RIGHT IDENTITIES, WRONG PLACES: REFERENCE ADDRESSING IN MULTI-REFERENCE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-reference image generation integrates multiple visual references under textual instructions. When the prompt specifies spatial relations among these references, the model must ensure not only identity fidelity, but also **reference addressing**: assigning each reference to its designated spatial slot. Existing methods and evaluations, however, largely overlook reference addressing. Through a controlled counterfactual audit, we show that this is a distinct capability gap: models can preserve the requested identities yet fail to execute their prompt-specified reference-to-slot assignment, independent of identity loss, reference count, and physical input-order dependence. We introduce **AddrBind**, an inference-time executor that aligns an explicit reference-assignment plan with the model's emerging layout and redistributes, rather than amplifies, existing reference-conditioning influence, softly abstaining where the assignment is ambiguous. Across five reference-conditioning configurations and two benchmarks, AddrBind improves slot and exact-assignment accuracy in every configuration, with relative gains of 30.8% and 50.8% on our primary CogCanvas backbone. Code will be released at `anonymous.4open.science/r/AddrBind-ICLR`.

1 INTRODUCTION

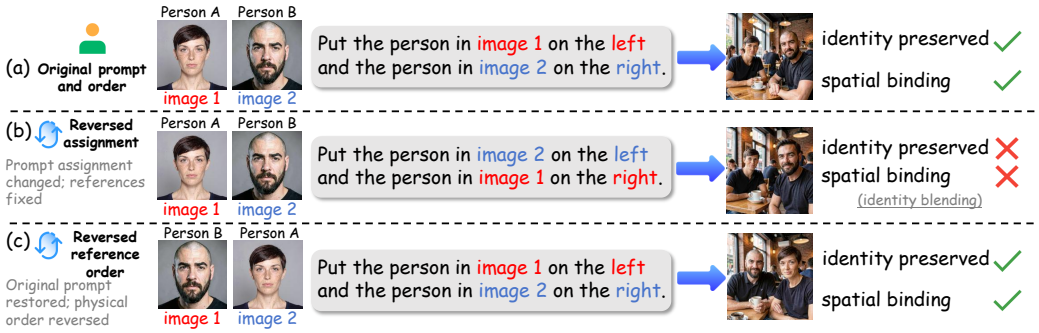


Figure 1: **Reference-addressing failure under changed instructions and physical input order.** (a) The model correctly places A left and B right. (b) Reversing the requested assignment with physical input order fixed fails to swap the identities and introduces blending. (c) Reordering references and updating prompt indices expresses the same assignment as (b), but yields correct placement.

Multi-reference image generation aims to synthesize images that follow textual instructions while preserving subject identities from multiple reference images (Xiao et al., 2024; Ma et al., 2024; Wang et al., 2025b; Chen et al., 2025; Wu et al., 2026). In applications such as **personalized group portraits** (Huang et al., 2026c) and **cast-based storyboards** (Huang et al., 2026b), users need to place specific people in prescribed positions. Referring to uploaded portraits by image index separates identity selection from appearance, allowing users to specify who goes where even when clothing or other visual attributes change. Consider two reference portraits and the instruction to place the person in Image 1 on the left and the person in Image 2 on the right. If the instruction is reversed, the generated identities should exchange positions. Figure 1 shows a failure of this basic

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.

requirement: the model follows the original assignment, but fails to follow the reversed assignment and introduces identity blending. Reordering the references and updating the image indices to express the same reversed assignment produces a successful arrangement. This contrast shows that the requested layout is within the model's generative capability, but that the assignment expressed in the prompt does not reliably govern which identity occupies each position. Accurately following such instructions requires **reference addressing**: determining which reference the prompt names and placing the corresponding identity in its requested spatial role.

The failure in Figure 1 highlights a distinction between identity fidelity and reference addressing. Identity fidelity does not imply successful reference addressing: a model can preserve every requested identity yet fail to place each one where the prompt assigns it. Personalization methods primarily focus on representing and preserving one or multiple referenced identities (Kumari et al., 2023; Ma et al., 2024; Jang et al., 2024; Wang et al., 2026), while unified or structured multi-reference models improve general cross-reference instruction following through multimodal training, indexed representations, or structured context (Wu et al., 2026; Peng et al., 2026). These advances do not by themselves establish whether changing the assignment in the prompt reliably changes where each identity appears. Spatial-control methods provide additional means of controlling placement through boxes, masks, points, or location canvases (Li et al., 2023; Xie et al., 2023; Wang et al., 2024; Bar-Tal et al., 2023; Yuan, 2026), but require spatial information beyond the textual instruction. The unresolved question is whether models can reliably execute the reference-to-slot assignment specified in the prompt.

Our diagnostic study across five reference-conditioning configurations confirms that the failure in Figure 1 extends beyond an isolated example (Section 2.2). Models can recover the requested identities while failing to follow changes to their spatial assignment, and several configurations are also sensitive to the physical input order. By controlling for identity loss, reference count, and physical input order, our audit isolates reference-addressing failure as a distinct capability gap. These findings motivate our central idea: make the prompt's reference assignment explicit and use it to direct each reference's conditioning influence toward the location where its corresponding subject is being formed.

To realize this idea, we propose **AddrBind**, an inference-time executor that aligns the requested assignment with the model's emerging layout. AddrBind first extracts which reference belongs in each spatial role and converts the spatial instructions into a coarse spatial prior. It then combines this prior with an initial observation of where the model is beginning to form each subject, identifying where each reference should exert its conditioning influence. Finally, it redistributes the model's existing reference-conditioning influence toward the assigned references at those locations. The routing preserves the total reference-conditioning budget, bounds the strength of the correction, and retains more of the original allocation where the assignment is ambiguous. This provides an explicit mechanism for executing the assignment without retraining or external spatial inputs.

We implement AddrBind across five native and adapter-based reference-conditioning configurations, improving slot and exact-assignment accuracy over native generation in every configuration on both CogCanvas and MultiBanana. On FLUX.2-klein-4B, AddrBind raises CogCanvas slot accuracy from 60.7% to 79.4% and exact-assignment accuracy from 38.8% to 58.5%, corresponding to relative gains of 30.8% and 50.8%, respectively. Our contributions are as follows:

(1) We identify *reference-addressing failure* as a distinct failure mode in multi-reference image generation. Our controlled counterfactual audit shows that this failure is distinct from identity loss, reference count, and physical input-order dependence, establishing the reference-to-slot assignment as a requirement separate from identity fidelity.

(2) We propose **AddrBind**, an inference-time executor that aligns an explicit reference-assignment plan with spatial evidence from the model's emerging layout during generation and executes the resulting reference-to-slot assignment by redistributing existing reference-conditioning influence, without retraining or user-provided spatial controls.

(3) We validate AddrBind across five native and adapter-based reference-conditioning configurations and two benchmarks, spanning different model scales and reference settings, and show consistent improvements in prompt-specified assignment while analyzing the effects on identity fidelity, surrounding prompt content, and image quality.

2 REFERENCE ADDRESSING: PROBLEM AND DIAGNOSIS

Multi-reference prompts specify which reference supplies each subject and where that subject should appear; we call this correspondence a **reference-to-slot assignment**. We use CogCanvas for controlled diagnosis and MultiBanana for cross-benchmark evaluation, restricting the study to open-weight models because AddrBind requires internal access to per-reference conditioning.

2.1 TASK DEFINITION AND EVALUATION

Formal setting. Let x be a natural-language prompt and $\mathcal{I} = (I_1, \ldots, I_R)$ an ordered tuple of R reference images. We assume the standard interface in which the reference index names the position of an image in the input tuple, so that "image r" refers to the r-th reference image I_r. The prompt specifies S distinct requested subjects, and m_s denotes the text span used to address subject s. Let $\pi^\star(s) \in \{1, \ldots, R\}$ identify the reference image supplying subject s, and let $\Omega_s^\star$ denote the evaluation slot corresponding to its requested spatial role. The complete assignment is

$$\mathcal{Z}^\star = \{(m_s, \pi^\star(s), \Omega_s^\star)\}_{s=1}^{S}.$$

We study the one-to-one setting with $R \geq S$, injective $\pi^\star$, and distinct identities among the requested references. Given only x and $\mathcal{I}$, successful reference addressing requires the identity from $I_{\pi^\star(s)}$ to occupy the spatial role represented by $\Omega_s^\star$ for every s. In this task setting, the annotated target slots are used only for evaluation and are not supplied to the model. Ground-truth spatial annotations are provided only to explicitly identified privileged controls in the experiments.

Evaluation measures. We distinguish correct placement from identity recovery. An output that faithfully reproduces both requested identities but exchanges their positions has complete identity coverage and zero slot accuracy.

Let Y denote a generated image. Detected subject instances are matched one-to-one to target slots using their center distances and a fixed matching radius. For slot $\Omega_s^\star$, $g(Y, \Omega_s^\star)$ returns the matched instance z, or $\varnothing$ if no instance is matched. For a detected instance z, $\mathrm{id}(z)$ denotes the index of its most similar supplied reference, with $\mathrm{id}(\varnothing) = \perp$. Slot accuracy measures the agreement between these nearest-reference labels and the requested assignment:

$$\mathrm{SlotAcc}(Y, \mathcal{Z}^\star) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}[\mathrm{id}(g(Y, \Omega_s^\star)) = \pi^\star(s)] \,.$$

Exact-assignment accuracy is the fraction of outputs for which $\mathrm{SlotAcc}(Y, \mathcal{Z}^\star) = 1$.

Identity coverage measures recovery irrespective of location using thresholded one-to-one identity matching. Let $\mathcal{R}^\star = \{\pi^\star(s) : s = 1, \ldots, S\}$ be the requested reference indices, and let $\widehat{\mathcal{R}}(Y)$ be the set of reference indices accepted by this matching procedure. Identity coverage is then $\mathrm{Cov}(Y) = |\mathcal{R}^\star \cap \widehat{\mathcal{R}}(Y)|/S$. Slot accuracy uses nearest-reference labels, whereas coverage requires accepted identity matches. Figure 2(c) uses this thresholded coverage. The cross-configuration audit in Table 1 additionally reports the fraction of distinct nearest-reference labels recovered across the requested slots. Consequently, slot accuracy does not independently certify identity fidelity. Appendix A specifies the complete protocol.

2.2 CONTROLLED INTERVENTIONS

We test two complementary requirements: changing the requested assignment should change which identity occupies each slot, whereas changing only the input presentation should preserve the requested placement. All paired comparisons fix the scene, reference set, initial noise, and sampling configuration; $K = S$ denotes the number of requested subjects.

Cyclic assignment: change the requested placement. We keep the physical input order fixed and rotate the identities requested at the same slots, e.g., changing "Image 1 on the left, Image 2 on the right" to the reversed assignment. Figures 1(a–b) and 2(a) illustrate how this tests whether the output follows the changed instruction.

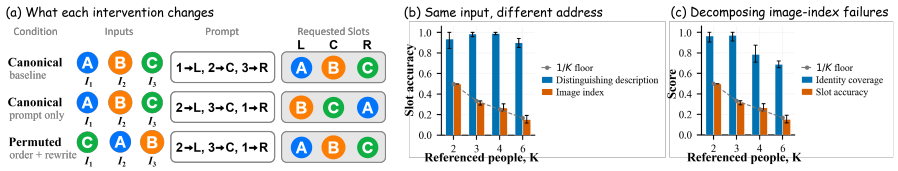


Figure 2: **Controlled reference-assignment audit on FLUX.2-klein-4B.** (a) Cyclic assignment changes only the prompt-specified assignment, whereas synchronized permutation changes physical input order while preserving that assignment. (b) Distinguishing descriptions remain reliable as K grows, while image-index expressions approach the $1/K$ floor. (c) Identity coverage remains above slot accuracy on the same image-index outputs; error bars show 95% seed-clustered intervals.

Table 1: **Native reference-assignment audit on CogCanvas.** Canonical uses the released request; Permuted changes physical input order while preserving the intended placement; Cyclic changes the assignment with physical input order fixed. Distinct labels measures the fraction of distinct supplied-reference labels recovered across requested slots, and Unfilled measures slots without a matched instance. Full diagnostics are reported in Appendix Table 5.

	Slot accuracy		Cyclic evaluation			
Configuration	Canonical ↑	Permuted ↑	Slot acc. ↑	Distinct labels ↑	Unfilled ↓	Tasks
FLUX.2-klein-4B	.598	.349	.259	.772	.073	150
FLUX.2-klein-9B	.636	.331	.234	.798	.091	150
OmniGen2	.375	.375	.350	.655	.220	100
SDXL + IP-Adapter	.157	.164	.170	.320	.632	150
Qwen-Image-Edit-2511	.357	.309	.341	.650	.212	150

Synchronized permutation: preserve the requested placement. We reorder the reference tuple and update its prompt indices while preserving the intended placement, as illustrated by the equivalent requests in Figure 1(b–c). This tests whether assignment execution remains reliable under a different input presentation.

Across the K cyclic shifts, any request-insensitive complete assignment that fills every slot with a supplied identity matches each slot's request in exactly one shift, yielding a mean slot accuracy of $1/K$. Figure 2 evaluates these complete cycles, whereas Table 1 uses one nonzero cyclic shift per task and is therefore not compared with the $1/K$ complete-cycle reference. We classify every requested slot as *correct*, *filled by another supplied identity*, or *unfilled*, so that wrong placement is not conflated with identity omission.

2.3 Audit Results

Reference count and identity loss do not explain the failure. Figure 2(b) reveals a clear separation between two ways of addressing the same reference set. Distinguishing descriptions remain reliable as the number of references grows, showing that the model can organize multiple identities when they are addressed by appearance. Image-index expressions instead remain near the $1/K$ floor, so the degradation is specific to resolving index-based references rather than increasing reference count alone. The complementary comparison in Figure 2(c) shows that identity coverage consistently exceeds slot accuracy on the same outputs. The requested identities are therefore often recovered even when their prompt-specified assignment is not executed.

Assignment failures across configurations. Table 1 shows low cyclic-assignment accuracy across all five evaluated configurations, extending the diagnosis beyond the backbone examined in Figure 2. Physical input-order sensitivity is pronounced on the two FLUX configurations: reordering the references and updating the prompt indices while preserving the intended identity placement reduces slot accuracy by 24.9 and 30.5 percentage points on FLUX.2-klein-4B and FLUX.2-klein-9B, respectively.

These results show that identity recovery alone does not establish whether the prompt governs the generated arrangement. They motivate making the requested reference assignment explicit and con-

necting it to spatial reference conditioning. AddrBind implements this connection by combining the assignment with early spatial evidence and directing reference influence toward the corresponding regions.

3 METHODS

A prompt such as "Image 2 on the left, Image 1 on the right" specifies a subject-reference assignment, but executing it requires determining where and how each reference should influence generation. AddrBind separates assignment planning from generation-time execution, first making the requested associations explicit from the prompt x and supplied references $\mathcal{I}$. To determine where these associations should act, it combines coarse spatial guidance from language with imperfect localization evidence from early subject-related attention. These signals guide a spatial redistribution of reference attention that preserves its total mass at each modified operation, with separate controls for spatial ambiguity and intervention magnitude. Figure 3 summarizes the pipeline, and Algorithm 1 states the complete procedure.

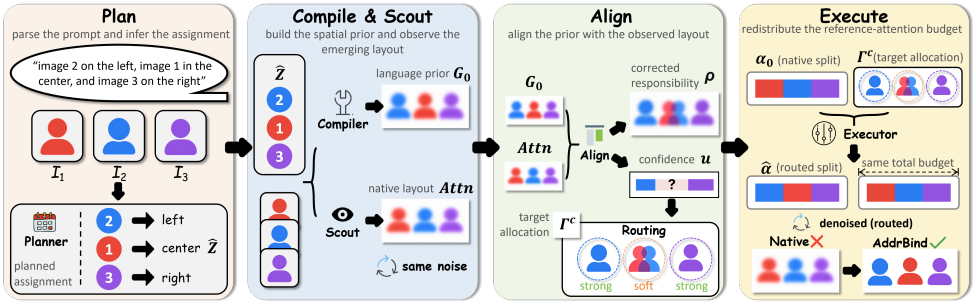


Figure 3: **AddrBind overview.** AddrBind extracts the prompt-specified reference assignment, compiles it into a spatial prior, and aligns that prior with the model's early layout. The executor converts the resulting confidence-weighted responsibilities into a budget-conserving redistribution of reference attention, attenuating intervention where the assignment is ambiguous.

3.1 ASSIGNMENT PLAN AND COMPILATION

Discrete reference-assignment plan. The executor first needs an explicit representation of which reference should influence which subject. A multimodal planner reads $(x, \mathcal{I})$ and produces a reference-assignment plan $\widehat{\mathcal{Z}} = \{\mathcal{D}_s\}_{s=1}^{\widehat{S}}$, with one record $\mathcal{D}_s = (\sigma_s, r_s, \ell_s, v_s)$ for each estimated distinct subject. Repeated mentions of the same subject are consolidated before compilation: σ_s is the union of token positions across these mentions, $r_s \in \{1, \dots, R\}$ identifies the assigned reference, and ℓ_s specifies the requested location. A deterministic validator sets $v_s = 0$ for records with unresolved references, invalid spans, or unsupported spatial phrases.

Deterministic compilation. Let $\mathcal{A} = \{s \in \{1, \dots, \widehat{S}\} : v_s = 1\}$ index the active records. The compiler then converts the planner's records into a spatial prior. For each $s \in \mathcal{A}$, the compiler retains r_s and selects the non-address tokens in σ_s as subject-related text for the scout observation. Address tokens such as an image index name which reference is meant rather than what it depicts, so they provide no appearance evidence for localization. Each spatial phrase is converted through a predefined lexicon into a smooth canvas mask, and overlapping masks are normalized across subjects to form the language-derived spatial prior G^0, where G^0_{sq} denotes the responsibility of subject s at canvas position q. G^0 expresses relative subject responsibility rather than foreground occupancy, partitioning the canvas among subjects instead of marking which pixels belong to a subject.

3.2 LAYOUT-INFORMED EXECUTION

Scene-corrected responsibility and soft abstention. The language prior G^0 specifies where each subject should appear, but not where the model is beginning to form it. AddrBind therefore runs

one denoising step without routing. It measures the average canvas-to-text attention response $\overline{\mathrm{Attn}}(q, \sigma_s)$ from canvas position q to the retained non-address tokens in each subject's token set σ_s. The scout trajectory is discarded, and generation restarts from the original sampling state with the recorded attention held fixed. Combining the requested and observed layouts gives the corrected responsibility

$$\rho_{sq} = \frac{G^0_{sq}\,\overline{\mathrm{Attn}}(q, \sigma_s)}{\sum_{j \in \mathcal{A}} G^0_{jq}\,\overline{\mathrm{Attn}}(q, \sigma_j)},$$

When the denominator vanishes, we set $\rho_{sq} = 0$ and $u_q = 0$. Thus, a subject receives high responsibility only where the requested spatial prior and the model's emerging spatial evidence agree. Otherwise, the routing strength is the normalized confidence $u_q = 1 - H(\rho_{\cdot q})/\log|\mathcal{A}|$, where H is Shannon entropy over subjects; for a single active subject, we set $u_q = 1$. Thus AddrBind intervenes strongly where subject responsibility is clear and attenuates routing where it is ambiguous.

Budget-conserving routing. Spatial responsibility estimates where an assignment should act, but does not itself determine how to change reference conditioning. The key is to control how attention is divided among references: increasing reference attention overall does not specify which identity should dominate at a location. AddrBind therefore changes the relative reference allocation while preserving its total attention mass. For each head h and canvas query q, let α^0_{hqr} denote the native attention share of reference block r, normalized across the R reference blocks; each block contains the keys of one reference image. The plan converts subject responsibilities into a reference-specific target:

$$\Gamma^c_{qr} = u_q \sum_{s \in \mathcal{A}} \rho_{sq} \mathbb{1}[r = r_s], \qquad \sum_{r=1}^{R} \Gamma^c_{qr} = u_q.$$

Thus, Γ^c_q allocates the fraction u_q according to the estimated subject responsibilities and their assigned references.

For a candidate scale $\lambda \in [0, 1]$ and fixed logit-shift bound D, define

$$\widehat{\alpha}_{hqr}(\lambda) = (1 - \lambda u_q)\alpha^0_{hqr} + \lambda \Gamma^c_{qr}, \qquad \delta_{hqr}(\lambda) = \log \frac{\widehat{\alpha}_{hqr}(\lambda)}{\alpha^0_{hqr}}.$$

We choose the largest feasible scale $\lambda_{hq} = \max\{\lambda \in [0, 1] : \|\delta_{hq}(\lambda)\|_\infty \le D\}$ and add $\delta_{hqr}(\lambda_{hq})$ uniformly to all keys in reference block r. Since $\lambda = 0$ leaves the allocation unchanged, a feasible scale always exists.

For shared attention, the blockwise shifts also preserve the relative token weights within each reference block. The same normalized construction applies to the branch coefficients of separable reference adapters such as IP-Adapter. The routing transform is equivariant to a common relabeling of reference coordinates: applying the same permutation to Γ^c and α^0 permutes the resulting δ accordingly. It therefore introduces no reference-index preference beyond that already present in the estimated plan and the model's native allocation. The early-layout quantities (ρ, u, Γ^c) are computed once and held fixed, whereas $(\alpha^0, \lambda, \delta)$ are recomputed from the current attention allocation at every routed call.

4 Experiments

Having diagnosed the assignment failures in Section 2, we evaluate whether AddrBind improves reference addressing, which components contribute to the correction, and how its benefits and costs vary across configurations and benchmarks.

4.1 Setup

Benchmarks and metrics. We use CogCanvas (Nguyen et al., 2026) as the primary benchmark, reporting slot accuracy, exact-assignment accuracy, identity coverage, and the quality and prompt-following metrics described in Appendix E. Broader generalization is evaluated on a multi-person subset of MultiBanana (Oshima et al., 2026), using its official metrics together with identity presence, with placement evaluated only where person-specific spatial ground truth is unambiguous.

Appendix C records how the released splits narrow to these sets. We additionally use the controlled cyclic evaluation from Section 2.2 to test whether methods follow prompt-specified reference-to-slot assignments or a fixed input order.

Backbones. The five configurations comprise four native visual-reference interfaces: FLUX.2-klein-4B, FLUX.2-klein-9B (Black Forest Labs, 2025), OmniGen2 (Wu et al., 2026), and Qwen-Image-Edit-2511 (Wu et al., 2025). The fifth uses an adapter-based interface, SDXL (Podell et al., 2024) with per-reference IP-Adapter branches (Ye et al., 2023). Native and AddrBind are paired on identical tasks, prompts, references, seeds, and sampling settings. We report AddrBind only for interfaces that expose separable per-reference conditioning.

Baselines. Native generation is the input-matched baseline and receives the same prompt and references as AddrBind. The spatial-control baselines include static GT-region routing and reference-aware adaptations of BoxDiff (Xie et al., 2023), DenseDiffusion (Kim et al., 2023), MultiDiffusion (Bar-Tal et al., 2023), and regional prompting (Chen et al., 2024), which apply each method's spatial-control rule to the corresponding per-reference conditioning channel. We evaluate these adaptations on FLUX.2-klein-9B, OmniGen2, and SDXL + IP-Adapter as representative native and adapter-based interfaces. MS-Diffusion (Wang et al., 2025b) and SIGMA-GEN (Saha et al., 2026) are evaluated with their released backbones and required layout interfaces, instantiated from the same ground-truth target regions.

Table 2: **Reference-assignment results on CogCanvas.** Native and AddrBind use only the prompt and ordered references; static GT-region routing and Oracle plan receive the ground-truth target regions and assignment; backbone-adaptable methods apply their spatial-control rule to per-reference conditioning channels, while released methods use their own backbones and layout interfaces.

Backbone	Arm	Slot ↑	Exact ↑	Cov. ↑	Empty ↓
FLUX.2-klein-4B	Native	.607	.388	.690	**.064**
	Static GT-region routing	.871	.731	.892	.049
	AddrBind	**.794**	**.585**	**.883**	.100
	Oracle plan	.800	.595	.887	.101
FLUX.2-klein-9B	Native	.635	.443	.843	**.070**
	Static GT-region routing	.901	.791	.944	.036
	AddrBind	**.826**	**.665**	**.950**	.089
	Oracle plan	.832	.675	.951	.090
OmniGen2	Native	.367	.181	.506	.292
	Static GT-region routing	.909	.842	.920	.059
	AddrBind	**.905**	**.840**	**.935**	**.069**
	Oracle plan	.909	.844	.939	.070
SDXL + IP-Adapter	Native	.160	.026	.083	.638
	Static GT-region routing	.137	.028	.043	.753
	AddrBind	**.459**	**.220**	**.300**	**.468**
	Oracle plan	.465	.230	.304	.462
Qwen-Image-Edit-2511	Native	.368	.128	**.408**	.231
	Static GT-region routing	.417	.164	.326	.211
	AddrBind	**.447**	**.199**	.351	**.221**
	Oracle plan	.452	.200	.359	.220
Backbone-adaptable methods					
FLUX.2-klein-9B	BoxDiff	.636	.446	.845	.072
	DenseDiffusion	.604	.412	.826	.081
	MultiDiffusion	.648	.459	.851	.068
	Regional prompting	.678	.487	.864	.032
OmniGen2	BoxDiff	.395	.196	.521	.272
	DenseDiffusion	.352	.170	.493	.300
	MultiDiffusion	.404	.203	.529	.265
	Regional prompting	.438	.228	.551	.241
SDXL + IP-Adapter	BoxDiff	.162	.026	.084	.641
	DenseDiffusion	.128	.021	.052	.679
	MultiDiffusion	.164	.033	.025	.624
	Regional prompting	.164	.038	.053	.601
Backbone-specific released methods					
Released backbone	MS-Diffusion	.905	.792	.818	.061
	SIGMA-GEN	.917	.812	.912	.043

Table 3: **Reference-assignment results on MultiBanana.** Native and AddrBind use identical prompts and ordered references. Seat and Exact measure person-specific placement, while Pres. and Recall measure identity presence.

Backbone	Arm	Seat ↑	Exact ↑	Pres. ↑	Recall ↑
FLUX.2-klein-4B	Native	.468	.281	.179	.220
	AddrBind	**.537**	**.363**	**.195**	**.240**
FLUX.2-klein-9B	Native	.470	.308	.258	.353
	AddrBind	**.548**	**.370**	**.268**	**.369**
OmniGen2	Native	.474	.265	.178	.267
	AddrBind	**.509**	**.353**	**.187**	**.308**
SDXL + IP-Adapter	Native	.153	.089	.043	.009
	AddrBind	**.212**	**.096**	**.046**	**.015**
Qwen-Image-Edit-2511	Native	.332	.198	**.140**	**.192**
	AddrBind	**.366**	**.224**	.134	.191

Seat/Exact measure per-slot/full-assignment correctness; Pres./Recall ignore location. Gray bold compares Native with AddrBind in Tables 2–3 and non-privileged variants in Table 4; privileged controls and released methods are unmarked.

Table 4: **FLUX.2-klein-9B component ablation.** Static GT uses the ground-truth assignment without layout feedback; No ambiguity gate fixes the entropy-based confidence to one; Oracle plan supplies the ground-truth assignment to the full executor.

Variant	Slot ↑	Exact ↑	Cov. ↑	Empty ↓	Duplicate ↓
Native	.636	.427	.836	**.069**	.222
Static GT routing	.896	.773	.949	.040	.138
No ambiguity gate	.819	.640	.924	.077	.144
AddrBind	**.832**	**.673**	**.948**	.072	**.138**
Oracle plan	.842	.693	.953	.076	.122

4.2 MAIN RESULTS

We first test whether reference-addressing failure can be corrected without user-provided spatial annotations. Table 2 compares native generation with AddrBind across five CogCanvas configurations and reports spatial-control baselines and oracle-plan controls on the same metrics. AddrBind improves slot and exact-assignment accuracy in every configuration; on FLUX.2-klein-4B, slot accuracy rises from .607 to .794. This result approaches the .800 oracle-plan variant, indicating that plan errors contribute little to the remaining placement error on this backbone. Static GT-region routing exceeds AddrBind on FLUX.2-klein-4B, FLUX.2-klein-9B, and OmniGen2 because it receives the ground-truth target regions that AddrBind must infer. In contrast, AddrBind exceeds static GT-region routing by .322 on SDXL and .030 on Qwen-Image-Edit-2511. The same pattern extends to MultiBanana, where Table 3 shows higher seat and exact-assignment accuracy on all five backbones.

Figure 4 shows qualitative examples: one shared $K = 2$ prompt across five backbones, the privileged controls at $K = 3$, and the correction at $K = 5$.

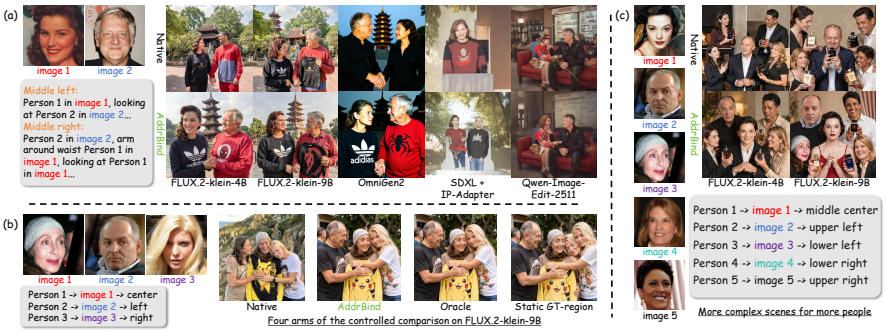


Figure 4: **Qualitative reference assignment.** Numbered references and prompts accompany outputs for $K = 2$ across five backbones, $K = 3$ with privileged controls on FLUX.2-klein-9B, and $K = 5$ on both FLUX backbones; methods share prompts, references, and sampling settings within a panel.

Component ablation. Table 4 evaluates privileged spatial supervision, ambiguity-aware abstention, and planner error on the 150-task FLUX.2-klein-9B audit subset under matched sampling settings. Static GT routing attains the highest slot accuracy at .896, exceeding AddrBind by .064; removing the ambiguity gate reduces slot accuracy by .013, while the oracle plan adds .010.

4.3 SCOPE AND COST

We examine model scale and a second benchmark; Appendix Table 7 reports the held-out $K = 4/5$ results.

Scale does not close the gap. The two FLUX backbones share an architecture and reference interface but differ by $2.25\times$ in parameters, which gives a direct test of whether scale closes the addressing gap. It does not. Scaling lifts native slot accuracy only from .607 to .635, and the AddrBind gain is essentially the same on both models, $+.191$ against $+.187$. The gap also widens as the number of references grows rather than being absorbed by capacity. On the 9B model, native slot accuracy falls from .714 at $K = 2$ to .516 at $K = 3$ while AddrBind moves only from .834 to .815, so the correction grows from $+.120$ to $+.299$. Appendix D.1 separates the two cardinalities. A model that generates more faithfully has not thereby learned which input a phrase names. Routing buys this correction by trading duplicated identities for unfilled seats, reducing the former from .206 to .138 and raising the latter from .070 to .089.

A second benchmark. MultiBanana (Oshima et al., 2026) asks the same question in prompts we did not write, using image-index expressions with K from 2 to 7 and R from 2 to 8. On the 491 tasks that pass its reference audit, FLUX.2-klein-4B routes 152 and abstains on the rest, and the abstained

outputs are unchanged from native. Where it routes, the correction reproduces: identity presence rises from .186 to .235 and per-seat placement from .532 to .619. Absolute presence remains low under both arms, so this is a directional replication on an external benchmark. Appendix F gives the full decomposition.

Trade-offs. The correction is not free, but the cost is small next to what it buys. Identity similarity improves on four of the five backbones and is unchanged on the fifth, so the assignment gain does not come at the expense of identity preservation, and DINO similarity improves on three. Prompt alignment, aesthetics, and background similarity move down only slightly. Two independent vision–language judges find the same pattern in the surrounding prompt content, where the largest absolute change reaches .076 on the smaller FLUX backbone and .158 on the larger, and the declines concentrate on OmniGen2 and SDXL. Appendix Tables 9, 10, and 11 report the complete paired and absolute results.

5 Related Works

Multi-reference personalization and reference assignment. Subject-driven generation represents user-specific concepts through learned embeddings or fine-tuned parameters (Gal et al., 2023; Ruiz et al., 2023), and multi-concept customization and image-prompt adapters extend this interface beyond one concept (Kumari et al., 2023; Ye et al., 2023). Multi-subject systems reduce identity blending through localized attention, layout guidance, or learned reference–target correspondence (Ma et al., 2024; Xiao et al., 2024; Wang et al., 2025b; She et al., 2025; Xu et al., 2026). Recent methods make correspondence explicit through location canvases, positional identifiers, or structured context (Yuan, 2026; Huang et al., 2026a; Peng et al., 2026). These approaches demonstrate the value of explicit correspondence, but generally obtain it from training annotations or user-provided layouts, tags, or schemas rather than natural-language reference expressions. Existing benchmarks expose swaps, dominance, blending, and identity collapse (Tian et al., 2026; Chen et al., 2026; Nguyen et al., 2026), but do not isolate image-index execution from physical input order. Our cyclic protocol targets this distinction, related to position bias in multi-image retrieval (Xian et al., 2026), while separating identity coverage from prompt-specified placement.

Grounded planning and attention-based control. Language-guided planners translate complex prompts into layouts or regional programs (Feng et al., 2023; Lian et al., 2024; Yang et al., 2024). Given such a plan, grounded layers, fused diffusion paths, box constraints, and attention modulation impose spatial control on pretrained models (Li et al., 2023; Bar-Tal et al., 2023; Xie et al., 2023; Kim et al., 2023; Chen et al., 2024), while attention guidance mitigates omission and attribute misbinding (Chefer et al., 2023). Causal analysis shows that identity can flow directly from a reference to output tokens (Ge et al., 2026), while MAST/LAMA allocates attention mass among masked reference groups (Kang et al., 2026). Unlike MAST/LAMA, which receives masks, AddrBind infers the reference-to-slot assignment from language and changes only the relative split among references while preserving the local reference budget. Existing grounded methods therefore enforce a correspondence once its target layout is supplied; AddrBind addresses the preceding problem of inferring and executing that correspondence from the prompt.

6 Conclusion

We study reference addressing: binding language-addressed visual references to spatial roles in multi-reference image generation. Our counterfactual audit separates it from identity fidelity, reference count, and physical input-order dependence, showing that identities can be recovered without executing their assignment. AddrBind aligns an explicit reference-assignment plan with the emerging layout and executes it through budget-conserving routing with ambiguity-aware abstention. Across models, reference cardinalities, and two benchmarks, it consistently improves prompt-specified assignment without retraining or spatial annotations, generally preserving identity similarity with backbone-dependent quality trade-offs. Controllable generation must preserve the right identities and bind each one to its instructed role.

AI USE STATEMENT

The generative AI tool is only used for the early collection of basic related literature, and is responsible for language editing and proofreading during the completion of the paper, including improving the grammar, clarity and wording of the author's written text. They are not used to generate research ideas, develop methods, conduct experiments, analyze results, or make scientific claims. All AI assisted editors are reviewed by authors to ensure that they do not change the technical meaning or introduce unsupported content. The author is fully responsible for the final content of this work.

ETHICS STATEMENT

This work uses public benchmark data, collects no new personal data, and involves no human subjects, crowdsourcing, or user study. The evaluation pipeline uses a released face detector and frozen ArcFace embeddings only to match generated instances against the references supplied with each benchmark task, and does not identify individuals outside those reference sets. AddrBind inherits the potential misuse risks of personalized image generation, including non-consensual identity manipulation; we will release the reference-addressing audit and scoring tools alongside the executor so that its behavior can be measured rather than assumed.

REPRODUCIBILITY STATEMENT

We will release the AddrBind executor, the reference-assignment scorer, and the scripts used to construct the evaluated task sets and the reported tables; the underlying benchmarks are public. The scoring protocol, including detection, matching radii, the coverage readouts, the calibrated thresholds, and the bootstrap procedures, is specified in Appendix A. The inference procedure is given in Appendix B, the task construction and split sizes in Appendix C, and the quality and judge implementations in Appendix E. The released configurations will record the planner and judge models, prompts, sampling settings, thresholds, the shift bound, and the per-image metric rows, so the reported aggregates can be recomputed without regenerating images. All reported contrasts are paired within task, with uncertainty resampled over the clustering unit named in the corresponding caption.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752, 2023.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025.

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. *ACM Transactions on Graphics*, 42(4):1–10, 2023.

Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024.

Ruoxi Chen, Dongping Chen, Siyuan Wu, Sinan Wang, Shiyun Lang, Petr Sushko, Gaoyang Jiang, Yao Wan, and Ranjay Krishna. Multiref: Controllable image generation with multiple visual references. *arXiv preprint arXiv:2508.06905*, 2025. URL `https://arxiv.org/abs/2508.06905`.

Zhihan Chen, Yuhuan Zhao, Yijie Zhu, and Xinyu Yao. When identities collapse: A stress-test benchmark for multi-subject personalization. *arXiv preprint arXiv:2603.26078*, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18225–18250, 2023.

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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Chris Ge, Rohit Gandikota, Antonio Torralba, and Tamar Rott Shaham. Vision-language binding in in-context image generation. *arXiv preprint arXiv:2605.24624*, 2026.

GLM Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Binyuan Huang, Yuning Lu, Weinan Jia, Hualiang Wang, Mu Liu, and Daiqing Yang. Rethinking position embedding as a context controller for multi-reference and multi-shot video generation. *arXiv preprint arXiv:2604.03738*, 2026a.

Junjia Huang, Binbin Yang, Pengxiang Yan, Jiyang Liu, Bin Xia, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamshot: Personalized storyboard synthesis with video diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36736–36746, June 2026b.

Meijia Huang, Ruida Li, Bing Ma, Liangwei Jiang, Shuo Fang, and Chenguang Ma. GroupPortrait: Multi-id portrait generation with high identity preservation and fine-grained control. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6226–6235, 2026c.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100895–100937, 2024. doi: 10.52202/079017-3200. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b6e67ae290635d0874c4cb43ba2a2cfb-Paper-Conference.pdf.

Dongkyung Kang, Jaeyeon Hwang, Junseo Park, Minji Kang, Yeryeong Lee, Beomseok Ko, Hanyoung Roh, Jeongmin Shin, and Hyeryung Jang. Mast: Mask-guided attention mass allocation for training-free multi-style transfer. *arXiv preprint arXiv:2604.12281*, 2026.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7701–7711, 2023.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023. doi: 10.1109/CVPR52729.2023.00192.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023. doi: 10.1109/CVPR52729.2023.02156.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hFALpTb4fR.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024. doi: 10.1145/3641519.3657469.

Long-Bao Nguyen, Quang-Khai Tran, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. Cogcanvas: A benchmark for evaluating multi-subject reference-based image generation. *arXiv preprint arXiv:2606.15867*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–460, 2026.

Jianing Peng, Mengyu Wang, Henghui Ding, Zixiang Li, Ting Liu, Xiaochao Qu, Luoqi Liu, Yao Zhao, and Yunchao Wei. Structgen: Disambiguating multi-reference image generation via structured context modeling. *arXiv preprint arXiv:2607.15619*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Oindrila Saha, Vojtech Krs, Radomir Mech, Subhransu Maji, Kevin Blackburn-Matzen, and Matheus Gadelha. SIGMA-GEN: Structure and identity guided multi-subject assembly for image generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4748b6b6ca49f04b6a8cfce1d5f9a70-Abstract-Conference.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Wenqing Tian, Hanyi Mao, Zhaocheng Liu, Lihua Zhang, Qiang Liu, Jian Wu, and Liang Wang. Multibind: A benchmark for attribute misbinding in multi-subject generation. *arXiv preprint arXiv:2603.21937*, 2026.

Weiyun Wang, Zhangwei Gao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=PJqP0wyQek.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024.

Zhongsheng Wang, Zhedong Lin, Qian Liu, Xinyu Zhang, and Jiamou Liu. Sdo: Subspace deconflicting operator for multi-adapter composition. *arXiv preprint arXiv:2608.13820*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Mingtao Xian, Yifeng Yang, Qinying Gu, Xinbing Wang, and Nanyang Ye. Logit-attention divergence: Mitigating position bias in multi-image retrieval via attention-guided calibration. *arXiv preprint arXiv:2605.11591*, 2026.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2024. doi: 10.1007/s11263-024-02227-z.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7418–7427, 2023.

Yiyan Xu, Qiulin Wang, Wenjie Wang, Yunyao Mao, Xintao Wang, Pengfei Wan, Kun Gai, and Fuli Feng. Unicustom: Unified visual conditioning for multi-reference image generation. *arXiv preprint arXiv:2605.12088*, 2026.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56704–56721, 2024.

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. URL https://arxiv.org/abs/2308.06721.

Longhui Yuan. Anyphoto: Multi-person identity preserving image generation with id adaptive modulation on location canvas. *arXiv preprint arXiv:2603.14770*, 2026.

A REFERENCE-ASSIGNMENT AUDIT AND SCORING PROTOCOL

This section specifies the evaluation conditions, scoring implementation, and complete behavioral measures used by the reference-assignment audit in Section 2. The controlled grid in Figure 2 evaluates complete cyclic assignment sets, whereas Table 1 extends the audit across reference-conditioning configurations.

A.1 SAMPLES AND PROMPT CONSTRUCTION

Controlled grid in Figure 2. The grid uses FLUX.2-klein-4B with $K \in \{2, 3, 4, 6\}$ references, three levels of visual describability, two referring forms, and eight seeds (7, 11, 23, 42, 101, 202, 313, and 505). For each reference set, the descriptive prompt identifies subjects by appearance, whereas the image-index prompt uses expressions such as "the person from image 2." Each referring form is evaluated under all K cyclic shifts: shift zero is the canonical assignment and the remaining $K - 1$ shifts rotate which reference is requested at each slot. This gives 720 logical assignment conditions. Only 640 images are generated because, at the visually indistinguishable level, rotating the descriptive assignment leaves the prompt string unchanged and the shared image is scored against each corresponding assignment. Panels (b) and (c) use the visually distinct reference sets: panel (b) compares descriptive and image-index expressions, and panel (c) compares slot accuracy with thresholded identity coverage on the same image-index outputs.

Cross-configuration audit in Table 5. For FLUX.2-klein-4B, FLUX.2-klein-9B, SDXL + IP-Adapter, and Qwen-Image-Edit-2511, the audit uses a frozen set of 150 CogCanvas tasks, split evenly between $K = 2$ and $K = 3$. OmniGen2 is scored on 100 $K = 2$ tasks because the corresponding $K = 3$ inputs exceed its five-reference limit. Every admitted task has three paired outputs: the canonical request, one synchronized permutation, and one nonzero cyclic assignment. The cyclic column of Table 5 excludes the canonical shift. The three outputs share the scene, reference set, initial noise, and sampling configuration.

A.2 AUDIT CONDITIONS

Conditions and slot outcomes. The *canonical* condition uses the released request and its original physical input order. The *permuted* condition reorders the reference images and updates the corresponding image indices in the prompt so that the intended identity placement remains unchanged. The *cyclic* condition keeps the reference images and their physical input order fixed while rotating the identities requested at the target slots. Each requested slot is classified by its spatially matched instance as *correct* when the instance's nearest-reference label agrees with the reference assigned to the slot, *other reference label* when a matched instance carries a different supplied-reference label, or *unfilled* when no instance is matched. The three fractions sum to one, up to rounding.

Reference value. A request-insensitive arrangement that stays fixed across all K cyclic requests scores $1/K$ on average, because each slot's fixed identity is requested at that slot in exactly one of the K assignments.

Behavioral measures. *Permutation consistency* measures agreement between the realized identity placements in the canonical and synchronized-permutation outputs, over slots that are spatially matched in both. Reference labels are expressed in a common identity indexing before comparison, because the numerical input indices change under permutation. *Comparable* reports the fraction of slots eligible for that comparison. *Fixed-order* reports the fraction of cyclic outputs whose realized identity mapping agrees with the canonical output on every slot occupied in both.

A.3 SCORING IMPLEMENTATION

Detection, registration, and matching. We use InsightFace `buffalo_l` for both face detection and ArcFace identity embeddings (Deng et al., 2019). Generated images are detected at 640×640 resolution, and each detected face is represented by its normalized ArcFace embedding and normalized bounding-box center. For each input reference, the scorer first attempts registration at 640×640,

Table 5: **Complete native reference-assignment audit on CogCanvas.** Canonical uses the released request. Permuted changes physical input order and prompt indices while preserving the intended identity placement. Cyclic changes the requested assignment with physical input order fixed. Other ref. label marks a spatially matched instance whose nearest-reference label differs from the one requested for that slot. Distinct labels is the threshold-free fraction of distinct nearest-reference labels recovered across the requested slots.

	Slot accuracy		Cyclic evaluation				Behavioral diagnosis			
Configuration	Canonical ↑	Permuted ↑	Correct ↑	Other ref. label ↓	Unfilled ↓	Distinct labels ↑	Permutation consistency ↑	Fixed-order	Comparable	Tasks
FLUX.2-klein-4B	.598	.349	.259	.668	.073	.772	.493	.392	.916	150
FLUX.2-klein-9B	.636	.331	.234	.675	.091	.798	.452	.483	.893	150
OmniGen2	.375	.375	.350	.430	.220	.655	.744	.615	.682	100
SDXL + IP-Adapter	.157	.164	.170	.198	.632	.320	.779	.763	.336	150
Qwen-Image-Edit-2511	.357	.309	.341	.447	.212	.650	.497	.517	.716	150

falls back to 320×320 when necessary, and retains the largest detected face if more than one candidate is found. Each spatial phrase is mapped to a fixed anchor in normalized canvas coordinates. Detected faces and requested slots are matched one-to-one by Hungarian assignment on Euclidean center distance, without using identity similarity in the spatial assignment. Matches outside the frozen radius are discarded and the corresponding slots are marked unfilled. The controlled grid uses its frozen radius factor of 1.2 relative to the closest anchor spacing, whereas the CogCanvas audit uses a normalized radius of $.35$. After spatial matching, the identity at a filled slot is the reference with maximum cosine similarity to the matched face. A slot is correct when this reference index equals the index requested at that slot; an empty slot is incorrect. Exact-assignment accuracy equals one only when every requested slot is correct.

Coverage and uncertainty. Figure 2(c) uses thresholded one-to-one identity coverage. The scorer forms the cosine-similarity matrix between spatially matched faces and the requested references, finds a maximum-cardinality one-to-one matching above a frozen K-specific threshold, and divides the number of accepted matches by K. The thresholds are calibrated before scoring from impostor-only matched sets and equal $.2275$, $.2362$, $.2400$, and $.2482$ for $K = 2, 3, 4, 6$, respectively. Table 5 uses a threshold-free diagnostic derived from the same spatial matches: it reports the fraction of distinct supplied-reference labels recovered across the requested slots. The two readouts answer different questions: the thresholded one asks whether each requested identity was recovered, the threshold-free one how many distinct references are represented among the matched faces. For the controlled grid, confidence intervals resample the eight seeds while retaining all logical rows induced by a shared generated image in the same cluster. For the CogCanvas audit, intervals use $4{,}000$ bootstrap draws over background clusters with seed 20260826, keeping the paired conditions for a task together.

A.4 Interpretation of the audit results

Different mixtures of labeling errors and unfilled slots. The cyclic results reveal different observable failure patterns. On FLUX.2-klein-4B and FLUX.2-klein-9B, respectively, $.668$ and $.675$ of requested slots receive another reference label, while only $.073$ and $.091$ are unfilled. Distinct-label coverage remains $.772$ and $.798$.

SDXL + IP-Adapter exhibits a different pattern: $.632$ of requested slots are unfilled and distinct-label coverage is $.320$. Its low cyclic slot accuracy therefore reflects substantial spatial matching or detection failures as well as incorrect reference labels.

Input-order sensitivity. Synchronized permutation reduces slot accuracy by 24.9 percentage points on FLUX.2-klein-4B and 30.5 points on FLUX.2-klein-9B. Their permutation-consistency scores are $.493$ and $.452$, with comparable-slot fractions of $.916$ and $.893$.

OmniGen2 has identical aggregate canonical and permuted accuracy ($.375$), yet permutation consistency is only $.744$: the same aggregate score can arise from different per-slot placements. SDXL + IP-Adapter has higher permutation consistency ($.779$), but only $.336$ of slots are comparable.

Persistence and implication. Fixed-order rates range from .392 to .763 across the configurations. Together with the thresholded coverage analysis in Figure 2(c), these results separate reference-assignment execution from identity recovery and input-order stability.

B AddrBind inference procedure

The planner is `InternVL3.5-8B` (Wang et al., 2025a), a frozen open-weight multimodal language model used off the shelf, run greedily at temperature zero under a fixed JSON contract that asks only for the prompt span, the reference index, and the requested region of each person; it sees the ordered references and the prompt, never the target regions or any annotation, and its prompt instructs it to read the index from the text rather than from reference appearance. The instruction template is recorded in the released configuration. Algorithm 1 states the complete inference procedure summarized in Section 3.

Algorithm 1 AddrBind inference

Input: prompt x, ordered references $\mathcal{I}$, model p_θ, initial noise z, and shift bound D
Output: generated image Y

Plan compilation and layout observation
$\widehat{\mathcal{Z}} \leftarrow \textsc{Plan}(x, \mathcal{I})$
$(\mathcal{A}, G^0) \leftarrow \textsc{Compile}(\widehat{\mathcal{Z}})$ ▷ validate and form G^0
if $\mathcal{A} = \varnothing$ **then**
  return $\textsc{Generate}(p_\theta, z, x, \mathcal{I})$ ▷ native output
end if
$\overline{\mathrm{Attn}} \leftarrow \textsc{Scout}(p_\theta, z, x, \mathcal{I})$ ▷ observe native layout
$\rho_{sq} \propto G^0_{sq} \overline{\mathrm{Attn}}(q, \sigma_s)$ ▷ fuse plan and layout
$u_q \leftarrow \textsc{Confidence}(\rho_{\cdot q})$ ▷ entropy-based attenuation
$\Gamma^c_{qr} \leftarrow u_q \sum_{s \in \mathcal{A}} \rho_{sq} \mathbb{1}[r = r_s]$ ▷ subject → reference
Budget-conserving execution
$Y \leftarrow \textsc{Generate}(p_\theta, z, x, \mathcal{I}, \mathrm{Route})$ ▷ same initial noise
return Y

procedure Route(logits at head h and canvas query q)
  $\alpha^0_{hq} \leftarrow \textsc{ReferenceSplit}(\text{logits}, \mathcal{I})$ ▷ native reference split
  $\lambda_{hq} \leftarrow \textsc{FeasibleScale}(\alpha^0_{hq}, \Gamma^c_q, u_q, D)$ ▷ largest feasible scale
  $\widehat{\alpha}_{hq} \leftarrow (1 - \lambda_{hq} u_q)\alpha^0_{hq} + \lambda_{hq}\Gamma^c_q$ ▷ conserve reference budget
  $\delta_{hqr} \leftarrow \log(\widehat{\alpha}_{hqr} / \alpha^0_{hqr}),\ r \in \{1, \ldots, R\}$ ▷ blockwise logit shift
  return logits with δ_{hqr} added to every key in reference block r
end procedure

C Evaluation-set construction

Table 6 records how each benchmark's task pool narrows from the released split to the set this paper evaluates on.

D Behavior across cardinalities

This section complements the main $K = 2/3$ evaluation with results at $K = 4/5$ across supported backbones and per-K diagnostics on FLUX.2-klein-4B. Table 7 reports slot accuracy and unfilled rates on the held-out $K = 4/5$ CogCanvas tasks. OmniGen2 is not evaluated because these tasks exceed its five-reference input limit, and Qwen-Image-Edit-2511 is evaluated only at $K = 2/3$ in Table 2.

At $K = 4/5$, AddrBind improves slot accuracy by $+.350$ on FLUX.2-klein-4B and $+.340$ on FLUX.2-klein-9B, while their unfilled rates change only from .046 to .059 and from .060 to .061, respectively. On FLUX.2-klein-4B, the learned plan reaches .618 slot accuracy compared with

Table 6: Evaluation-set construction. "Eligible" denotes an unambiguous prompt-level reference assignment; "Pre-audit" is the frozen pool passed to the reference-image audit, and "Final" the tasks the audit retains. K is the number of target identities and R that of input references. The 48 CogCanvas development examples are excluded and $K = 4/5$ are held out. The CogCanvas audit excludes no rows at any K, with $R = 2K + 1$: K people, K objects, and one shared background. The MultiBanana audit retains 491 of its 645 pre-audit tasks.

Dataset split	K	R	Released	Eligible	Pre-audit	Final
CogCanvas test_2p	2	5	548	511	487	487
CogCanvas test_3p	3	7	356	346	322	322
CogCanvas test_4p	4	9	212	203	203	203
CogCanvas test_5p	5	11	156	149	149	149
CogCanvas total	2–5	5–11	1,272	1,209	1,161	1,161
MultiBanana	2–7	2–8	2,229	705	645	491

Table 7: Results on the held-out $K = 4/5$ split (352 tasks, with $R = 9$ and $R = 11$), using the same scorer and paired cluster bootstrap as the main evaluation. †Not evaluated because $R = 9/11$ exceeds OmniGen2's five-reference limit. *Not evaluated at $K = 4/5$.

	Slot accuracy			Unfilled ↓		
Backbone	Native	AddrBind	AddrBind − Native	Native	AddrBind	Tasks
FLUX.2-klein-4B	.267	**.618**	+.350 [+.322, +.379]	.046	.059	352
oracle plan	—	*.637*	+.369 [+.340, +.401]	—	*.058*	352
FLUX.2-klein-9B	.304	**.644**	+.340 [+.310, +.373]	.060	.061	352
OmniGen2†	not evaluated: $R = 9/11$ exceeds its five-reference limit					
SDXL + IP-Adapter	.006	.022	+.015 [+.006, +.027]	.954	.930	352
Qwen-Image-Edit-2511*	not evaluated at $K = 4/5$					

.637 for the oracle plan. SDXL + IP-Adapter remains dominated by missing instances: Native and AddrBind achieve .006 and .022 slot accuracy, with unfilled rates of .954 and .930.

D.1 PER-K DIAGNOSTICS ON FLUX.2-KLEIN-4B

The FLUX.2-klein-4B row of Table 7 pools the two cardinalities onto one number; the diagnostics below separate them again and add the quantities that decide where the pooled gain comes from. They cover the 809 tasks at $K = 2/3$ and the 352 tasks at $K = 4/5$ behind the FLUX.2-klein-4B pair of Table 2.

Assignment gains across K. Native slot accuracy falls from .695 at $K = 2$ to .474, .270, and .264 at $K = 3, 4, 5$, while AddrBind reaches .813, .766, .653, and .571 on the same tasks; exact-assignment accuracy moves from .542 to .667, .155 to .460, .010 to .261, and .000 to .107. The paired slot-accuracy gains are +.118, +.292, +.383, and +.306, respectively. The learned plan remains close to the oracle plan, with its largest gap, .027 [.013, .042], at $K = 4$.

Cost and compensating gains. Routing raises *unfilled*, the fraction of requested seats with no matched face, from .081 to .119 at $K = 2$ and from .037 to .071 at $K = 3$; the changes at $K = 4/5$ are smaller. Three diagnostics move the other way. *Duplicate fraction*, the fraction of seat-matched face pairs above the identity threshold, falls from .172 to .101, from .268 to .181, and from .252 to .183 at $K = 2$, 4, and 5. *Excess*, the number of detected faces beyond the matched seats, falls from 1.267 to 1.106, from 2.882 to 1.970, and from 2.342 to 1.510 at $K = 3$, 4, and 5. *Route-gap*, the mean margin by which a matched face's best-matching reference outscores the reference named by the prompt over filled seats, falls from .041 to .007, from .089 to .020, from .133 to .042, and from .115 to .052. Thus, the assignment gains at lower K coincide with more unfilled slots, while filled slots match the requested references more closely across all cardinalities.

E CogCanvas quality and prompt-following metrics

CogCanvas defines six quality and prompt-following metrics but does not release its evaluation code, so the six columns of Table 8 are our own implementations: identity similarity from the same frozen ArcFace pipeline that produces slot accuracy (Deng et al., 2019), subject and background similarity from DINOv2 ViT-L/14 (Oquab et al., 2024), text alignment from CLIP ViT-L/14 (Radford et al., 2021), aesthetic score from the LAION v2 linear head on that same backbone (Schuhmann et al., 2022), and attribute VQA answered independently by two open vision–language models, `Qwen3.8-27B` and `GLM-4.6V-Flash` (GLM Team, 2025).

The attribute-VQA column is the attribute component of a dependency-weighted question decomposition: each prompt is decomposed into yes/no questions over entities, attributes and relations, and an attribute question is scored only when the entity question it depends on was answered yes. How many images contribute to that column therefore depends on the method (152 to 326 of 641 across the rows of the table), so its means are not over a fixed image set. The values shown are the primary judge's; the second judge, which shares no model with the planner, is more generous in level but orders these rows the same way. Table 9 below gives the paired AddrBind − native differences on every backbone. Together the four tables give the complete quality picture behind the summary in Section 4.3: Table 8 reports absolute values, Table 9 the paired changes across all five backbones, and Tables 10 and 11 the paired and absolute prompt-content scores from the two independent judges.

Table 8: CogCanvas quality and prompt-following metrics, our implementation, for the FLUX.2-klein-4B pair and the SDXL block of Table 2. The layout-control rows receive target regions and use separable IP-Adapter branches; native generation and AddrBind do not. Best in bold with a gray background within each block; the oracle-plan row and the released systems are not marked.

Method	ID-Sim ↑	DINO-Sim ↑	Attr-VQA ↑	CLIP-T ↑	Aesth. ↑	BG-Sim ↑
FLUX.2-klein-4B, input-matched pair						
Native	.240	.261	**.337**	**.254**	**5.62**	**.320**
AddrBind	**.467**	**.304**	.217	.239	5.47	.087
AddrBind, oracle plan	.471	.303	.214	.239	5.46	.086
SDXL with separable IP-Adapter branches						
Native	.048	.293	.051	.204	4.66	.053
Static GT-region routing	.040	.386	.092	.238	4.94	**.128**
BoxDiff	.048	.293	.067	.204	4.66	.052
DenseDiffusion	.046	.322	.068	.214	4.63	.072
MultiDiffusion	.031	**.409**	**.152**	**.292**	**5.32**	.093
Regional prompting	.044	.373	.094	.234	4.86	.104
AddrBind	**.091**	.344	.089	.158	4.53	.028
Released end-to-end systems, original backbones						
MS-Diffusion	.222	.369	.074	.245	5.21	.026
SIGMA-GEN	.653	.313	.087	.193	5.32	.038

Paired quality changes across backbones. Table 9 compares each AddrBind arm with its paired native arm on identical tasks, seeds, and sampling settings. Identity similarity rises wherever the change resolves, and the two columns that fall consistently are CLIP alignment and background similarity, both by small amounts. The face-rate column is a control rather than a result: it reports how many faces the identity metric is computed over, and on SDXL native generation frequently omits faces, so that column moves far more there than on the joint-attention backbones.

Paired prompt-content changes. Table 10 reports paired score differences between AddrBind and native on the same CogCanvas tasks. The two judges agree on direction in every row: reference-slot accuracy improves on all five backbones, object relations decline on all five, and the declines are sharpest on OmniGen2 and SDXL.

Table 9: **Paired quality changes on CogCanvas.** Entries are AddrBind minus native. Columns report face rate, identity similarity, DINO similarity, CLIP alignment, aesthetic score, and background similarity. A dash denotes an interval containing zero.

Backbone	ΔFace	ΔID	ΔDINO	ΔCLIP	ΔAesth.	ΔBG
FLUX.2-klein-4B	—	+.227	+.042	−.015	−.154	−.233
FLUX.2-klein-9B	+.001	+.245	+.035	−.016	−.126	−.269
OmniGen2	+.004	**+.414**	−.028	−.087	—	−.049
SDXL + IP-Adapter	**+.194**	+.048	+.051	−.046	−.128	−.025
Qwen-Image-Edit-2511	—	—	−.034	−.008	—	−.082

Table 10: **Assignment gains and paired prompt-content changes.** Ref. slot measures identity-specific assignment; the remaining columns report count, slot, and object scores from two independent VLM judges.

		Primary VLM judge			Secondary VLM judge		
Backbone	ΔRef. slot	ΔCount	ΔSlot	ΔObject	ΔCount	ΔSlot	ΔObject
FLUX.2-klein-4B	+.187	−.004	−.035	−.076	−.010	−.065	−.057
FLUX.2-klein-9B	+.191	−.135	−.023	−.044	−.158	−.011	−.034
OmniGen2	+.538	−.187	+.079	−.467	−.310	+.212	−.419
SDXL + IP-Adapter	+.299	+.171	+.211	−.250	+.173	+.083	−.323
Qwen-Image-Edit-2511	+.079	+.019	+.009	−.137	−.013	+.003	−.067

Absolute placement-exam scores. Table 11 reports the absolute scores underlying the paired contrasts in Table 10. Each contrast uses the same tasks, seed, and sampling configuration and is resampled over 29 background clusters with 4,000 bootstrap draws. OmniGen2 is evaluated on the 487 $K = 2$ tasks admitted by its reference limit, and Qwen-Image-Edit-2511 is paired on the tasks where its conserving write passed a pre-registered numerical criterion. The absolute numbers confirm that the paired declines are small: on FLUX.2-klein-4B the primary judge moves from .689 to .621 while reference-slot accuracy gains .187, and the learned plan matches the oracle plan to within .002. The one large absolute drop is on OmniGen2, whose object-relation score falls by .173 under the primary judge.

Table 11: **Absolute placement-exam scores on CogCanvas.** Scores are dependency-weighted over rule-generated count, slot, object, and auxiliary questions. Native and AddrBind use paired task sets.

	Primary judge		Secondary judge	
Backbone / method	Native	AddrBind	Native	AddrBind
FLUX.2-klein-4B	.689	.621	.783	.723
oracle plan	—	*.619*	—	*.724*
FLUX.2-klein-9B	.729	.650	.802	.740
OmniGen2	.560	.387	.667	.503
SDXL + IP-Adapter	.235	.258	.441	.414
Qwen-Image-Edit-2511	.392	.342	.533	.526

F MULTIBANANA: THE AUDITED PAIR IN FULL

The pair uses FLUX.2-klein-4B with the released executor, seed 7, 28 steps, and the frozen v5 planner output on the 491 tasks retained by the reference audit. Of 645 prompt-level candidates, 113 were excluded because a reference shows more than one face and 58 because a reference shows none at either detection scale; these criteria overlap for 17 tasks. Every routed call satisfies conservation (maximum residual 1.0×10^{-6}, floor never active).

Identity presence is the fraction of named persons whose best face match exceeds a threshold τ_P calibrated from the audited reference embeddings alone under a matched null at $\alpha = .05$ ($\tau_2 = .257$ to $\tau_7 = .339$); set recall is the fraction of tasks in which every named person is present. Per-seat and exact placement are defined only on the 146 tasks (117 routed) whose prompt gives a person

an absolute canvas relation; among the remaining tasks, 130 prompts have complete but exclusively relative spatial language.

The judge is the benchmark's own Qwen3-VL-8B script (Qwen Team, 2025) with its regexes, 1–10 per dimension; its means are over the images whose verdict parsed (452 native, 451 AddrBind) and its differences are paired over the 436 tasks where both did (132 routed). Intervals are task-level bootstrap, 4,000 draws, seed 20260826. On the 339 non-routed tasks the two images are byte-identical, so every image metric differs by exactly zero there, and the judge's Average on those identical inputs, $-.035$ $[-.263, +.203]$ over 304 pairs, is the scale of its own noise.

Routing is concentrated in one family: of the 134 prompts that place a person relative to an object, 125 are routed; of the other 357 (background, style, local-edit, replacement and one addition prompt), 27 are.

Table 12: MultiBanana, FLUX.2-klein-4B, AddrBind against native. The four identity and placement rows on the left restate the FLUX.2-klein-4B pair of Table 3 for a single backbone and add the paired difference; the two face-count rows, the routed subset, and the judge rows appear only here. Left: all 491 audited tasks. Right: the 152 tasks the compiler routed (on the other 339 the AddrBind image is the native image). Δ is the paired difference, and * marks a task-level bootstrap interval excluding zero; the interval bounds are omitted for width. Placement rows use the 146 (117 routed) tasks with an absolute region; judge rows are the benchmark's own Qwen3-VL-8B on a 1–10 scale.

	All audited tasks (491)			Routed subset (152)		
Metric	Native	AddrBind	Δ	Native	AddrBind	Δ
Identity and placement, frozen scorer						
Presence per person ↑	.179	.195	$+.015^*$	.186	.235	$+.050^*$
Set recall ↑	.220	.240	$+.021^*$	.235	.302	$+.066^*$
Faces per image	3.11	3.09	$-.020$	3.37	3.30	$-.066$
Excess faces ↓	.644	.629	$-.014$	.796	.750	$-.046$
Per-seat ↑ (146 / 117 tasks)	.468	.537	$+.069^*$	.532	.619	$+.087^*$
Exact ↑ (146 / 117 tasks)	.281	.363	$+.082$	.308	.410	$+.103$
Benchmark judge, 1–10 (436 / 132 pairs)						
Instruction Alignment	4.50	4.39	$-.12$	4.20	3.63	$-.58^*$
Reference Consistency	4.56	4.52	$-.05$	4.27	3.90	$-.36$
Background–Subject Match	3.68	3.53	$-.18$	3.19	2.82	$-.38$
Physical Realism	4.27	4.17	$-.14$	3.60	3.55	$-.08$
Visual Quality	5.60	5.48	$-.13$	5.33	5.04	$-.25$
Average	4.52	4.42	$-.13$	4.12	3.79	$-.33$

On the routed subset the benchmark's judge scores Instruction Alignment $-.58$ $[-1.01, -.15]$ on a ten-point scale, with its other four dimensions negative in point estimate and not separable from zero.