bind improves for all six identity pairs. On the same D/Y subset, identity top-1 accuracy decreases by 1.04 percentage points.

Figure 3 compares methods with matched prompts, reference order, and seeds across papercraft in Santorini, a realistic patterned-clothing portrait, autumn fashion, and a studio portrait. In rows 2–3, M6 corrects the misplaced plaid shirt and brown coat seen with

Table 1: Two-subject D/Y comparison ($N = 144$ images per method). Identity is correct-reference top-1, and M0 has no reference identity. Bold indicates the best result, and underlining indicates the second-best result. Tied second-best results are all underlined.

Method	Bind↑	PR↑	ID Top-1↑	FaceValid↑
Text-only SDXL (M0)	.479	.250	–	.917
FastComposer (M1)	.486	.139	.462	.903
MuDI (M2)	.465	.181	.559	.986
Nested Attention direct (M3)	.559	<u>.354</u>	.542	.941
B1: Identity-only baseline (M4)	.538	.306	**.847**	<u>.990</u>
B2: Reference from step 0 (M5)	**.663**	**.410**	<u>.840</u>	.969
R: + attribute router (M6)	<u>.628</u>	<u>.354</u>	.837	**.997**

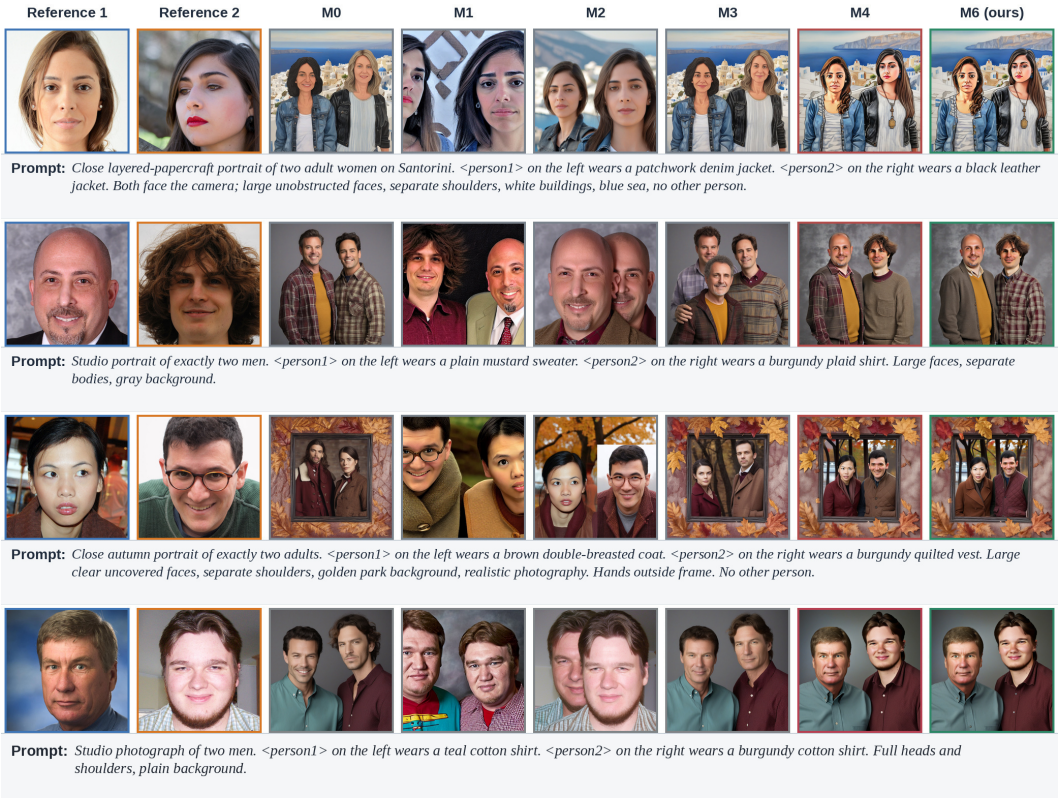

Figure 3: **Two-subject qualitative comparison.** Ordered references and outputs from six methods under matched prompts and seeds.

B1 (M4), while retaining recognizable identities in the requested order. Row 3 contrasts a brown coat with a burgundy quilted vest; row 4 uses teal and burgundy cotton shirts at seed 42. These examples illustrate the contribution of attribute routing; exact prompts appear below each row.

Figure 4 shows twelve settings across three disjoint reference pairs. Navy/orange raincoats and tan-trench/burgundy-quilted clothing remain visually distinct in the rainy-street and subway examples. Comic-book and watercolor outputs also retain recognizable identities and contrasting clothing colors, illustrating compositional range across scenes and styles. Supplementary Table 6 lists the exact prompts and seeds.



Figure 4: **Generalization across scenes, styles, attributes, and three disjoint reference pairs.** Each panel is generated independently. The text below each image gives the scene, style, and ordered P1/P2 clothing phrases.



Figure 5: **Scene-driven occupational homogenization.** Matched studio and fire-station outputs; checks and crosses indicate correct and incorrect clothing assignments.

4.3 Two-Subject Failure Analysis

Nested Attention identifies self-attention as a source of same-domain subject leakage (Patashnik et al., 2025). IARouter changes cross-attention but leaves self-attention unchanged, allowing information transfer that can homogenize clothing under strong scene priors.

Figure 5 compares neutral studio and fire-station scenes while fixing the ordered references, occupation phrases, seed, initial noise, framing, and identity constraints. Output 1 shows both requested coats in the studio; adding a drill, fire engines, and rescue equipment in Output 2 shifts the medical coat toward firefighter clothing while retaining recognizable identities and hair colors. Green checks and red crosses mark matched and mismatched clothing. We repeat this intervention across eight fixed seeds; the figure shows one example.

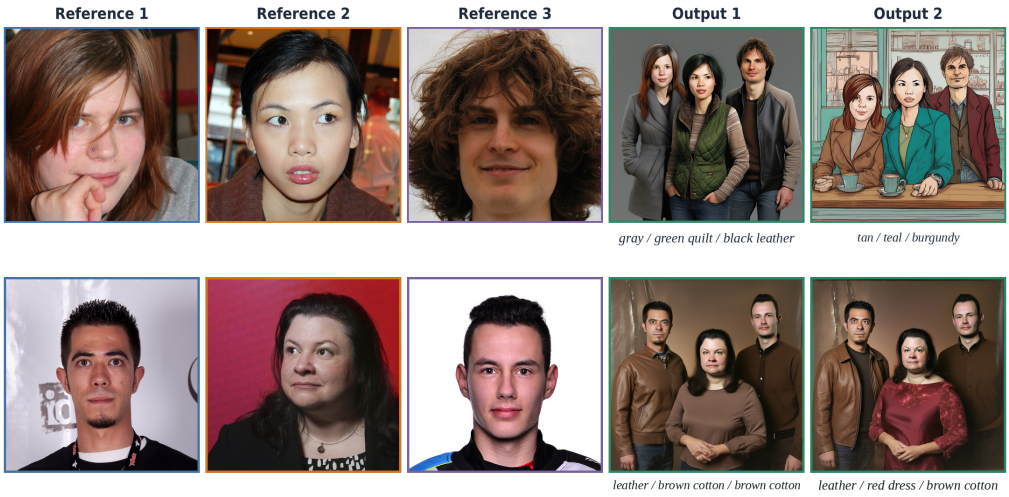

Figure 6: **Three-subject qualitative results.** Ordered reference triples and outputs with varying color, material, and clothing assignments.

4.4 Qualitative Three-Subject Extension and Scope

The index R permits three slots without retraining, although more subjects compete for spatial support and conditioning. Figure 6 shows two ordered reference triples with independently generated outputs; labels give clothing assignments from left to right. Row 1 varies color and material. Row 2 keeps a leather jacket on the left and a brown cotton shirt on the right, while changing the center subject from a brown cotton shirt to a red dress. We inspect person count, identity order, clothing, faces, and anatomy.

These and the supplementary three-subject examples are qualitative and excluded from Table 1. Quantitative evidence remains restricted to $R = 2$: E1 tests prompt and identity variation, and E2 compares routing variants under matched conditions. With a nonbinding attribute-presence threshold, Bind measures target-versus-decoy ranking rather than absolute clothing recognition.

5 Discussion, Limitations, and Conclusion

IARouter provides a training-free way to preserve individual identities and assign clothing attributes to the intended people. It separates identity and attribute control, directing reference information to heads and clothing descriptions to bodies while keeping the pretrained model frozen. This design addresses a central challenge in multi-subject generation: preserving a person's appearance does not by itself ensure that the person receives the requested attributes.

Two-subject experiments show that attribute routing improves binding and complete prompt response over the identity-only baseline. Qualitative results further illustrate its use across scenes, clothing combinations, and rendering styles. Together, these findings support explicit spatial routing as a practical approach to compositional personalization without additional training.

Reliable person localization remains essential, and strong scene cues can still cause clothing to mix. Fine clothing details and interactions with objects also remain challenging. Future work will explore finer spatial control and broader evaluation, including human judgments and quantitative tests with more subjects.

Reproducibility Statement

The main paper specifies the method, routing schedule, evaluation denominators, and result counts. The supplementary material provides the execution matrix, reference order, exact prompts and seeds, metric definitions, statistical tests, failure handling, and parameter sensitivity. We retain the configuration files and row-level results needed to reproduce the reported tables.

Ethics Statement

Personalized generation can be misused for impersonation, harassment, or non-consensual synthetic media. References should be used with appropriate licenses or consent, and synthetic outputs should retain provenance and clear labels. A public release should include safeguards against public-figure misuse and document demographic limitations. We evaluate compositional fidelity, not the inference of identity or personal traits.

Generative AI Disclosure

We used generative AI tools to refine hypotheses and framing, search for and identify related literature, review the methodology and experiment design, assist with deriving and writing mathematical proofs, assist implementation and debugging, assist with prompt refinement, experiment execution, and inspection of generated outputs, support qualitative analysis and translation, revise figures, suggest parameters, and assist with drafting and revising parts of the manuscript. Survey design, interviews, and transcription do not apply to this work. All reported images were produced by the named models with the settings described in the paper. The authors checked the implementation, equations, numerical results, citations, text, and figures and take responsibility for the complete paper and its artifacts.

References

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. In *ACM SIGGRAPH*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

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. In *ICLR*, 2023.

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. In *NeurIPS*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Nested attention: Semantic-aware attention values for concept personalization. *arXiv preprint arXiv:2501.01407*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

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 *CVPR*, 2023.

Mor Ventura, Michael Toker, Or Patashnik, Yonatan Belinkov, and Roi Reichart. DeLeaker: Dynamic inference-time reweighting for semantic leakage mitigation in text-to-image models. *arXiv preprint arXiv:2510.15015*, 2025.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. In *NeurIPS*, 2023.

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.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Supplementary Material

A Supplementary Overview

This appendix describes inference and person localization, followed by the experimental design, evaluation rules, and parameter sensitivity. We do not repeat the main-paper qualitative panels. Exact prompts, seeds, reference order, and evaluation settings are provided below for reproducibility.

B Unified Formulation and Inference Algorithm

Let q index a spatial query and $r \in \{1, \dots, R\}$ index an anonymous person slot. The input prompt uses numbered tokens such as `<person1>` and `<person2>`. Token r, reference I_r, and slot r share the same fixed order. Let $\mathcal{A}_r(P)$ contain the complete body-local phrase attached to slot r, and let (K_r^e, V_r^e) be its frozen reference-token bank. Each query reads each bank independently:

$$v_{q,r}^* = \mathrm{softmax}_m\left(\frac{q(K_r^e)^\top}{\sqrt{d}}\right) V_r^e, \qquad \widehat{v}_{q,r}^* = \frac{\|V_{s_r^{\mathrm{id}}}^{\mathrm{text}}\|_2}{\|v_{q,r}^*\|_2 + \epsilon} v_{q,r}^*. \tag{13}$$

The identity residual is $\delta v_{q,r} = \widehat{v}_{q,r}^* - V_{s_r^{\mathrm{id}}}^{\mathrm{text}}$. After the shared prefix, head ownership $u_{q,r}^{\mathrm{head}}$ restricts this residual to the corresponding person:

$$\Delta\phi_q^{\mathrm{id}} = \gamma_{\mathrm{id}}(n) \sum_{r=1}^{R} u_{q,r}^{\mathrm{head}} a_{q,s_r^{\mathrm{id}}} \delta v_{q,r}, \quad \gamma_{\mathrm{id}}(n) = \mathbb{1}[n \geq 5]. \tag{14}$$

For body-local phrases, we first compute their existing attention budget and the normalized allocation within each phrase:

$$B_q^{\mathrm{attr}} = \sum_{r=1}^{R} \sum_{s \in \mathcal{A}_r(P)} a_{q,s}, \qquad \eta_{q,s|r} = \frac{\exp z_{q,s}}{\sum_{v \in \mathcal{A}_r(P)} \exp z_{q,v} + \epsilon_\eta}. \tag{15}$$

The support m_q^{body} and the stabilized ownership $\rho_{q,r}^{\mathrm{body}}$ are defined in Equation 5; zero ownership support disables attribute routing at that query. The denominator stabilizers satisfy $\epsilon_\rho > 0$ and $\epsilon_\eta \geq 0$. The router uses strength $\lambda(n)$ to compute

$$\widetilde{a}_{q,s}^R = \big(1 - \lambda(n) m_q^{\mathrm{body}}\big) a_{q,s} + \lambda(n) m_q^{\mathrm{body}} B_q^{\mathrm{attr}} \rho_{q,r}^{\mathrm{body}} \eta_{q,s|r}, \quad s \in \mathcal{A}_r(P). \tag{16}$$

The default schedule is $\lambda(n) = \mathbb{1}[n \geq 5]$; Section F varies its strength and onset. With the stabilizer settings specified below, routing preserves the total phrase budget up to numerical tolerance:

$$\sum_{r=1}^{R} \sum_{s \in \mathcal{A}_r(P)} \widetilde{a}_{q,s}^R \approx B_q^{\mathrm{attr}}. \tag{17}$$

Equation 12 combines the three paths: unchanged ordinary text, routed body-local attributes, and head-local query-dependent reference Values. The unconditional classifier-free-guidance branch and all U-Net self-attention layers remain unchanged. The general index R defines the operator; our quantitative validation covers $R = 2$. Algorithm 1 summarizes inference: localize people without references, freeze their ownership maps, then restart from the same noise and activate both routes after the shared five-step prefix.

Algorithm 1: IARouter inference

Input: prompt P; references $\{I_r\}_{r=1}^{R}$; initial noise z_T; frozen SDXL and reference encoder.
1. Parse identity occurrences s_r^{id} and body-local phrase sets $\mathcal{A}_r(P)$.
2. Preprocess and encode every reference independently; cache its reference-token bank.
3. Starting from z_T, run a 30-step reference-free localization pass. Keep both routes off and record generic-person attention and predicted-x_0 images at the fixed checkpoints.
4. Run the frozen detector and segmenter on these records. Associate people across checkpoints, construct $u_{q,r}^{\mathrm{head}}$ and $u_{q,r}^{\mathrm{body}}$, and fix the slot order. If this assignment is unreliable, keep both routes off.
5. Restart from the same z_T. Run steps 0:4 with both routes off to obtain $z^{(5)}$.
6. During steps 5:29, compute standard text attention, read each reference, and apply Equations 14 and 16.
7. Complete the same scheduler trajectory and decode the final latent once.
Output: personalized image.

The implementation follows several constraints. Disabling attribute routing reduces the method to the identity-only baseline B1. Identity and attribute residuals are zero during steps 0–4. The attribute router changes neither identity-token weights nor tokens outside the selected attribute phrases, and it leaves the unconditional branch unchanged. The implementation uses $\epsilon_\rho = 10^{-8}$ in the ownership denominator and computes within-phrase weights with stable softmax, corresponding to $\epsilon_\eta = 0$. The stabilized normalization need not sum exactly to one; we check Equation 17 with an absolute error tolerance of 10^{-6}. References, person slots, and ownership masks use the same order.

C Person Localization and Ownership

We compute ownership with a separate reference-free diffusion pass rather than the five-step personalized prefix alone. This pass uses the same prompt, seed, and initial noise as personalized generation, but keeps identity injection and attribute routing off for all 30 steps. At fixed checkpoints, it records cross-attention for generic person words and predicted-x_0 images.

After the pass, a frozen open-vocabulary detector proposes person boxes (Liu et al., 2024), and a frozen segmentation model produces their masks (Kirillov et al., 2023). Face visibility, head location, and pose help determine whether a person can receive identity information reliably. Generic-person cross-attention provides additional localization evidence but does not define the target masks by itself. We associate detections across checkpoints and classify them as *core*, *auxiliary*, or *unresolved*. Only stable core people enter routing. We then freeze their ownership masks and restart personalized generation from the same initial noise.

Head masks control identity residuals, while body masks control body-local phrases. We resize both soft masks to each U-Net resolution with area-preserving interpolation. Before reading any reference, we order the person slots by the horizontal coordinate of each body-mask centroid and resolve ties deterministically. Small or distant detections may remain auxiliary and receive no identity injection. If the number or order of primary people is uncertain, both routes remain off.

Figure 7 illustrates how reference information and spatial ownership interact. Panel (a) reads each reference independently, matches the retrieved Value's norm, and forms an identity residual. Panel (b) uses head ownership to gate that residual and body ownership to redistribute attribute attention, then combines both paths with ordinary text conditioning. The references match the pair-4, seed-647319357 M6 example in Figure 2. The masks come from its reference-free localization pass, using the same prompt, seed, and initial noise without reference images.

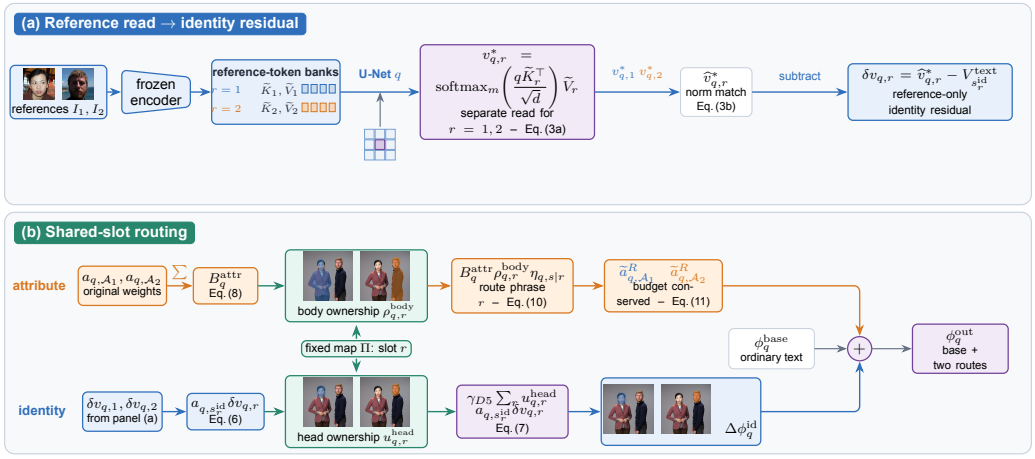

Figure 7: **Reference reading and shared-slot routing.** (a) Identity residual construction. (b) Ownership-guided identity and attribute routing.

D Experimental Protocol

D.1 Fixed Inputs and Settings

Before generation, we fix four groups of information. The reference specification gives the identity IDs, preprocessing procedure, and data split. The prompt specification gives the exact text, identity-token positions, attribute phrases, template family, and target person count. The generation specification gives the method, model version, scheduler, resolution, number of steps, CFG, seed, and reference order. The evaluation specification gives the detector, segmenter, face recognizer, CLIP model, thresholds, crop rules, metrics, and denominator policy. Table 2 summarizes the reference groups, prompts, seeds, methods, and image counts for each experiment.

D.2 Experiment Grid

Table 2: Experiment grid. Within each row, we evaluate every combination of the listed factors. E2 shares matching E1 conditions for M4 and M6. The full-grid evaluation of B2 tests reference control from step 0 on the same 144 D/Y conditions as Table 1.

Stage	Reference groups	Prompts	Seeds and methods	Images
E1	six pairs in Table 3	N01–N03, S01–S03, D01–D04, Y01–Y02 (12 prompts; Table 4)	four two-subject seeds; M0–M3, B1/M4, and M6	1,728
E2	pair groups 1, 3, 5	N01, S01, D01, D03, D04, Y01	seeds 2657029907, 2329732732; M4–M6	108
Full-grid evaluation of B2	all six pairs	D01, D02, D03, D04, Y01, Y02	all four two-subject seeds; B2/M5	144

E1 uses four seeds: `2657029907`, `2329732732`, `4008999317`, and `647319357`. E2 shares 72 matching M4/M6 conditions with E1 and adds 36 M5 conditions. The full-grid B2 evaluation shares 24 matching M5 conditions with E2 and adds 120 conditions, for a complete $6 \times 6 \times 4$ grid. We retain every fixed seed, including unsuccessful generations.

Within each condition, every method receives the same positive prompt, negative prompt, and seed. M1–M3, B1/M4, B2/M5, and M6 also receive the same ordered reference pair from Table 3, whose Person 1/2 entries match the identity-token and ownership-slot order; M0 receives no reference. B1 and M6 additionally share the same initial noise, first five scheduler updates, $z^{(5)}$, and frozen ownership maps. E2 uses only the three variants reported in the paper.

Table 3: Ordered two-subject reference pairs. The table order matches the identity-token and ownership-slot order. We form pairs from consecutive accepted FFHQ-Wild-500 candidates in ascending pool order and keep this order for all methods.

Group	Person 1	Person 2
Pair 1	ffhq_00082	ffhq_00083
Pair 2	ffhq_00084	ffhq_00089
Pair 3	ffhq_00092	ffhq_00093
Pair 4	ffhq_00096	ffhq_00097
Pair 5	ffhq_00099	ffhq_00100
Pair 6	ffhq_00101	ffhq_00102

All internal SDXL runs generate 1024×1024 images with 30 denoising steps and CFG 5.0. The default routing onset is step 5; the attribute-onset sweep is described in Section F. The default negative prompt is: *tiny face, distant person, profile, looking away, sunglasses, hat, mask, occlusion, blur, cropped head, extra person, crowd, duplicate, deformed face.* FastComposer follows its official configuration: 512^2, 50 steps, CFG 5.0, and $\alpha = 0.8$. MuDI uses 1024^2, 50 steps, CFG 5.0, $\gamma = 1$, and a 2,010-step rank-32 LoRA for each fixed pair. We record the code and checkpoint versions together with the prompt, seed, reference order, and generation settings for each condition.

D.3 Prompt and Attribute Coverage

We divide prompts into four fixed groups: N has no body-local contrast, S gives both people the same attribute, D assigns different categories or materials, and Y adds a stylized rendering. We treat carried objects as diagnostics rather than body-local evidence, and we do not rely on color alone. Table 4 gives the exact positive prompts for the quantitative tests and additional qualitative displays. It shows the literal identity tokens `<person1>` and `<person2>`; the default negative prompt appears above, and differences in negative prompts across qualitative examples are summarized before Table 5.

Table 4: Exact two-subject positive prompts used by E1 and by the displayed prompt-coverage matrices; E2 uses the six entries named in Table 2. E1 uses N01–N03, S01–S03, D01–D04, and Y01–Y02 (12 prompts); the remaining 29 entries are used only for qualitative display.

Prompt	Exact text
N01	a high-quality travel photograph of exactly two adults and no other person, <person1> standing on the left and <person2> standing on the right on the Great Wall of China at sunrise, separated bodies, clearly visible faces, sweeping mountain ridges, warm golden light, natural proportions, high detail
N02	a cinematic travel portrait of exactly two adults and no other person, <person1> on the left and <person2> on the right at the Trocadero with the Eiffel Tower clearly visible behind them during blue hour, separated bodies, unobstructed faces, elegant city lights, realistic photography, high detail
N03	a candid full-body photograph of exactly two adults and no other person, <person1> walking on the left and <person2> walking on the right through Central Park in autumn, separated silhouettes, clearly visible faces, orange and red foliage, soft afternoon sunlight, natural motion, realistic detail
N04	Close-up color portrait of two adults in front of library bookshelves. <person1> is the woman on the left. <person2> is the man on the right. Match faces and hair. Full heads, shoulders visible.
N05	Realistic portrait of one woman and one man against a plain terracotta wall. <person1> is the woman on the left. <person2> is the man on the right. Match each reference face and hair. Waist-up, space above heads, separate shoulders, soft daylight.
N06	Close head-and-shoulders lakeside portrait of exactly two adults. <person1>, an adult woman, is on the left; <person2>, an adult man, is on the right. Both face the camera; large unobstructed faces fill the frame, alpine lake softly behind them, realistic photography.

Table 4 (continued)

Prompt	Exact text
N07	Close photo of exactly one adult woman and one adult man at a quiet cafe terrace. <person1> is the woman on the left. <person2> is the man on the right. Large uncovered faces, separate shoulders, quiet cafe terrace behind them, realistic photography. Keep each reference identity and sex; do not swap them.
N08	Realistic portrait of one woman and one man in an empty tropical glasshouse. <person1> is the woman on the left. <person2> is the man on the right. Match each reference face and hair. Waist-up, space above heads, separate shoulders, soft daylight.
N09	Realistic portrait of exactly two adults in an empty tropical glasshouse. <person1> is on the left. <person2> is on the right. Match each reference face and hair. Chest-up, separate shoulders, soft daylight.
N10	Color photo of exactly two adults standing shoulder to shoulder in a library. <person1> left is a blonde woman. <person2> right is a bald man with a short beard. Match faces and hair. Close-up, equally large faces, complete heads.
S01	a heroic editorial photograph of exactly two firefighters and no other person outside a modern fire station, <person1> on the left wearing a navy firefighter turnout coat with wide reflective stripes and <person2> on the right wearing the same navy firefighter turnout coat with wide reflective stripes, separated bodies, clearly visible faces, fire engine in the background, realistic dramatic light
S02	a professional hospital portrait of exactly two doctors and no other person, <person1> on the left wearing a crisp white medical coat over teal scrubs and <person2> on the right wearing the same crisp white medical coat over teal scrubs, separated bodies, clearly visible faces, bright modern hospital atrium, clean realistic photography
S03	a documentary photograph of exactly two construction workers and no other person at a modern bridge worksite, <person1> on the left wearing a fluorescent orange safety vest over gray work coveralls and <person2> on the right wearing the same fluorescent orange safety vest over gray work coveralls, separated bodies, clearly visible faces, realistic daylight, high detail
S04	Color studio photo of two adults. <person1> left wears a navy firefighter coat. <person2> right wears a navy firefighter coat. Both coats have narrow silver reflective bands. Preserve each face, hair and sex. Full heads, chest-up, gray background.
S05	Realistic portrait of exactly two women outside a fire station. <person1> left wears a navy firefighter coat with yellow reflective bands. <person2> right wears the same navy firefighter coat with yellow reflective bands. Preserve both female faces and hair, large clear faces, separate shoulders, no helmets and no other person.
S06	Tight head-and-shoulders hospital portrait of exactly two adults. <person1> on the left wears a white medical coat. <person2> on the right wears the same white medical coat. Two large uncovered faces fill the frame, separate shoulders, bright clinical background, realistic photography. Identity and gender must match each corresponding reference.
S07	Chest-up realistic portrait on gray. <person1> left wears a solid-orange zip jacket. <person2> right wears the same solid-orange zip jacket. Preserve reference face, sex, age, hair color, and hairstyle. Uncovered heads, separate shoulders, waist and hands outside frame. No hats, headbands, belts, aprons, vests, gray panels, or others.
S08	Close realistic kitchen portrait. <person1> left wears a white double-breasted chef jacket. <person2> right wears the same white double-breasted chef jacket. No aprons. Keep reference identity, sex, age, and natural hair. Uncovered faces, separate shoulders, no hats, exactly two people.
S09	Close headshot of exactly two hikers. <person1> left wears a red hiking jacket. <person2> right wears a red hiking jacket. Large uncovered faces, separate shoulders, blurred forest, realistic photo. Keep identity and gender as references.
D01	a cinematic editorial photograph of exactly two professionals and no other person during a city emergency drill, <person1> on the left wearing a dark firefighter turnout coat with yellow reflective bands and <person2> on the right wearing a white medical coat over blue scrubs, separated bodies, clearly visible faces, emergency vehicles softly blurred in the background, realistic high detail

Table 4 (continued)

Prompt	Exact text
D02	a high-quality documentary portrait of exactly two public safety professionals and no other person in a spacious civic square, <person1> on the left wearing a navy police uniform jacket with silver buttons and <person2> on the right wearing a tan firefighter turnout coat with fluorescent bands, separated bodies, clearly visible faces, realistic daylight, high detail
D03	an elegant wedding photograph of exactly two adults and no other person at the Trocadero with the Eiffel Tower behind them, <person1> on the left wearing an embroidered ivory wedding gown and <person2> on the right wearing a tailored black tuxedo with satin lapels, separated full bodies, clearly visible faces, soft sunrise light, luxurious realistic detail
D04	a sophisticated autumn fashion portrait of exactly two adults and no other person in Central Park, <person1> on the left wearing a brown herringbone double-breasted coat and <person2> on the right wearing a burgundy diamond-quilted sleeveless vest, separated bodies, clearly visible faces, golden foliage, realistic editorial photography
D05	Portrait photo of two adults. On the right, <person2> wears a red puffer jacket. On the left, <person1> wears a white wool coat. Match faces and hair. Chest-up, full heads, separate shoulders, gray background.
D06	Studio photo of exactly two adults standing side by side. <person1> on the left wears a white blazer. <person2> on the right wears a black blazer. Match faces and hair. Large faces, complete heads, plain gray background.
D07	Close-up studio photograph of two adults. <person1> left wears a brown herringbone blazer. <person2> right wears a burgundy quilted jacket. Preserve each face, hair and sex. Full heads and shoulders, gray background.
D08	Realistic studio portrait. <person1> left wears a solid navy denim jacket, no red. <person2> right wears a solid scarlet jacket, no blue or green. Preserve each reference face, sex, and natural hair. Chest-up, uncovered faces, separate shoulders, gray background. No hats, patterns, medical clothing, or other people.
D09	Close realistic studio portrait of exactly two adults. <person1> on the left wears a red marching-band jacket with gold trim. <person2> on the right wears a white astronaut flight jacket with blue panels. Large uncovered faces, separate shoulders, plain light-gray background. Keep each reference identity, sex, and hair; no hats, helmets, or other people.
D10	Realistic studio portrait of exactly two adults. <person1> left wears a cable-knit sweater. <person2> right wears a leather jacket. Preserve each face, sex, and hair. Large clear faces, separate shoulders, neutral gray background, no other person.
D11	Color photo of exactly two people. <person1>, the woman on the left, wears a cobalt-blue jacket. <person2>, the man on the right, wears an orange jacket. Preserve faces and hair. Chest-up, full heads, gray studio background.
Y01	a delicate watercolor painting of exactly two adults and no other person in a Kyoto garden, <person1> on the left wearing a blue vertical-striped blazer and <person2> on the right wearing a red floral embroidered cardigan, separated bodies, recognizable faces, cherry blossoms, expressive watercolor washes, fine paper texture
Y02	an intricate layered papercraft scene of exactly two adults and no other person overlooking Santorini, <person1> on the left wearing a patchwork denim jacket and <person2> on the right wearing a black leather biker jacket, separated bodies, recognizable faces, white buildings and blue sea, handcrafted paper texture
Y03	Oil-painted portrait of two adults. On the right, <person2> wears a red puffer jacket. On the left, <person1> wears a white coat. Match faces and hair. Chest-up, full heads, separate shoulders, gray background.
Y04	Close product portrait of exactly two Pop figure-style vinyl characters standing side by side on a clean display pedestal. <person1> is on the left and <person2> is on the right. Both large faces and upper bodies fill the frame, two separate figures, no other figure, simple colorful backdrop, polished 3D product photography.
Y05	Product photo of exactly two full-body Pop vinyl figures standing far apart on a pink studio background. <person1> is the dark-haired woman on the left. <person2> is the brown-haired bearded man on the right. Preserve their faces, sex, hair, and his short beard. Each figure stands centered on its own separate black round display base; both bases are fully visible. No touching or extra figure.

Table 4 (continued)

Prompt	Exact text
Y06	Charcoal portrait of two adults. On the right, <person2> wears a black leather jacket. On the left, <person1> wears a white coat. Match each reference face, hairstyle and sex. Chest-up, full heads, separate shoulders, plain background.
Y07	Close retro-comic portrait of exactly two adults at a subway entrance. <person1> left wears a mustard jacket. <person2> right wears a teal jacket. Preserve each reference face, sex, and natural hair. Equal-size uncovered faces, separate shoulders, bold ink and halftone color.
Y08	Retro-comic portrait of two adults. <person1> left wears a mustard sweater. <person2> right wears a teal polo shirt. Preserve faces and natural hair. Complete heads, chest-up, white background, bold ink.
Y09	Retro-comic portrait. <person1>, a short-haired man on the left, wears a mustard sweater. <person2>, a long-haired woman on the right, wears a teal polo shirt. Two adults, chest-up, full heads, white background.
Y10	Retro-comic illustration of two adults. <person1> left is a man wearing a mustard shirt. <person2> right is a woman wearing a teal shirt. Match reference faces and hair. Chest-up, full heads, white background.
Y11	Retro-comic illustration of two adults. <person1> left is a woman wearing a mustard shirt. <person2> right is a brown-haired man wearing a teal shirt. Match reference faces and his short beard. Chest-up, full heads, white background.

D.3.1 Visual coverage of the four prompt strata

Figures 8, 9, 10, and 11 show no body-local contrast, shared clothing attributes, distinct categories or materials, and stylized contrasts, respectively. Each figure is a 4×7 matrix. The first two columns show the ordered references, and the remaining five show M6 outputs. Table 5 maps each output column to its prompt and seed, including row-specific wording. Negative prompts vary across qualitative examples, with differences in exclusions for rendering style, clothing, accessories, and anatomy. All four rows within each output column share the listed seed and the same negative prompt. The 32 aligned references are distinct and do not appear in earlier paper figures. Before generation, we checked that each reference has one clear, unobstructed face without eyewear. These 80 images provide qualitative coverage and do not enter Table 1 or any aggregate metric.

Table 5: Prompt and seed for each output column in the supplementary matrices. Each entry has the form `prompt ID / seed`; Table 4 gives the full prompt text used for generation.

Type	Column 1	Column 2	Column 3	Column 4	Column 5
N	N04/856101	N05/836503	N06/616103	N07/716401	N08/836502
S	S04/31416019	S06/616202	S07/181000019	S08/1016304	S09/686205
D	D05/31416019	D06/109000013‖	D07/856101	D08/106000007	D10/83105901
Y	Y02/4008999317	Y03/31416019	Y04/2828282	Y06/109000013	Y08--Y11/109000013°

All four rows in each output column share the seed shown above. ‖D Column 2 specifies two women in row B instead of two adults. °Y Column 5 uses Y10, Y09, Y08, and Y11 in rows A, B, C, and D, respectively.



Figure 8: **N: no body-local contrast.** Rows A–D use four new ordered reference pairs. The five output columns show a library, terracotta studio, lakeside, cafe terrace, and empty tropical glasshouse.



Figure 9: **S: shared body-local attribute.** Each output gives both people the same clothing phrase. The five columns request firefighter coats, medical coats, solid-orange jackets, chef uniforms, and red hiking jackets.



Figure 10: **D: distinct categories or materials.** The five columns test white wool coats versus red puffer jackets, white versus black blazers, brown herringbone blazers versus burgundy quilted jackets, navy denim versus scarlet clothing, and knit versus leather.

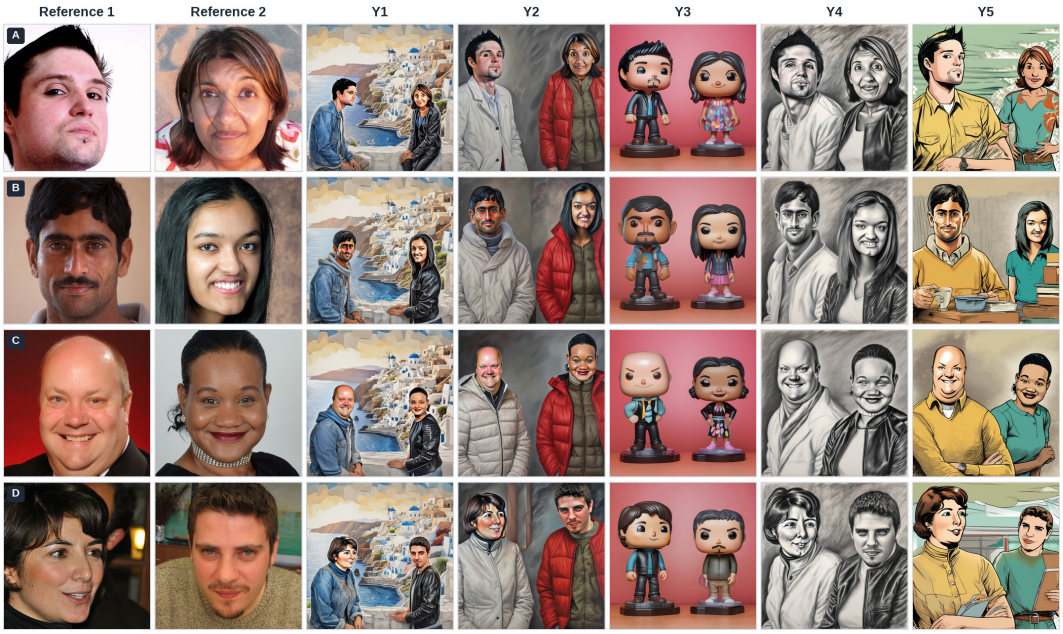

Figure 11: **Y: stylized contrast.** The columns use papercraft, oil painting, Pop-figure product photography, charcoal drawing, and retro comic rendering. All outputs prioritize large, unobstructed faces. Table 5 lists their inputs.

Figure 12 extends the qualitative analysis to seven disjoint reference triples, absent from the main text. Rows 1–3 vary color, rows 4–5 material, and rows 6–7 occupation. Outputs 1–3 assign one shared, two distinct, and three distinct attributes in cel-shaded, oil-painted, and photographic styles, respectively. Labels specify the left-to-right assignments; occupations may differ between rows. We visually screened recorded prompt, reference, and seed combinations for person count, identity, hair, gender presentation, clothing order, facial quality, and anatomy. These selected examples do not enter aggregate metrics.

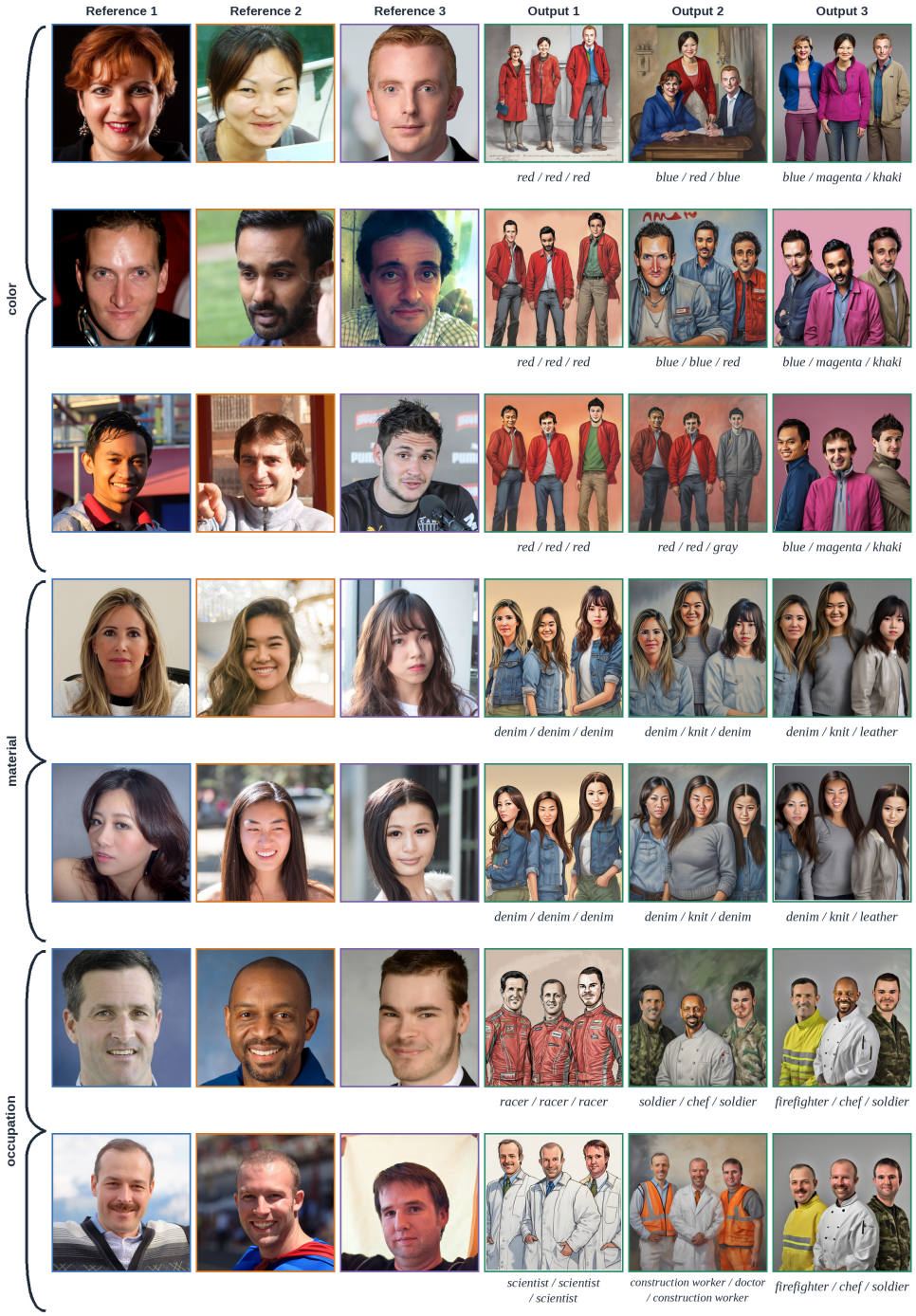

Figure 12: **Three-subject qualitative results.** Seven ordered reference triples with shared and distinct color, material, or occupation assignments.

D.4 Metrics and Denominators

For person r in image i, $S^+_{i,r}$ is the ArcFace similarity to the assigned reference, and $S^-_{i,r}$ is the largest similarity to another reference. We define the identity margin as $M_{i,r} = S^+_{i,r} - S^-_{i,r}$. The top-1 identity decision is correct when $M_{i,r} > 0$. A person without a valid face counts as an identity failure.

For D/Y outputs, $y_{i,r} = 1$ when the assigned phrase scores above the same-image decoy phrase for person r. We compute

$$\text{Bind} = \frac{1}{\sum_i R_i} \sum_i \sum_{r=1}^{R_i} y_{i,r}, \qquad \text{PromptResponse} = \frac{1}{N} \sum_i \prod_{r=1}^{R_i} y_{i,r}. \tag{18}$$

The fixed presence threshold is -1, so Bind measures relative target-versus-decoy ranking rather than absolute phrase presence. Invalid detections remain in the denominators. We estimate uncertainty with paired 95% bootstrap intervals clustered by identity tuple (10,000 draws; seed 20260813).

E Scene- and Attribute-Diverse Additional Reference Pairs

Main-paper Figure 4 uses three ordered reference pairs that do not overlap: `ffhq_00082/ffhq_00116` in row A, `ffhq_00097/ffhq_00114` in row B, and `ffhq_00099/ffhq_00119` in row C. The outputs vary scenes, clothing, and rendering styles, including realistic photography, Pointillism, comic art, and watercolor. A1 uses a beach boardwalk, A2 uses a botanical conservatory, A3 contrasts corduroy and leather in a realistic gallery, and B4 applies Pointillism to a carnival scene. Every output uses a distinct seed. Table 6 lists the exact prompts and seeds.

Table 6: Exact prompts and seeds for Figure 4.

Cell	Seed	Exact positive prompt
A1	32452867	Chest-up beach-boardwalk photo. `<person1>`, bald man left, wears a plain brown coat. `<person2>`, dark-haired woman right, wears a plain burgundy jacket. Wooden boardwalk and ocean behind. Large clear faces, natural anatomy, frame ends above the waist. Exactly two adults.
A2	610000103	Chest-up color photo of exactly two adults in a botanical conservatory. `<person1>`, bald man left, wears a forest-green wool coat. `<person2>`, dark-haired woman right, wears a sunflower-yellow jacket. Large uncovered faces, separate shoulders, glass roof and plants behind, hands out of frame. No other person.
A3	83149363	Realistic chest-up gallery portrait of exactly two adults. `<person1>` bald man left wears a mustard-yellow corduroy jacket. `<person2>` dark-haired woman right wears a dark-brown leather jacket. Clear faces, plain gray wall, hands outside frame, no scarves, no hoodies.
A4	300000047	Waist-up comic-book illustration. `<person1>`, bald man left, wears a blue jacket. `<person2>`, dark-haired woman right, wears a black jacket. Colorful city mural behind. The entire image, including both people, is hand-drawn comic art. Large clear faces. Exactly two adults.
B1	14142139	Chest-up rainy city photo of two adults filling frame. Wet street reflections and storefront lights visible behind. `<person1>`, auburn-haired woman left, wears a closed navy raincoat. `<person2>`, older gray-haired man right, wears a closed bright-orange raincoat. Large reference-faithful faces, hands outside frame, no crowd or other person.
B2	32452843	Chest-up cafe photo. `<person1>`, red-haired woman left, wears a gray-and-black V-pattern herringbone coat. `<person2>`, older gray-haired man right, wears a dark-green diamond-quilted sleeveless vest. Cafe tables and a warm counter behind. Exactly two adults.
B3	1415926	Tight library photo of two adults with large faces. `<person1>`, an auburn-haired woman with side bangs on the left, wears a crimson-red jacket with dense tiny yellow polka dots covering all its fabric. `<person2>`, an older gray-haired man on the right, wears a plain mustard-gold coat. Sharp reference-faithful faces, no other person.
B4	83121403	Pointillist portrait of exactly two adults at a carnival. `<person1>` red-haired woman left wears a bright red coat. `<person2>` older gray-haired man right wears a solid black coat. Equal-size chest-up figures, separate shoulders, Ferris wheel behind, no other person.
C1	550000007	Waist-up coastal rescue photo of exactly two women. `<person1>` left wears plain teal medical scrubs. `<person2>` right wears a bright-orange rescue jacket. Both face the camera with natural forward-looking eyes and large identity-faithful faces. A white rescue tent and the ocean are behind them. No other person.

Table 6 (continued)

Cell	Seed	Exact positive prompt
C2	11235813	Full-body photograph of exactly two adult women on an empty subway platform. `<person1>` stands on the left wearing a plain tan trench coat. `<person2>` stands on the right wearing a burgundy diamond-quilted jacket. Both stand side by side and face the camera with clear faces, cinematic station lighting, no other person.
C3	560000011	Blue-hour rooftop portrait. `<person1>` on the left wears a plain blue jacket. `<person2>` on the right wears a plain orange jacket. Both look at the camera with natural eyes. A city skyline is behind them. Exactly two women.
C4	300000017	Waist-up watercolor portrait by a canal. `<person1>`, younger woman left, wears a blue coat. `<person2>`, older woman right, wears a black jacket. An arched bridge is behind them. The entire image is watercolor. Large clear faces. Exactly two adults.

F Routing Strength and Onset Sensitivity

We test reference pairs 1, 3, and 5, prompts D01, D04, and Y01 (Table 4), and seeds 2657029907 and 2329732732: 18 images and 36 subject records per setting. The attribute schedule is $\lambda(n) = \lambda \mathbb{1}[n \geq n_0]$. We vary strength $\lambda \in \{0, .5, 1\}$ at onset $n_0 = 5$, or $n_0 \in \{5, 10, 20\}$ at $\lambda = 1$. Identity routing always starts at step 5.

References, prompts, seeds, initial noise, $z^{(5)}$, ownership maps, identity onset, scheduler, CFG, decoder, and negative prompts remain fixed within matched conditions. We reuse B1/M4 for $\lambda = 0$ and M6 for $(\lambda, n_0) = (1, 5)$; the other settings add 54 images.

Figure 13 shows Bind rising from .472 to .500 to .556 as λ increases; delaying attribute onset lowers it to .500 and .472. Table 7 reports PR of .222 at the default setting and .167 otherwise. Identity top-1 ranges from .861 to .889, FaceValid remains 1.0, and CLIP varies by less than .009. On this subset, the default setting achieves the highest Bind and PR among the tested configurations, supporting its use in our experiments.

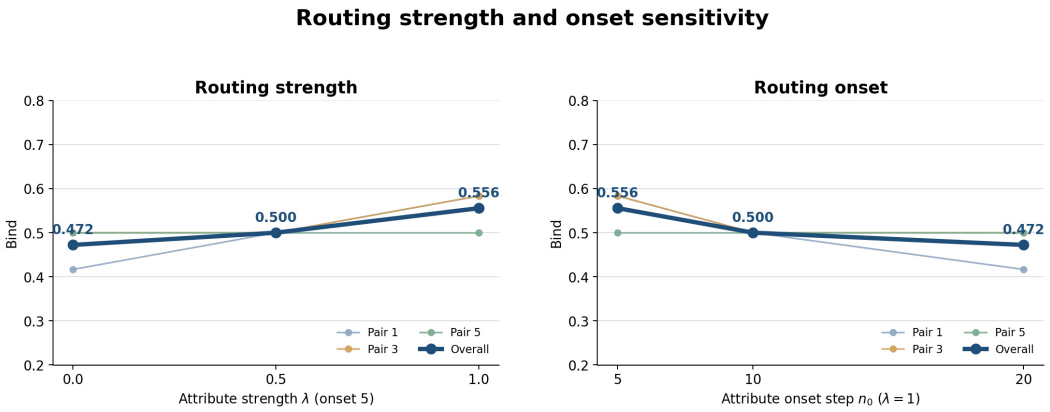

Figure 13: **Routing strength and onset sensitivity.** Thick curves aggregate 36 subject records; thin curves show 12 records per reference pair. Left: varying attribute strength at $n_0 = 5$. Right: varying attribute onset at $\lambda = 1$.

Table 7: Automatic metrics for the routing sensitivity experiment. Bind, ID top-1, and FaceValid use 36 subject records per setting; PR, TargetPerson, CLIP, and LPIPS use 18 images. LPIPS is measured against the matched $\lambda = 1, n_0 = 5$ output.

Variant	λ	n_0	Bind	PR	ID top-1	FaceValid	TargetPerson	CLIP mean±std	LPIPS-to-V1
V0	0.0	5	.472	.167	.889	1.000	.833	.3039 ± .0223	.0810 ± .0341
V05	0.5	5	.500	.167	.861	1.000	.778	.2984 ± .0195	.0426 ± .0252
V1	1.0	5	.556	.222	.861	1.000	.833	.2957 ± .0228	.0000 ± .0000
V10	1.0	10	.500	.167	.889	1.000	.778	.2975 ± .0215	.0645 ± .0324
V20	1.0	20	.472	.167	.889	1.000	.778	.3020 ± .0232	.0786 ± .0338