FOLLOW ME IF YOU CAN: Auditing Understanding-to-Generation Transfer

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding-only training can change image generation through shared parameters, but the resulting scores support different claims about transfer. We develop an audit framework separating source validity, target influence, and attribute correspondence. Fixed-checkpoint image and instruction controls test the source gain; paired trained and frozen states measure generation influence; conditioning swaps probe the requested value expressed by the update. We characterize a blind spot of norm-matched swaps: changes along a shared direction yield identical interventions even when their amplitudes depend on the prompt. Across six unified architectures, counting and spatial experiments reveal different answer–generation response patterns. In a focused counting audit, a $+4.98$ percentage-point answer gain becomes $+0.37/-0.85$ under digit requests at fixed weights. A standard low-rank adaptation (LoRA) reference also improves answers without a consistent generation-score gain. Joint readouts distinguish recipient improvement from donor-value following: a $+4.88$ percentage-point recipient gain accompanies a -0.59 donor change. Mixed same-budget prompt-contrast calibration bounds inference from non-following. The audit ties each interpretation of transfer to the comparison that supports it.

1 INTRODUCTION

A capability claim connects a name to observed behavior. Assessing that claim requires examining the conditions under which we observe the behavior. Unified multimodal models answer questions about images and generate images through shared parameters (Pan et al., 2025; Lin et al., 2025; Deng et al., 2025). An understanding-only update can therefore change both tasks. Establishing capability transfer requires evidence about what the answer gain measures and which requested attribute changes in generation.

Endpoint changes leave the update's visual content underdetermined. Answer gains can combine visual recognition, answer priors, and response form. Generation scores can combine object presence with count or relation agreement. Paired checkpoints, prompts, and decoding randomness isolate the update's score effect, while attribution to a learned visual attribute remains open.

We distinguish three claims. *Source validity* asks whether the answer gain tracks the intended visual attribute across image and instruction controls. *Target influence* asks whether the same trained state changes generation under paired prompts and noise. *Attribute correspondence* asks whether the resulting image responds to a particular requested value. A donor intervention replaces the recipient's conditioning change with one elicited by an attribute-changing prompt. Joint readouts test requested-value sensitivity at that interface and magnitude. An improvement in following the recipient request can coexist with unchanged donor readouts.

We evaluate counting and spatial endpoints and donor replacements across six unified architectures. Removable square cores and a separate additive-LoRA reference provide two update parameterizations. Each contrast compares paired states within its training cohort. Figure 1 locates the training and conditioning interventions.

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

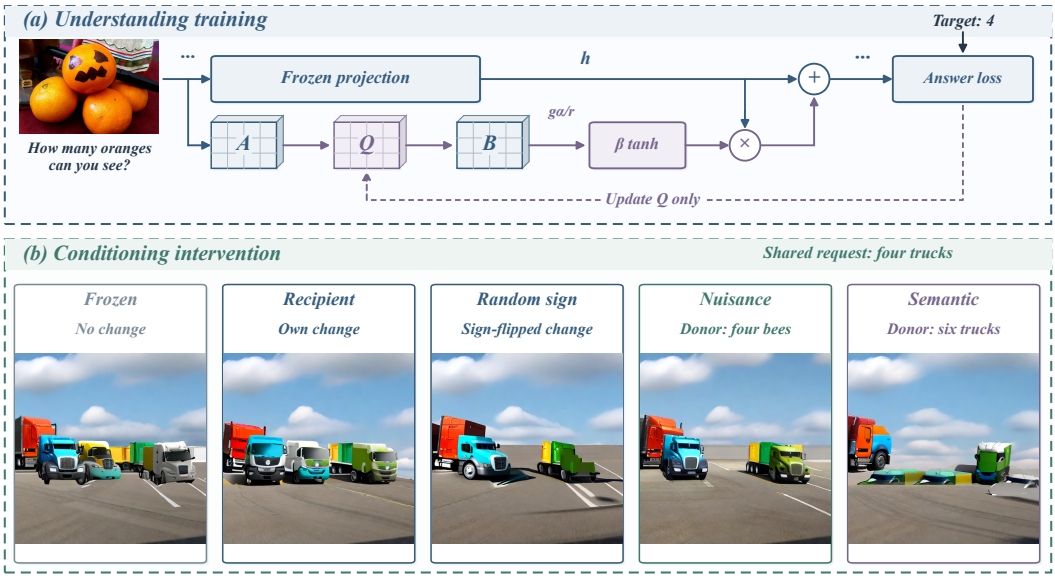

Figure 1: **Training and conditioning probe two stages of transfer.** (a) Only Q learns. (b) Decoder calls share the recipient request, frozen weights, and noise; donor changes match the recipient norm (Eq. 3).

Our audit changes the interpretation of observed gains. In a focused counting audit, digit-only instructions change a $+4.98$ pp answer advantage to $+0.37/-0.85$ pp at fixed weights. Across six architectures, paired endpoint changes establish generation influence with different source–target relationships. Joint readouts separate recipient improvement from donor-value following. A directional characterization of norm matching explains which update differences a swap retains, and same-budget calibration bounds its interpretation.

2 Related Work

Shared architectures and transfer. Unified models connect visual question answering to synthesis through autoregressive tokens, flow matching, or diffusion (Chen et al., 2025b; Xie et al., 2026; Yang et al., 2026b; Swerdlow et al., 2025; Xin et al., 2025); other systems pass semantic representations to diffusion decoders (Pan et al., 2025; Lin et al., 2025; Hu et al., 2024). These designs establish paths through which an understanding update can affect generation. HBridge selectively connects heterogeneous experts (Wang et al., 2026a), while Uni-X separates modality-specific ends to reduce gradient conflict (Hao et al., 2026). Transferability compares capability-specific LoRA updates and reports positive generation transfer under some model and task conditions (Kang et al., 2026). UniSandbox tests reasoning and factual knowledge transfer (Niu et al., 2025); native-model work studies when shared knowledge and specialized paths yield synergy (Wu et al., 2026). Cross-branch steering tests transferable semantic directions (Wang & Li, 2026); reasoning-guided generation tests whether images follow textual reasoning (Yang et al., 2026a). UniG2U-Bench evaluates generation-to-understanding effects across tasks (Wen et al., 2026). Joint objectives also add generation feedback (Yan et al., 2026; Su et al., 2026). Beyond unified generation, selective distillation assigns heterogeneous teachers different supervision roles in clinical prediction and audio-visual localization (Xu et al., 2026a;b). Our audit complements endpoint comparisons with fixed-checkpoint image and instruction controls and dual readouts of requested attributes.

Intervention instruments. Low-rank adaptation trains small updates to a frozen backbone (Hu et al., 2021). Related approaches share cross-modal adapters (Lu et al., 2024), compose task-specific LoRAs (Huang et al., 2023a), or reduce their interference (Zhang et al., 2025). Fixed-map square cores also appear in LoRA-XS; initialized bases and multiplicative operators parameterize updates differently (Bałazy et al., 2025; Meng et al., 2024; Lian et al., 2022; Liu et al., 2022). Our fixed-map square core serves as a removable state at the same projection sites for answering and generation.

A standard additive-LoRA reference checks whether the central source/target contrast depends on that parameterization. Causal tracing and activation patching use internal interventions to study model behavior (Meng et al., 2022; Zhang & Nanda, 2024; Wu et al., 2024). We intervene at the generator's conditioning interface. Interventions can produce representations outside a model's natural activation distribution (Grant et al., 2026). We control realized perturbation magnitude, but this does not establish distributional fidelity. Interactions among mediators further complicate pathway attribution (Vaidyanathan et al., 2026). Our donor contrast concerns the implemented carrier replacement; it does not identify the pathways that mediate natural understanding-to-generation transfer.

Evidence and attribute measurement. Complementary-image questions expose answer priors in visual question answering (Goyal et al., 2017). We pair image-dependence tests with answer-instruction changes on fixed checkpoints. Generation benchmarks measure object alignment and question-based faithfulness (Ghosh et al., 2023; Hu et al., 2023). T2I-CompBench tests compositional alignment (Huang et al., 2023b); its successor supplies our generation scores (Huang et al., 2025). SpatialGenEval extends spatial evaluation to information-dense scenes (Wang et al., 2026b). We use controlled count and relation requests to isolate attribute values. Scoring the same output against recipient and donor requests tests value correspondence under a specified intervention.

3 METHOD

We audit an understanding-trained update with fixed-checkpoint source controls, paired state removal, and dual-readout conditioning swaps (Figure 1). Square cores make the update removable; a separate additive-LoRA reference checks another parameterization.

3.1 CORE PARAMETERIZATION

For projection ℓ, let $h_\ell = \mathrm{Base}_\ell(x)$ include the pretrained weight and bias. Frozen maps $A_\ell \in \mathbb{R}^{r\times d_{\mathrm{in}}}$ and $B_\ell \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ surround a trainable square core $Q_\ell \in \mathbb{R}^{r\times r}$. The projection outputs

$$\widetilde{h}_\ell = h_\ell + \rho_\ell(x; Q_\ell), \qquad \rho_\ell = h_\ell \odot \beta \tanh\Big(g_\ell \frac{\alpha}{r} B_\ell Q_\ell A_\ell x\Big). \tag{1}$$

Here $\beta > 0$ bounds the residual, $\odot$ multiplies coordinates, $g_\ell = 0.5$, and $\alpha = r = 64$. Frozen A_ℓ and B_ℓ use normalized, high-energy rows and columns of the pretrained weight. Understanding answers train only $q_s = \{Q_\ell\}_\ell$, starting at zero. For each source task, answering and generation share these cores and maps; removal recovers the pretrained projection.

The bound $|\rho_{\ell,j}| \le \beta |h_{\ell,j}|$ controls local rescaling; this does not establish a bound on generation-trajectory differences. Frozen access maps make Q removable. Additive LoRA learns both maps as a separate parameterization reference (Hu et al., 2021; Bałazy et al., 2025).

3.2 STATE REMOVAL AND RETENTION

For understanding and generation metrics M_U and M_G, we measure

$$\Delta_U = M_U(q_s) - M_U(0), \qquad \Delta_G = M_G(Y^{\mathrm{act}}) - M_G(Y^{\mathrm{rem}}), \tag{2}$$

where active and removed outputs share the prompt, initial noise, solver, and inference randomness. A negative generation change indicates harm on the measured endpoint.

Source development data select depth groups before target evaluation. Equal-budget random groups, leading singular components and their complement, and core negation probe state dependence. Appendices A.19 and A.5 define these influence contrasts and realized magnitudes along divergent inference paths.

3.3 DONOR REPLACEMENT

At a static conditioning interface, let $c_q^+(p)$ and $c_q^-(p)$ denote conditional and unconditional carriers. The trained conditional change is $\delta(p) = c_{q_s}^+(p) - c_0^+(p)$. For recipient p_r and donor p_d, inject

$$\widehat{c}^+(p_r; p_d) = c_0^+(p_r) + \frac{\|\delta(p_r)\|_F}{\|\delta(p_d)\|_F}\delta(p_d), \qquad \widehat{c}^-(p_r; p_d) = c_0^-(p_r). \tag{3}$$

The scale $\kappa = \|\delta(p_r)\|_F / \|\delta(p_d)\|_F$ matches the donor change to the recipient norm and requires positive energy in both changes. Native-precision checks compare the realized change magnitude with the recipient reference.

The carrier can be a visual-token state, conditioning embedding, image-token logit, or conditional velocity. We preserve recipient noise and the unconditional branch. Norm matching equates perturbation size within an interface, not the expressive power of different interfaces or the naturalness of the hybrid state. For dynamic carriers, recipient and donor passes share the current image state and generation step. Appendix A.13 gives native implementations, alignment, and magnitude gates.

Setting $p_d = p_r$ gives *Own*, the conditional-only control. *Swap* substitutes a semantic donor that keeps the objects and changes the count or 2D relation. At fixed recipient request and noise, we report $\Delta_a = M_a(Y^{\text{swap}}) - M_a(Y^{\text{own}})$ for $a \in \{r, d\}$, where M_r and M_d score the recipient and donor requests. Both readouts score each output. Their Own baseline differs from the full-minus-frozen endpoint contrast in Eq. 2.

In exact arithmetic, norm matching gives the swap a directional interpretation. Write $\delta(p) = m(p)v(p)$, with $m(p) = \|\delta(p)\|_F > 0$ and $\|v(p)\|_F = 1$. Equation 3 replaces $m(p_r)v(p_r)$ with $m(p_r)v(p_d)$: it retains the recipient magnitude and substitutes the donor direction. If all changes share a direction, Own and Swap coincide even when their amplitudes vary with the request. The prompt-independent case $\delta(p) = u \neq 0$ is a special instance. Such an update can still affect generation relative to frozen or improve adherence to the recipient request. The swap leaves these same-direction effects untested. Donor copying can also reflect conditioning replacement. Following is therefore neither necessary nor sufficient for broader capability transfer.

Nuisance donors change objects while preserving the attribute. Unconditional-only compares branch effects, full-state replay checks interface reconstruction, and random-sign changes probe direction dependence. Comparing these controls with the semantic donor tests attribute specificity (Appendix A.13). The embedding-interface counting audit also replaces the learned donor change with the frozen prompt contrast $c_0^+(p_d) - c_0^+(p_r)$, scaled to the same recipient magnitude. A prompt nuisance contrast changes the object at that budget. These controls calibrate local carrier sensitivity; they do not identify the pathway of a learned attribute (Appendix A.12).

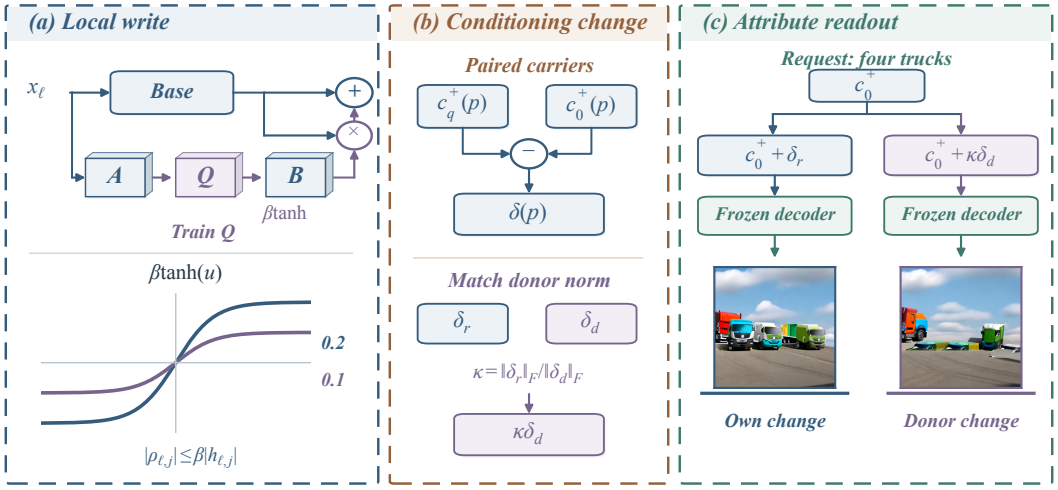

Figure 2: **From local write to attribute readout.** (a) Q_ℓ writes a bounded residual. (b) κ matches the donor change to the recipient norm. (c) Own and Swap share the request, frozen decoder, and noise.

4 EXPERIMENTS

4.1 MODELS AND TRAINING

Backbones and supervision. Primary reference comparisons use BLIP3-o 4B with diffusion (Chen et al., 2025a), JanusFlow-1.3B with flow matching (Ma et al., 2025), and UniWorld-V1 with joint Qwen/T5 conditioning and FLUX (Lin et al., 2025). DiMOO, Janus-Pro, and Show-o2 extend the endpoint and donor comparisons to six architectures (Appendix A.20). Cores share the rank-64 operator and ordered supervision; Appendix A.1 specifies model-specific sites and budgets.

Tasks and metrics. Counting uses 1,024 held-out HowMany-QA questions with answers 1–4, distributed alongside TallyQA (Acharya et al., 2019). Spatial understanding uses 512 Visual Spatial Reasoning (VSR) judgments (Liu et al., 2023). Generation uses 1,000 T2I-CompBench++ numeracy and 1,000 spatial prompts, half 2D and half 3D (Huang et al., 2025). We report answer accuracy (%) and official generation scores ($\times 100$). Donor counting reports exact-count hit rates, the fraction with unscaled official numeracy score $s = 1$.

Cohorts. Original cores provide instruction, removal, retention, and direction tests. Separate reference cores provide endpoints and donors; additive LoRA checks another parameterization. Appendix A.1 defines the training protocols.

4.2 PAIRED EVALUATION

Generation follows each model's native decoder with shared prompts and paired noise within model. The BLIP3-o interface study resets decoder randomness after conditioning extraction. Greedy understanding decoding follows each model's native answer length; decoding settings and a UniWorld-V1 length check appear in Appendix A.20.

Both original-core audits cross three answer instructions with original, constant-gray, and label-blindly permuted images while retaining questions and reference answers. We repeat these controls on the BLIP3-o reference checkpoints used for endpoints and donors (Appendix A.11). We also evaluate 122 disjoint complementary-image count pairs; both answers must be correct (Goyal et al., 2017). Lexical scoring checks answer-format sensitivity. A post hoc analysis separates generation prompts requesting counts 1–4 from 5–8 (Appendix A.15).

4.3 READOUTS AND ANALYSIS

Each donor experiment evaluates 128 recipient/semantic/nuisance prompt triples per task and run. Conditional-only retains the recipient's own change. Counting requests span 1–6; only 26 pairs have both counts within the supervised answer range 1–4. We report all pairs and post hoc count groups separately (Appendix A.16). Spatial donors change 2D relations: 56 pairs compare distinct directions, 59 compare direction with proximity, and 13 compare evaluator-equivalent proximity expressions. Directional support requires an official score above 0.5; the object-parser check retains 126 spatial pairs (Appendix A.17). Focused source controls, the relation-family breakdown, and the embedding-interface calibration use BLIP3-o. Counting transitions use the donor cohort for each backbone. Tables and appendices retain the cohort identities.

Tables report aggregate scores; transition matrices report joint percentages. Paired changes use unrounded scores. Percentage points describe accuracy and exact-hit changes; generation points use the $100\times$ scale. Training runs are independent replicates, with prompts repeated within run. Understanding development accuracy alone selects source depth groups. Appendix A.21 reports run-level effects and exploratory uncertainty.

5 RESULTS

5.1 SOURCE VALIDITY

Source gains depend on answer elicitation at fixed weights. In the reference source audit, which uses the endpoint and donor checkpoints, digit requests A/B move the counting gain from $+4.98$

to $+0.37/-0.85$ percentage points. True/false requests move the spatial gain from $+1.51$ to $-0.20/+0.10$ (Appendix A.11). The source score therefore supplies an instruction-dependent answer advantage for the same update whose generation effect we measure.

The separate original-core cohort reproduces this dependence. Digit requests move its counting gain from $+5.10$ to $+0.07/-0.76$; true/false requests move the spatial gain from $+1.56$ to $-0.05/+0.15$ (Table 1). On intact images, digit requests improve both states' counting accuracy while largely removing or reversing the trained advantage. Both digit requests also reverse the full-core advantage on 122 natural image pairs (Figure 3). Benchmark and lexical scoring agree on every original-image A/B response, which rules out parser choice as an explanation of that instruction effect.

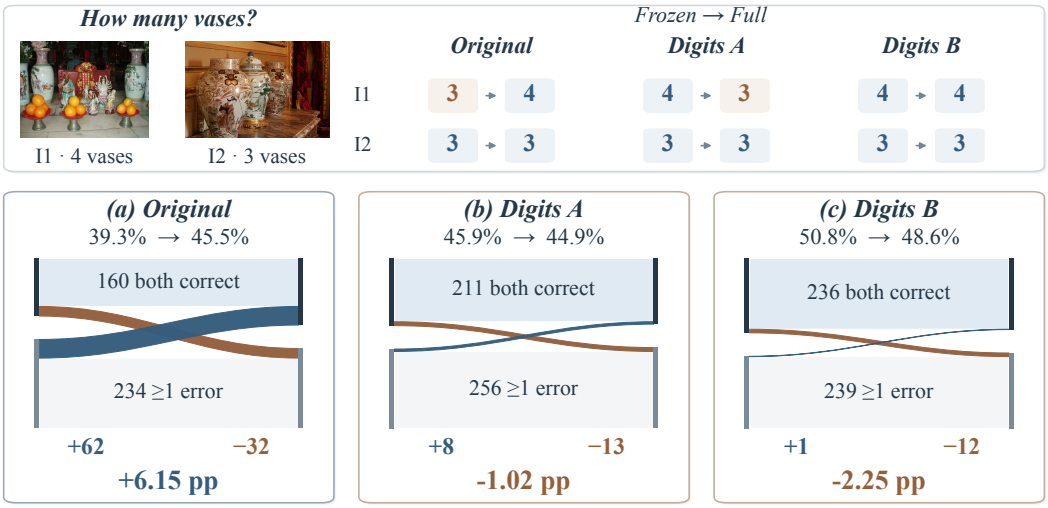

Figure 3: **Digit requests reshape the training advantage.** Top: frozen/full answers to two images; blue/orange marks correct/incorrect. Bottom: both-correct transitions over 122 image pairs and four runs. Ribbon widths encode counts; blue/orange marks gains/losses. Percentages give both-correct rates. A/B request digits (Appendix A.6).

Table 1: **Answer accuracy under instruction and image controls.** Accuracy (%); superscripts give full minus frozen. A/B request digits for counting and true/false for spatial relations.

	Original question		*Answer-only A*		*Answer-only B*	
Image condition	Frozen	Full	Frozen	Full	Frozen	Full
Counting: original	59.18	**64.28**$^{+5.10}$	66.80	**66.87**$^{+0.07}$	69.04	**68.29**$^{-0.76}$
Gray images	22.75	**24.15**$^{+1.39}$	13.96	**14.92**$^{+0.95}$	9.67	**10.72**$^{+1.05}$
Permuted images	3.71	**3.59**$^{-0.12}$	3.32	**3.49**$^{+0.17}$	4.10	**4.44**$^{+0.34}$
Spatial: original	79.88	**81.45**$^{+1.56}$	81.45	**81.40**$^{-0.05}$	83.20	**83.35**$^{+0.15}$
Gray images	5.27	**4.54**$^{-0.73}$	52.15	**52.39**$^{+0.24}$	51.17	**50.59**$^{-0.59}$
Permuted images	43.16	**48.14**$^{+4.98}$	50.20	**50.20**$^{0.00}$	49.41	**49.22**$^{-0.20}$

Image dependence and training advantage follow different patterns. Intact images yield higher answer accuracy than gray or permuted images under every instruction for the original cores. Yet digit-request counting gains grow under these image controls, and the spatial gain on original questions rises from $+1.56$ to $+4.98$ on permuted images. The controls distinguish use of image evidence from attribution of a training gain to the named visual attribute. Across both core cohorts, the visual contribution to the gain remains unresolved.

5.2 Target Influence

Answer and generation changes form several sign patterns across backbones and tasks (Table 2; Figure 5). Counting includes opposite-sign changes and declines at both endpoints. The largest

Table 2: **Understanding and generation across backbones.** Answer accuracy (%) and generation score (×100); superscripts give full minus frozen.

	Counting		*Spatial relations*			
Backbone / core	Answer	Gen.	Answer	Gen. 2D	Gen. 3D	Gen. Avg.
BLIP3-o: Frozen	59.18	39.54	79.88	14.07	26.61	20.34
BLIP3-o: **Full core**	**64.16**$^{+4.98}$	**38.65**$^{-0.89}$	**81.40**$^{+1.51}$	**14.33**$^{+0.26}$	**27.16**$^{+0.55}$	**20.74**$^{+0.41}$
JanusFlow: Frozen	66.50	49.46	65.04	25.59	27.59	26.59
JanusFlow: **Full core**	**64.75**$^{-1.76}$	**49.90**$^{+0.44}$	**62.84**$^{-2.20}$	**25.02**$^{-0.57}$	**28.13**$^{+0.54}$	**26.58**$^{-0.01}$
UniWorld-V1: Frozen	63.28	61.92	82.81	32.54	39.35	35.94
UniWorld-V1: **Full core**	**61.21**$^{-2.08}$	**61.80**$^{-0.12}$	**83.74**$^{+0.93}$	**32.51**$^{-0.03}$	**39.29**$^{-0.06}$	**35.90**$^{-0.04}$
DiMOO: Frozen	31.45	63.85	49.02	45.44	39.13	42.29
DiMOO: **Full core**	**30.86**$^{-0.59}$	**64.17**$^{+0.32}$	**49.02**$^{0.00}$	**47.05**$^{+1.61}$	**39.61**$^{+0.48}$	**43.33**$^{+1.04}$
Janus-Pro-1B: Frozen	63.38	28.34	60.35	11.39	24.36	17.87
Janus-Pro-1B: **Full core**	**63.38**$^{0.00}$	**26.18**$^{-2.16}$	**61.33**$^{+0.98}$	**13.89**$^{+2.50}$	**23.61**$^{-0.75}$	**18.75**$^{+0.87}$
Show-o2-1.5B: Frozen	50.98	60.37	67.58	25.60	37.69	31.64
Show-o2-1.5B: **Full core**	**61.62**$^{+10.64}$	**60.13**$^{-0.24}$	**69.92**$^{+2.34}$	**24.91**$^{-0.69}$	**37.19**$^{-0.49}$	**31.05**$^{-0.59}$

observed counting answer gain (+10.64 pp) accompanies a generation decline (−0.24 points). Generation scores also change at unchanged answer accuracy, and task responses differ within backbones. These paired native-inference contrasts establish target influence with unresolved visual content. Appendix Table 12 separates exact-count hits for requests of 1–4 and 5–8 objects, exposing variation within the aggregate score.

Figure 4 shows paired outputs from the separate original-core cohort; their automatic scores alone leave the changed visual property unresolved.

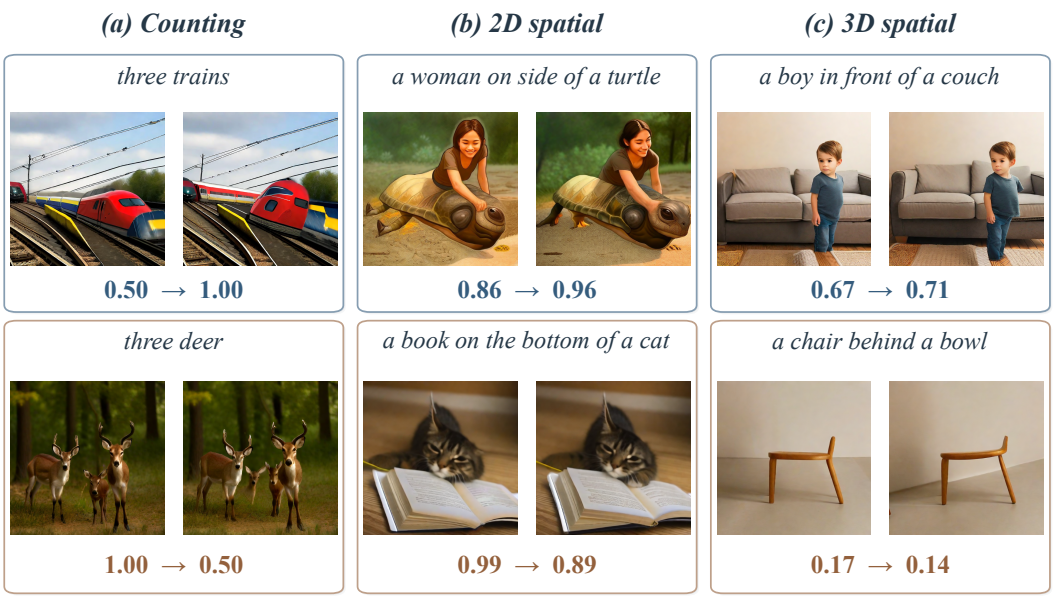

Figure 4: **Paired images reveal what a score change measures.** Frozen (left) and full (right) share the prompt and noise. Each task shows a score increase and decrease; Appendix A.22 gives the selection rule and full gallery.

A source decrease still permits measurement of target influence; a claim of failed transfer requires evidence of source acquisition. These comparisons characterize the training protocol across backbones, with source learning and native carriers varying together.

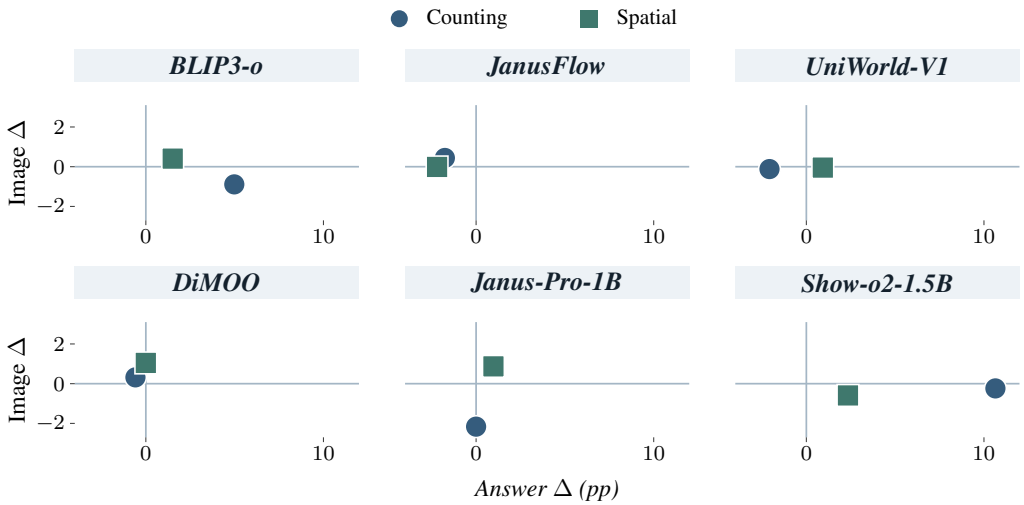

Figure 5: **Understanding and generation respond differently across backbones.** Full minus frozen: answer accuracy (pp) versus generation score ($\times 100$). Circles: counting; squares: spatial. Points show the aggregate changes in Table 2.

5.3 Attribute Correspondence

Aggregate counting responses span all four combinations of recipient and donor signs (Table 3). Joint readouts therefore distinguish recipient benefit from following the inserted count. Counting means show recipient gains alongside donor declines ($+4.88/-0.59$ for BLIP3-o and $+2.73/-2.73$ for JanusFlow), while both readouts rise slightly in UniWorld-V1 ($+0.39/+0.59$). Changes use exact-hit percentage points. JanusFlow's donor decrease holds in every run; the small UniWorld-V1 means represent few net hits, with signs varying across runs. Calibration bounds the interpretation of non-following (Section 6).

Table 3: **Attribute readout under donor replacement.** Exact-count hits (%) and spatial scores ($\times 100$) on 128 pairs. Own retains the recipient change; Swap inserts a semantic donor change. Superscripts give Swap minus Own.

	Recipient		*Donor*	
Backbone	Own	Swap	Own	Swap
Counting				
BLIP3-o	16.21	**21.09**$^{+4.88}$	9.77	**9.18**$^{-0.59}$
JanusFlow	21.09	**23.83**$^{+2.73}$	13.67	**10.94**$^{-2.73}$
UniWorld-V1	52.54	**52.93**$^{+0.39}$	9.57	**10.16**$^{+0.59}$
DiMOO	53.91	**53.12**$^{-0.78}$	11.72	**6.25**$^{-5.47}$
Janus-Pro-1B	16.41	**14.84**$^{-1.56}$	5.47	**9.38**$^{+3.91}$
Show-o2-1.5B	46.88	**46.09**$^{-0.78}$	7.81	**7.03**$^{-0.78}$
Spatial relations				
BLIP3-o	11.06	**11.54**$^{+0.47}$	7.86	**9.62**$^{+1.76}$
JanusFlow	19.47	**19.05**$^{-0.42}$	13.49	**13.00**$^{-0.50}$
UniWorld-V1	28.18	**28.38**$^{+0.20}$	20.90	**20.49**$^{-0.41}$
DiMOO	43.30	**41.93**$^{-1.37}$	22.32	**22.27**$^{-0.05}$
Janus-Pro-1B	10.21	**12.36**$^{+2.15}$	10.78	**10.20**$^{-0.57}$
Show-o2-1.5B	24.33	**24.89**$^{+0.55}$	26.26	**25.68**$^{-0.58}$

6 ANALYSIS

Update controls. In the original-core cohort, state controls separate retention of answer gains from generation effects. Full-minus-frozen counting changes have opposite answer and generation signs in each training run (Table 14). Source-selected depth groups retain 1.76 more answer points than equal-budget random groups, with a $+0.01$ generation-score difference. Against fixed $-Q$, the full core gains 3.05 answer points and loses 1.27 generation points; both directions hold in every run. Leading-component and complementary states show further endpoint interactions (Appendix A.19). These controls locate dependence on the trained state.

LoRA endpoint changes. A standard additive-LoRA reference gains answer accuracy under both answer requests without a consistent image-score gain. Mean answer accuracy rises by 7.28 points with a digit request and 4.88 with a word request, while strict word-format compliance does not improve (Appendix Table 4). Paired image scores vary in direction and average -0.03 points (Appendix A.7). This cohort trains a different operator and many more parameters than the square core, so it tests parameterization dependence; it does not rank adaptation methods.

The LoRA image controls narrow the source interpretation further. On reserved images, normally paired training beats shuffled-image training, but the gap includes damage caused by the shuffled arm. Normal training also improves answers over frozen under corrupted evaluation images (Appendix A.8). The pairing contrast supports sensitivity to image supervision, but these controls do not quantify the visual-counting contribution to the answer gain.

Outcome structure. Recipient and donor exact hits form mutually exclusive outcomes. One counting comparison changes recipient and donor hit rates by $+4.88$ and -0.59 percentage points (Table 11). Most of its new recipient hits previously matched neither count. This gain improves the unchanged request; dual scoring distinguishes it from following the inserted value.

Relation groups distinguish a spatial score gain from directional support. The focused spatial audit gains 1.76 donor-score points overall but only $+0.01$ on 56 direction-only pairs. The aggregate gain therefore arises chiefly outside those pairs. One of 224 direction-only outputs exceeds the official 0.5 donor threshold. Lower scores leave weaker geometric evidence unresolved. The other 72 pairs involve proximity or evaluator-equivalent expressions. Figure 8 shows response magnitudes; Appendix A.17 retains relation groups and parser checks.

Interface sensitivity. At the learned counting budget, the frozen prompt contrast raises donor hits by 1.37 pp and exceeds nuisance by 1.95 pp. One update reverses both contrasts. All magnitude gates pass (Table 9). This mixed response contrasts with the learned donor's -0.59 pp mean; non-following still cannot establish attribute absence.

Nuisance controls leave attribute specificity unresolved: spatial donor changes are $+1.64$ under nuisance and -0.41 under semantic replacement (UniWorld-V1; Table 10). This carrier retains recipient T5 and pooled conditioning while replacing the Qwen change. Non-following at this partial interface cannot establish absence from the full decoder. The embedding calibration tests a different carrier and therefore leaves this interface's sensitivity unresolved.

7 CONCLUSION

Understanding-to-generation transfer calls for distinguishing source validity, target influence, and attribute correspondence. Image and instruction controls expose the conditions of an answer advantage. Paired states measure generation influence, and joint readouts separate recipient improvement from donor-value following. Norm matching tests directional variation at the recipient magnitude while discarding same-direction amplitude differences. Six-architecture experiments and a separate LoRA reference reveal different source–target relationships across tasks and requested values. Mixed calibration leaves non-following inconclusive about attribute availability.

These distinctions concern how observations become knowledge. A score records an outcome under specified conditions; a transfer claim interprets the relation between outcomes. Controls give that interpretation empirical content by testing alternative explanations. Connecting each claim to its conditions of observation makes the language of shared competence answerable to evidence.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval and discovery, and with editing and polishing the manuscript.

ETHICS STATEMENT

The study uses public model checkpoints and public or procedurally generated benchmarks, with automated evaluation of generated attributes. Public images retain their original privacy and representation risks. Generated images may inherit biases from the frozen backbones. The experiments concern capability transfer under controlled evaluation settings.

REPRODUCIBILITY STATEMENT

The experiments and appendix specify evaluation splits, intervention definitions, and statistical units. Scripts reconstruct the main tables from saved predictions and official per-item scores. The project records training configurations and software environments; Appendix A.1 specifies the cohort protocols.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8076–8084, 2019.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, pp. 3194–3201. IOS Press, 2025. doi: 10.3233/FAIA251185.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 11330–11356, 2026.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *International Conference on Learning Representations*, volume 2026, pp. 123737–123758, 2026.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023a.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023b.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Jiwon Kang, Heeji Yoon, Jaewoo Jung, Jaewon Min, Minkyeong Jeon, Biyeon Hwang, Sangwon Jung, and Seungryong Kim. Transferability between understanding and generation in unified multimodal models. In *European Conference on Computer Vision*, pp. 626–647. Springer, 2026.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in neural information processing systems*, 35:1950–1965, 2022.

Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. In *International Conference on Learning Representations*, volume 2024, pp. 24577–24596, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Xiang Wang, Zhifei Zhang, He Zhang, Zhe Lin, Yuqian Zhou, Qing Liu, Shiwei Zhang, Yijun Li, Shaoteng Liu, Haitian Zheng, et al. Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14767–14778, 2026a.

Yu Wang and Sharon Li. Do unified multimodal models think in one space? a lens through cross-branch steering. *arXiv preprint arXiv:2607.26411*, 2026.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026b.

Zimo Wen, Boxiu Li, Wanbo Zhang, Junxiang Lei, Xiaoyu Chen, Yijia Fan, Qi Zhang, Yujiang Wang, Lili Qiu, Bo Li, et al. Unig2u-bench: Do unified models advance multimodal understanding? *arXiv preprint arXiv:2603.03241*, 2026.

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system, 2026. URL https://arxiv.org/abs/2609.01607.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher D Manning, and Christopher Potts. pyvene: A library for understanding and improving pytorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 158–165, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Yi Xu, Cheng Chen, and Mufan Cao. Orthkd: Extracting generalized clinical knowledge from heterogeneous teachers for lightweight deployment. *arXiv preprint arXiv:2607.25545*, 2026a.

Yi Xu, Cheng Chen, and Wenzhuo Lei. If you hear it, help find it: Orthogonal knowledge distillation for open-vocabulary audio-visual event localization. *arXiv preprint arXiv:2609.23376*, 2026b.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Haochen Wang, Zhendong Wang, Bin Lin, Hao Li, Xinyan Xiao, et al. Unified multimodal models as auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41903–41912, 2026.

Cheng Yang, Chufan Shi, Bo Shui, Yaokang Wu, Muzi Tao, Huijuan Wang, Ivan Yee Lee, Yong Liu, Xuezhe Ma, and Taylor Berg-Kirkpatrick. From reasoning to pixels: Benchmarking the alignment gap in unified multimodal models. *arXiv preprint arXiv:2602.08336*, 2026a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026b.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation. *arXiv preprint arXiv:2504.07448*, 2025.

A Implementation and Scope

A.1 Training and Evaluation

Both BLIP3-o cohorts use 72 rank-64 square cores, totaling 294,912 trainable scalars, with one source record per optimizer update. The launch protocol specifies 8,192 counting updates and 4,096 spatial updates. Counting source sampling balances answers 1 through 4; the reference cohort reuses the recorded source order before its label permutation. The model's original processor determines the understanding-image size; the launcher's image-size field records the nominal setting. BLIP3-o uses a frozen 4-bit NormalFloat backbone with bfloat16 computation; counting and spatial set $\beta = 0.2$ and 0.1, respectively. JanusFlow uses 48 query/value sites (196,608 core scalars) with the same ordered records and optimizer settings. UniWorld-V1 has 229,376 core scalars.

The reference normal-label and label-shuffled checkpoints explicitly record AdamW with learning rate 10^{-4} and weight decay 0.01. The original depth-group archive records checkpoints and source orders. The optimizer settings above refer to the independently recorded reference cohort; Table 14 specifies the original archive's scope.

The original BLIP3-o evaluation passes each benchmark question through its chat template without an added answer-only suffix. Greedy decoding permits at most 32 new tokens. The instruction controls append only the fixed strings in Appendix A.6 at evaluation time. Training supervises the answer tokens and keeps the pretrained model frozen. Generation clears the BLIP3-o unconditional cache before each paired image. Primary generation uses each native decoder: BLIP3-o at 512×512, 30 diffusion steps, guidance 3; JanusFlow at 384×384, 30 flow steps, guidance 5; and UniWorld-V1 at 512×512, 28 flow steps, guidance 3.5. Greedy answer decoding permits 32 tokens for BLIP3-o and 128 for JanusFlow and UniWorld-V1; Appendix A.20 checks a shorter UniWorld-V1 budget.

A.2 Core Operator and Local Bound

Figure 2 in the main text expands the two intervention sites shown in Figure 1.

The square-core interventions use Eq. 1. Only the shared square cores learn; setting them to zero recovers the pretrained projection. The implementation evaluates the shared contribution in native precision and adds it to the pretrained output. Numerical diagnostics reconstruct this same operation before measuring the output difference. The exact-arithmetic bound is

$$|\rho_{\ell,j}(x; Q_\ell)| \le \beta |h_{\ell,j}|. \tag{4}$$

Finite-precision reflection can violate a requested sign change even when this bound holds. The sign-control implementation preserves the original coordinate when the reflected difference is unrepresentable, and reports the resulting reduction in directional intervention.

A.3 Sign Controls

The validation protocol reserves 64 target-input prompts after the 128 already observed during parameter-matching development, excluding the generation test prompts. It checks local magnitude preservation, projection site coverage, and nonzero directional change.

The tied-core output-sign control intervenes at a different level. At the same visited input x, compute native zero-state output h_0 and learned-state output h_Q, then form $d = h_Q - h_0$ in 32-bit floating point. For fixed channel signs $s_j \in \{-1, +1\}$, request

$$\widehat{h}_j = h_{0,j} + s_j d_j. \tag{5}$$

We cast the request to the native output precision. If the reflected coordinate cannot reproduce $s_j d_j$, we retain $h_{Q,j}$ and record the fallback. Per-call checks require equal realized coordinate magnitudes and report the fraction of perturbation energy whose direction changes. Each condition uses its own visited trajectory. The control acts on outputs at each visited input, preserving the learned update's coordinate magnitudes and channel support. This output intervention is distinct from a fixed parameter-state replacement.

The reflection condition requests $s_j = -1$ for every channel. For run index u and random-control index $i \in \{0, 1, 2\}$, a CPU random generator uses index $1000 + u + 10000i$. It assigns signs in

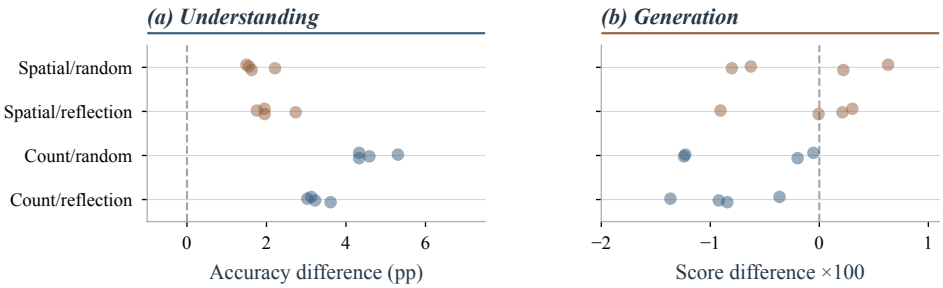

Figure 6: **Direction controls yield different endpoint responses.** Full minus each control; dots show training runs. Magnitude matching applies locally.

sorted projection-site order. We fix the signs before endpoint evaluation and share them across all understanding and generation examples. The random-sign endpoint first averages the three patterns within each training run, then averages across runs. A positive identity sign leaves that coordinate unchanged. Native-precision fallback can also retain a negative-sign coordinate.

Validation checks both preserved magnitude and realized direction change. We record site-level reference, controlled, and changed-direction energy, fallback counts, and call coverage. Across 32 validation conditions, summing over all 72 sites gives controlled/reference energy ratio 1 in every condition. Changed-direction energy spans 31.319–98.563% of reference energy; fallback affects 0.352–0.766% of nonzero coordinates. These held-out checks establish local energy equality at visited inputs. They test directional sensitivity while leaving trajectory norms and perceptual distances unconstrained.

A.4 Local Direction Responses

Figure 6 shows all four run effects for the contrasts between the full-core and local-direction conditions in Table 14. Each generation condition contains 1,000 prompts. Counting understanding contains 1,024 questions and spatial understanding contains 512. The inference unit remains the training run.

A.5 Fixed Parameter Negation

Fixed parameter negation replaces every learned Q_ℓ with $-Q_\ell$ while preserving the maps, gates, and decoder. The oddness of $\tanh$ reverses Eq. 1's update at an identical input in exact arithmetic. Each state then follows its own native inference path. We fix this control before validation and apply it directly in the native precision.

At each projection site, we accumulate squared native update magnitudes over all calls and elements. The square-root ratio of these sums and the ratio of pooled per-element root-mean-square (RMS) magnitudes must each lie in $[0.8, 1.2]$, pooling 32 validation inputs per task. Understanding validation excludes training and test image basenames; generation validation excludes main-benchmark prompts. All 72 sites pass for every run on both capabilities. Across the counting runs, understanding ratios span 0.976–1.070 for integrated L_2 and 0.966–1.058 for pooled RMS; generation ratios span 0.966–1.018 for both measures. These checks compare accumulated local updates along each state's path. Fixed negation retains the learned core's information and tests its direction at comparable aggregate update norms.

The endpoint evaluation uses the original 1,024 counting questions and 1,000 generation prompts per run. Full-minus-negated answer changes are $+2.83$, $+3.42$, $+2.93$, and $+3.03$ pp; generation changes are -1.56, -1.08, -1.12, and -1.34 points on the $100\times$ scale. Their means are $+3.05$ pp and -1.27 points. Negated-minus-frozen means are $+2.05$ pp for answers and $+0.02$ generation points.

Spatial endpoints use 512 questions and 1,000 generation prompts per run. Full-minus-negated answer changes are $+2.34$, $+2.73$, $+1.56$, and $+1.95$ pp; generation changes are $+0.43$, $+0.04$,

-0.70, and $+0.60$ points. Their means are $+2.15$ pp and $+0.09$ points. Negation yields 79.30% answer accuracy, with generation scores 13.96 for 2D and 26.97 for 3D (Table 14).

A.6 Fixed Answer Instructions

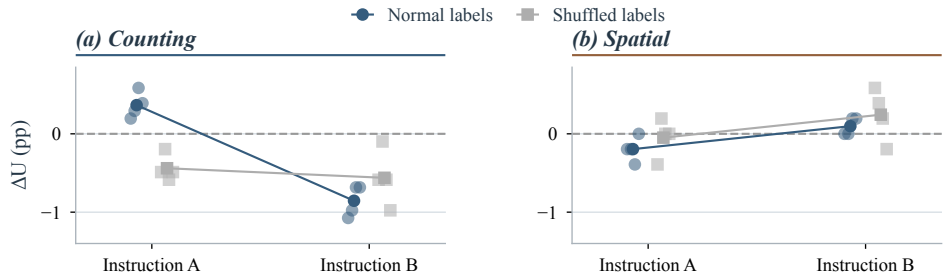

Figure 7: **Reference-core gains depend on the instruction.** Both instructions request answer-only outputs. Markers show training runs; lines join means.

We append the following strings to the original understanding questions. Counting A: "Answer with a single non-negative integer only." Counting B: "Return only the number of objects asked about, as digits. Do not explain." Spatial A: "Answer with exactly one word: true or false." Spatial B: "Return true if the statement matches the image, otherwise false. Do not explain." These strings contain no reference answer.

The original-image counting panel evaluates the same four checkpoints as Table 14. Original questions and requests A/B each use 1,024 images under frozen and full states, yielding 24,576 responses. The original-question responses reproduce all 8,192 historical answers exactly. Both answer-only requests together yield 16,384 responses with complete agreement between benchmark and lexical correctness.

On original questions, conservative lexical scoring requires an unambiguous count and counts flagged responses as incorrect, retaining every answer in the denominator. Its frozen/full accuracies are $53.22/62.82\%$, giving $+9.59$ pp across four runs. The lexical audit flags 19.24% of frozen responses and 4.17% of full-core responses. The benchmark-minus-lexical accuracy difference falls from 5.96 to 1.46 pp, reconciling the benchmark gain: $9.59 - 4.49 = 5.10$ pp. This identity describes scoring sensitivity; the instruction interventions measure responses to changed questions.

The natural-pair subset contains 244 distinct image filenames and 122 questions. Before the instruction-panel inference, we group records by question and select at most one pair with different reference counts, using fixed random ordering and excluding previously used images. Selection uses questions and annotations only. Both-image accuracy applies the same benchmark parser as individual accuracy. The full-minus-frozen pair changes are $+6.15$, -1.02, and -2.25 pp under the original question, A, and B. A conservative lexical check yields $+10.25$, -1.02, and -2.25 pp; all ambiguous answers remain incorrect. The paired images can differ in scene properties beyond the target count.

Figure 3 illustrates the first stored pair at the first training run with an original-question gain, an A loss, and no B gain. This post hoc illustration rule requires unambiguous parsed counts and agreement between both parsers for every displayed answer. It selects the vase pair at index 20. The photographs preserve their original content and aspect ratios. The lower panels retain all 122 pairs across all four runs.

The complete counting image audit contains 73,728 responses across four checkpoints, three image conditions, three instructions, and two core states. We fix a label-blind permutation with no image left in its original position; questions and reference answers retain their original association. We verify every gray image's dimensions and constant RGB value. Sixteen instruction-B responses across gray and permuted conditions contain "2 3" for reference answer 2. The benchmark parser accepts them;

the conservative lexical check marks them incorrect. We keep all responses. Lexical scoring lowers each affected endpoint by 0.10 pp and leaves its full-minus-frozen change unchanged. Image controls measure responses to corrupted inputs against original references.

The spatial audit adds 36,864 responses across the same conditions and four original spatial checkpoints, reproducing all 4,096 historical answers. Its original-image accuracy gains are $+1.56/-0.05/+0.15$ pp for the original question/A/B. All 24,576 spatial A/B responses across image conditions follow the requested true/false format, with agreement between benchmark and lexical correctness. On original questions and images, conservative lexical accuracy rises from 57.23% to 77.34%, a 20.12 pp gain; under A/B it matches the benchmark changes. Under the original question, permuted-image gains range from 4.69 to 5.27 pp across runs, exceeding the corresponding original-image gains of 1.17–1.95 pp. These same-checkpoint comparisons establish instruction and image-condition dependence in both tasks.

The additional reference-core analysis in Figure 7 preserves every question, including incorrect answers. All 36,864 reference-core responses satisfy the lexical answer-only check, with no disagreement between the benchmark parser and lexical correctness audit. The agreement covers the complete answer-only panel under the four fixed instruction strings.

A.7 LoRA Answers and Format

The BLIP3-o learning reference tests acquisition under additive LoRA. Standard additive LoRA trains both rank-64 maps at the same query/value sites, with 14,745,600 trainable scalars versus 294,912 in the fixed-map core. Each normal-label and shuffled-label run uses 8,192 source records in recorded order, AdamW, learning rate 10^{-4}, and weight decay 0.01. This comparison changes the trainable variables, parameter count, and output operator together.

We evaluate all registered runs on the same 1,024 counting questions. The digit request appends "Answer with a single digit."; the word request appends "Answer with one lowercase number word." Greedy decoding starts after the model's assistant prefix and permits 32 new tokens. The audit retains complete generated answers and accepts answer-only digit or number-word aliases for semantic correctness, ignoring case. Strict format compliance separately checks the requested representation. These instructions differ from counting A/B in Appendix A.6.

Table 4: **LoRA answer accuracy and format.** Percentages on 1,024 counting questions. Accuracy accepts digit/word aliases; format checks the requested form. Superscripts give normal-label LoRA minus frozen.

	Digit request		***Word request***	
Training condition	Accuracy	Format	Accuracy	Format
Frozen	66.89	99.02	68.95	2.34
Normal labels	**74.17**$^{+7.28}$	**100.00**$^{+0.98}$	**73.83**$^{+4.88}$	**0.00**$^{-2.34}$
Shuffled labels	49.39	99.98	52.83	0.00

Table 5: **LoRA endpoint changes by run.** LoRA minus frozen: answer accuracy (pp) and numeracy score ($\times 100$).

Endpoint	Run 17	Run 29	Run 43	Run 71
Digit-request answer	+7.23	+7.23	+6.93	+7.71
Word-request answer	+4.59	+5.18	+4.39	+5.37
Generation score	-0.78	+1.46	-0.79	-0.02

Normal-label LoRA improves answer-alias accuracy under both requests (Table 4). Every normal-label answer to the word request uses a different format, separating answer-alias correctness from requested-format compliance. Shuffled-label effects vary across runs. All runs share the same frozen model and identical frozen answers.

Table 6: **LoRA answers under image controls.** Alias accuracy (%) on 1,024 questions; superscripts give LoRA minus the corresponding frozen baseline.

	Digit request		*Word request*	
Image	Frozen	LoRA	Frozen	LoRA
Original	66.89	**74.17**$^{+7.28}$	68.95	**73.83**$^{+4.88}$
Gray	24.12	**31.88**$^{+7.76}$	21.78	**31.10**$^{+9.33}$
Permuted	3.71	**14.92**$^{+11.21}$	3.71	**16.85**$^{+13.13}$

For the normal-minus-frozen answer gains, two-sided run sign-flip tests yield $p = 0.125$; Holm adjustment over the fixed nine-comparison family gives $p = 1$. These values summarize run-level sign consistency on the previously examined question panel.

Image controls. We reevaluate the same checkpoints with neutral gray images (RGB 127 in each channel) and a fixed label-blind permutation of held-out images with no image left in its original position. Questions and original reference answers stay fixed. All arms use the same image assignment; we score each arm against its corresponding frozen baseline. The two controls, two instructions, three states, and four training runs yield 49,152 additional answers. Table 6 reports LoRA-minus-frozen gains and the absolute answer-alias accuracies. The gains persist under both image controls. Gray inputs introduce a distribution shift, and permuted inputs can contain relevant objects by chance; these diagnostics test corrupted images.

A.8 Training Pairing Controls

The confirmation panel contains 512 images from the official TallyQA test set, with 64 questions in each simple/complex $\times$ answer-1/2/3/4 stratum. Its image identifiers exclude prior registered project use; pretrained-model exposure remains unknown. We reuse the normal-label additive-LoRA checkpoints and train an image-shuffled arm with the same questions, answers, sample order, optimizer, and 8,192-update budget. This arm differs from label shuffling. Evaluation crosses three model states, three image conditions, two answer requests, and four training runs, yielding 36,864 complete answers. Table 7 reports the complete comparison.

Table 7: **Training pairing and image dependence.** Accuracy (%) on 512 reserved images; superscripts give paired training minus frozen.

	Digit request			*Word request*		
Training condition	Original	Gray	Permuted	Original	Gray	Permuted
Frozen	57.81	24.41	4.30	58.98	23.63	3.91
Shuffled images	48.97	34.62	32.42	49.80	33.50	33.50
Paired images	**67.24**$^{+9.42}$	**32.47**$^{+8.06}$	**16.94**$^{+12.65}$	**66.85**$^{+7.86}$	**31.88**$^{+8.25}$	**19.38**$^{+15.48}$

Let A^c_m denote answer-alias accuracy for model m under evaluation-image condition c, where $m \in \{N, P, Z\}$ denotes paired training, shuffled-image training, or the frozen baseline. We distinguish the normal-training gain $A^c_N - A^c_Z$ from the training-pairing contrast $A^c_N - A^c_P$. On original images, the latter decomposes as

$$A^{\text{real}}_N - A^{\text{real}}_P = (A^{\text{real}}_N - A^{\text{real}}_Z) + (A^{\text{real}}_Z - A^{\text{real}}_P). \tag{6}$$

The digit-request terms are $9.42 + 8.84 = 18.26$ pp; word-request terms are $7.86 + 9.18 = 17.04$ pp, with each term rounded independently. Thus the paired-versus-shuffled gap includes shuffled-training damage.

The interaction $(A_N - A_P)^{\text{real}} - (A_N - A_P)^{\text{gray}}$ is 20.41 pp for digit requests and 18.65 pp for word requests. Both interactions are positive in every training run. In contrast, normal-minus-frozen gains on gray images remain 8.06 and 8.25 pp. These contrasts answer different questions: intact

training pairing contributes to real-image performance, while gains over frozen persist under corrupted evaluation images. The measured interaction is between training pairing and test-image condition. We reuse each trained arm's four checkpoints across evaluation-image conditions and answer requests; all conditions use one frozen checkpoint. The new images supply a separate evaluation panel for those same training replicates.

A.9 LoRA Generation and Core Basis

The paired LoRA generation panel contains 24,000 item scores: four training runs, normal/shuffled training, three conditions (active, removal, replay), and 1,000 prompts. Normal-minus-removal effects are -0.78, $+1.46$, -0.79, and -0.02 score points in run order, with mean -0.03. Shuffled-label training has mean effect -10.77; one run contributes a -38.72 change and answers every source question with 1. Normal-versus-shuffled differences therefore include control deterioration. Saved replay checks reproduce the active outputs within this LoRA protocol.

A separate fixed-core comparison derives frozen maps from singular value decomposition. Its 24,576 answers yield mean normal-minus-zero changes of -0.07 pp for digit requests and -0.17 pp for word requests. Both mean changes are negative under this fixed-core basis configuration. The SVD-basis, additive-LoRA, and original-core cohorts use distinct checkpoints.

A.10 LoRA Replay Checks

During an active additive-LoRA trajectory, replay records each realized local update $\rho^{\mathrm{act}}_{\ell,k}$, where k indexes a matched forward call. The replay trajectory restores that recorded value:

$$h^{\mathrm{rem}}_{\ell,k} = \mathrm{Base}_\ell(x^{\mathrm{rem}}_{\ell,k}), \qquad h^{\mathrm{rep}}_{\ell,k} = \mathrm{Base}_\ell(x^{\mathrm{rep}}_{\ell,k}) + \mathrm{stopgrad}(\rho^{\mathrm{act}}_{\ell,k}). \tag{7}$$

We match call order, task, and layer. Replay checks whether the recorded local contributions recover the active outputs along that matched call sequence.

A.11 Donor-Checkpoint Source Controls

The BLIP3-o reference updates in Tables 2 and 3 also undergo the fixed-checkpoint source controls. Native original-request answers replay exactly before the instruction and image comparisons. Each counting or spatial cell retains the same 1,024 or 512 questions, respectively, with the original references and decoding budget. Table 8 reports the paired accuracy changes.

The digit-request counting gains are small or negative on intact images, while gray-image gains remain positive. The spatial gain under the original request is larger on permuted images than on intact images. Thus the source measurement limits apply to the reference updates themselves. This links the source and donor audits without treating the answer gain as a validated visual capability or combining it with the original-core cohort.

Table 8: **Source controls on donor checkpoints.** Full minus frozen accuracy (pp). A/B request digits for counting and true/false for spatial relations.

Image	***Original request***	***Answer-only A***	***Answer-only B***
Counting			
Intact	+4.98	+0.37	-0.85
Gray	+1.29	+0.81	+1.10
Permuted	-0.12	+0.20	+0.44
Spatial relations			
Intact	+1.51	-0.20	+0.10
Gray	-0.54	+0.20	+0.29
Permuted	+5.13	+0.00	-0.29

A.12 Intervention Calibration

Prompt contrasts test whether the registered BLIP3-o carrier can express count changes at the learned intervention magnitude. We retain all 128 counting pairs and the same reference updates used in the source controls, endpoints, and donor readouts. For each pair, the frozen conditional contrast $c_0^+(p_d) - c_0^+(p_r)$ replaces the learned change at its recipient norm. Prompt nuisance changes the object while retaining the recipient count. Both interventions preserve the recipient request, unconditional branch, decoder noise, and native precision. Every realized magnitude lies within the registered 20% energy tolerance. A separate natural donor call uses its own request and both frozen branches; it checks base generation and does not share the intervention budget.

The prompt attribute contrast raises donor hits by 1.37 pp on average and exceeds prompt nuisance by 1.95 pp. Both differences reverse for one reference update; the donor-change range is $[-4.69, +3.91]$ pp. The learned donor raises recipient hits by 4.88 pp while donor hits change by -0.59 pp. These comparisons separate response to the prompt contrast from response to the learned change at the same carrier and magnitude. They provide partial calibration, not a guarantee of sensitivity for every update. Natural donor requests achieve 14.06% donor exact hits, which also limits the strength of the base capability check. Neither result certifies that the source update learned a visual count representation or that non-following implies its absence from generation.

Table 9: **Prompt contrasts at the learned magnitude.** Exact-count hits (%) on 128 pairs; donor changes from Own are in pp. Ranges span reference updates. Natural donor uses its own request outside the matched comparison.

Condition	***Recipient***	***Donor***	***Donor change***	***Observed range***
Recipient request, fixed noise and matched magnitude				
Own change	16.21	9.77	0.00	—
Learned donor	21.09	9.18	-0.59	$[-3.12, +2.34]$
Prompt attribute	15.43	11.13	+1.37	$[-4.69, +3.91]$
Prompt nuisance	16.41	9.18	-0.59	$[-3.91, +2.34]$
Random sign	15.82	8.79	-0.98	$[-3.91, +2.34]$
Separate base capability check				
Natural donor	14.06	14.06	—	—

A.13 Donor Controls

Branch controls separate conditional and unconditional changes; nuisance and sign controls test the specificity of semantic replacement. Table 10 reports their changes from Own alongside the semantic comparison in Table 3.

The donor study uses normal-label reference cores, distinct from the original retention checkpoints. Each capability uses 128 paired recipients per training run. Counting recipients come from simple count/object prompts with counts one through six. A semantic donor keeps the object and changes the count; a nuisance donor keeps the count and changes the object. Spatial recipients use 2D prompts. A semantic donor keeps the subject/object pair and changes the relation; a nuisance donor keeps the relation and changes at least one object. For each recipient, the implementation sorts eligible donors by prompt text and record identifier and takes the first in each class. The counting panel reuses 98 distinct semantic donors and only nine nuisance donors. Shared donors introduce dependence across recipient examples. Counting recipients span one through six, whereas source training answers span one through four; this panel also differs from the 1,000-prompt generation benchmark.

The exact-count requests compete on each image; we interpret their marginal hit-rate changes as one joint outcome. We classify each paired conditional-only and semantic-donor output as a recipient hit, donor hit, other detected count, or missing object. Table 11 reports joint outcome proportions. In BLIP3-o, Other-to-Recipient and Donor-to-Recipient transitions account for 8.59% and 1.95% of paired observations. The corresponding proportions in JanusFlow are 7.42% and 1.37%. UniWorld-V1 changes recipient and donor hit rates by $+0.39$ and $+0.59$ pp. These transitions describe a redistribution of exclusive labels.

For BLIP3-o, we capture the conditional and unconditional embeddings before diffusion. Equation 3 rescales donor changes using the Frobenius norm over all conditional embedding entries. The native-precision squared energy ratio lies in $[0.8, 1.2]$ for every semantic and nuisance donor. The conditional-only control uses the recipient's full conditional embedding and zero-state unconditional embedding. The random-sign control reflects the recipient conditional change along channel-wise signs, fixed across examples, with the native-precision fallback in Eq. 5. Both keep the recipient zero-state unconditional embedding. The unconditional-only control combines the zero-state conditional embedding with the full-state unconditional embedding. Interface replay restores both captured full-state embeddings before decoding.

We reset decoder randomness at the interface using $8087 + i$ for recipient index i. This excludes encoder random-number consumption from decoder noise and defines a separate inference protocol from the earlier full/zero panel. The recipient prompt and initial noise stay fixed across conditions; the evolving decoder states can differ. We score each output against the original recipient attribute and the semantic donor attribute using the official evaluator. Counting reports $100 \,\mathrm{mean}\, \mathbf{1}[s = 1]$ for exact-count hits and $100 \,\mathrm{mean}\, \mathbf{1}[s > 0]$ for object presence, using the official single-object numeracy score $s \in \{0, 0.5, 1\}$. For this restricted setting, s equals half the sum of the two indicators; the spatial panel reports continuous relation scores. All contrasts compare the same recipient and scoring target across conditions.

A.14 JanusFlow Conditioning

The JanusFlow study retains the same 128 recipient, semantic-donor, and nuisance-donor prompt triples per task as BLIP3-o and evaluates all eight conditions on four reference cores per task. Native decoding uses 384×384 images, 30 flow steps, and guidance 5. At each step, we capture the visual-token hidden state before decoder alignment. Frozen and full passes share the current recipient latent and timestep; their difference defines the recipient change. Donor passes use that same latent and timestep with the donor prompt. We rescale the donor change to the recipient norm and add it to the frozen recipient carrier, keeping the unconditional carrier frozen.

Random-sign controls use one fixed sign pattern across flow steps. Conditional-only, unconditional-only, and interface replay respectively restore the conditional, unconditional, and both full-state carriers. Interface replay reproduces all 128 full-state images per run; native replay also reproduces the first 16 frozen and full images per run. All semantic, nuisance, and random-sign interventions pass both the runtime norm-ratio check and the squared-norm ratio $[0.8, 1.2]$ check at every visited step. These checks cover $4 \times 128 \times 30$ steps per condition and task. The dynamic carrier and BLIP3-o's static conditioning provide separate implementations of the recipient/donor comparison.

Counting semantic donors change recipient hits by -4.69, $+2.34$, $+8.59$, and $+4.69$ pp over conditional-only across the four runs. Donor changes are -2.34, -4.69, -3.13, and -0.78 pp, giving a mean of -2.73 pp. The semantic donor mean also lies below frozen generation: 10.94% versus 13.28%. Table 10 compares frozen, branch, nuisance, and direction controls using the same paired targets. Figure 8 shows the spatial item-level response distributions.

A.15 Presence and Exact Counts

The original-core 1,000-prompt counting panel mixes requested counts and scene complexity. We partition all prompts in a post hoc descriptive analysis: 116 contain only a count from 1–4 and one object-category word, 124 have the same form with counts 5–8, and 760 add context or object categories. The partition uses complete prompt strings and retains all four runs. Mean generation scores change from 61.64 to 60.67, 48.79 to 48.79, and 34.66 to 33.15, respectively. Weighting these groups by their sizes recovers the full-panel change from 39.54 to 38.28.

For the two simple-prompt groups, the official score equals half the sum of detected object presence and exact-count correctness indicators (Huang et al., 2025). In the 1–4 group, object presence changes from 90.52% to 89.44% and exact-count hits from 32.76% to 31.90%. The composite score decreases in every run; exact-count hits decrease in two, stay equal in one, and increase in one. The 5–8 group retains mean presence of 91.13% and exact-count hits of 6.45% under both states. These components separate detection from count agreement within the evaluator. The input groups describe prompt composition; their differences combine counts, object categories, and scene context.

Applying the same partition to the reference cohorts yields mean score changes of -0.22, $+0.81$, and -1.27 for BLIP3-o, and -1.08, $+1.51$, and $+0.50$ for JanusFlow, in the same group order. Table 12 separates presence and exact-count changes for the simple prompts. On counts 1–4, exact-count hits decrease in two runs, stay equal in one, and increase in one for those two cohorts. On counts 5–8, they increase in three runs and decrease in one for BLIP3-o; JanusFlow has three increases and one unchanged run. The additional architectures also separate detection from exact counts. DiMOO increases exact hits in both simple-count groups. Janus-Pro loses 6.90 pp of exact hits for counts 1–4 despite gaining object presence, while its score loss for counts 5–8 comes from detection. Show-o2 gains 0.86 pp of exact hits for counts 1–4 and loses 3.23 pp for counts 5–8. Thus, the complete-panel score combines distinct changes in detection, count agreement, and prompt composition across all six backbones.

A.16 Donor Responses by Count

We partition all 128 prompt pairs by whether recipient and donor counts fall within the supervised range 1–4 (Table 13). This post hoc analysis applies the same input-defined groups to every displayed backbone. Objects and scenes also vary across groups.

In BLIP3-o, relative to conditional-only, semantic donors reduce donor hits in all four runs of the 26-pair group requesting counts 1–4 for both targets. The other three groups have positive mean donor changes: $+0.48$ pp for 1–4 to 5–6, $+2.59$ pp for 5–6 to 1–4, and $+3.57$ pp for 5–6 to 5–6.

Within the 26 pairs requesting 1–4 for both targets, semantic donors improve recipient hits over nuisance in all four runs. Donor hits rise in one run and fall in three. The 29 pairs requesting 5–6 for the recipient and 1–4 for the donor have the largest mean donor improvement over nuisance, with three positive run differences and one zero. Their mean donor hit rate also exceeds frozen generation. Thus, the all-pair recipient benefit coexists with a positive donor response in one requested-count group. Each range group contains its original prompts and object categories.

JanusFlow donor hits decline on average in all four groups relative to conditional-only: -2.88, -1.44, -3.45, and -4.76 pp in table order. Within counts 1–4, two runs lose donor hits and two remain unchanged; mean recipient hits rise by 3.85 pp. The common all-pair recipient/donor separation therefore coexists with different count-range responses across models. UniWorld-V1 loses donor hits within 1–4 but gains them in the other groups. DiMOO loses donor hits in the first three groups and retains them in the fourth. Janus-Pro gains donor hits in every group involving a count above four, whereas Show-o2 has opposite donor signs in the two groups with a recipient count above four. The aggregate donor contrast therefore need not describe each requested-count group.

A.17 Spatial Relation Sensitivity

The six spatial panels share the same 128 prompt pairs. A post hoc partition by input relation strings retains every pair: 56 compare distinct directional predicates (left, right, top, or bottom), 59 compare a direction with proximity, and 13 compare proximity expressions. The latter use *near*, *next to*, or *on side of*, which share the same geometric branch in the official evaluator (Huang et al., 2025). The partition uses inputs and evaluator rules, independently of the observed score changes. The following evaluator audit uses the BLIP3-o reference cohort; its four training runs reuse the same prompts.

When both parsed object categories appear in the detections, the evaluator computes $s = 0.25c_1 + 0.25c_2 + 0.5g$, then sets values below 0.5 to zero. Here $c_1, c_2 \in [0, 1]$ are detection confidences and $g \in [0, 1]$ is the geometric score. A final score above 0.5 therefore requires $g > 0$. We use this sufficient condition to count detector-supported directions. Distinct directional predicates are mutually exclusive for the same pair of detected boxes. Every directional output has at most one supported target; none scores exactly 0.5. The threshold leaves weaker geometric evidence unresolved. The saved official scores omit boxes and confidences, so we retain the combined score for continuous comparisons.

We also audit the evaluator's first/last-noun object extraction using its deployed spaCy 3.5.3 and en_core_web_sm 3.5.0. All 56 directional pairs extract the intended object categories. Two other pairs misparse *key* as an adjective. For *a boy on side of a key*, the evaluator selects person/side; for *a boy near a key*, it selects person/person and reuses the same person box. The resulting donor score measures person detection, independently of the key. For *a turtle on the left of a key* and *a turtle*

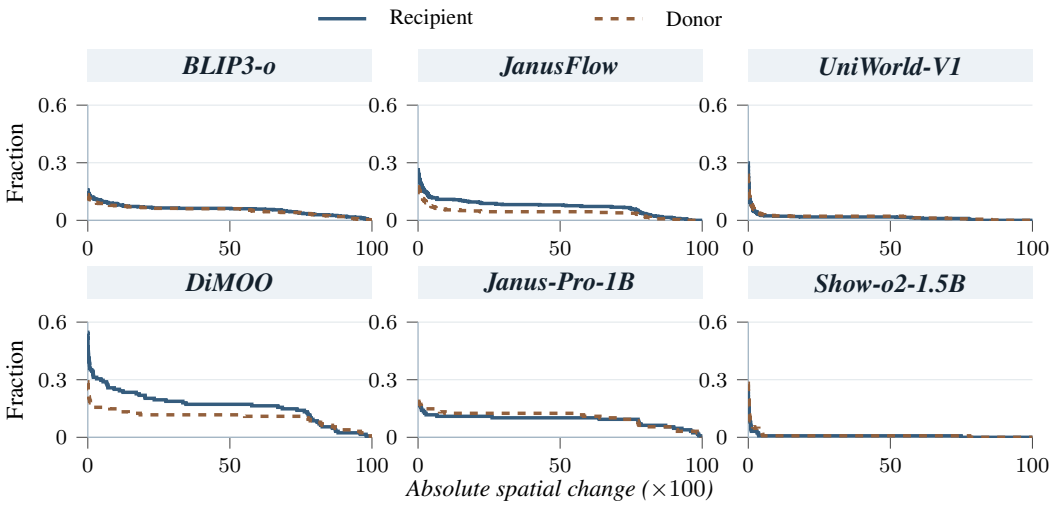

Figure 8: **Spatial sensitivity across six backbones.** Fraction of items above each absolute Swap minus Own score-change threshold (×100). Both readouts score the same images; Table 3 gives signed changes.

on side of a key, it selects turtle/left and turtle/side. We retain these prompts in the official-score aggregates and distinguish those aggregates from directional support. A separate post hoc check excludes these two pairs and yields a 1.50-point donor-score gain over conditional-only across the remaining 126 pairs. The four gains are 0.82, 0.14, 1.69, and 3.34 points. The two excluded pairs contribute 0.28 points to the original 1.76-point mean, using all 128 pairs as the denominator. Table 3 retains the complete panel. The score gain persists with correct object extraction, while the directional subset identifies which attribute the scorer supports.

A.18 LoRA Output Diagnostics

Pre-addition magnitude matching does not ensure matched native outputs. In the additive-LoRA audit, 16 held-out prompts per run give a random-control root-mean-square change only 0.492–0.626 times the selected state's change. All four runs fail the native-output gate ($\pm 2\%$ in root-mean-square magnitude). A second audit uses eight prompts per run and rescales controls at a fixed layer input, passing the gate at all 72 sites. The match holds at those fixed inputs; each condition subsequently follows its own trajectory.

The corresponding 256-prompt generation panel gives mean zero and selected scores of 55.08 and 55.23, a paired change of $+0.15$ on the official $100\times$ scale. This panel has its own prompts and protocol; we keep it separate from the 1,000-prompt LoRA comparison in Appendix A.9.

Joint output matching. A subsequent additive-LoRA direction study passes its source-retention continuation check on reserved images, then attempts joint native-output matching at all 72 sites. All 12 combinations of four runs and three random directions fail the joint fit/validation criterion (RMS tolerances 2% and 20%). The registered continuation rule stops generation scoring for this control family.

A.19 Core Retention and Components

These interventions use the original BLIP3-o fixed-core cohort. We rank six depth groups by the source development-accuracy drop on removal, then compare the top two with two random groups and the remaining four. Singular-component controls retain the 64 largest components across cores or their complement. For either endpoint metric M, the interaction

$$I_M = M(Q) - M(Q_{\text{top}}) - M(Q_{\text{comp}}) + M(0) \tag{8}$$

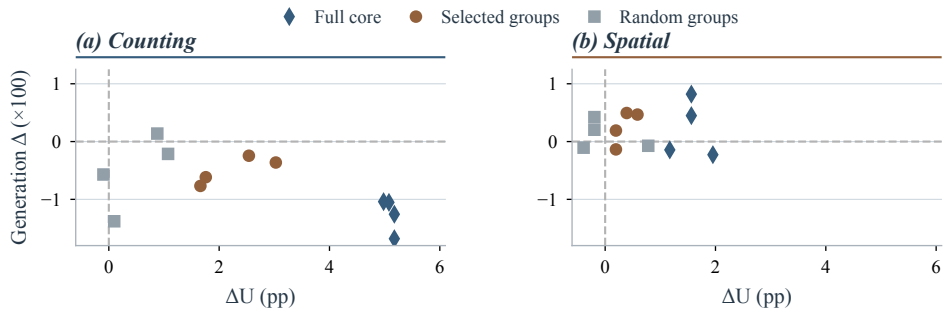

Figure 9: **Understanding retention does not order generation effects.** Points show runs and retained states. Selected/random retain two of six groups; zero denotes frozen.

measures nonadditivity of these retained states. The original counting cores have interactions of $+2.93$ answer points and -0.94 generation-score points. These endpoint interactions do not identify semantic interactions between visual attributes.

The six depth groups each contain six consecutive transformer layers and both query/value projection sites. Development evaluation shuffles the order of full-state and group-removal conditions using the training-run index. The random comparator samples one of the 14 two-group subsets other than the selected pair, using a pseudorandom index of 9000 plus the training-run index. It may overlap the selected pair in one group.

For the spectrum-matched random state, each site's core supplies its numerical rank and nonzero singular values. Independent Gaussian matrices and reduced QR decompositions supply left and right orthonormal factors; their product with the original singular values gives the random core. The implementation constructs this state in float64 and casts back to the native dtype. The matched quantity is the pre-cast parameter spectrum; the nonlinear operator determines the resulting native-output perturbations.

A.20 Native Interfaces

Table 2 compares six native answer and generation interfaces. Table 12 applies the same requested-count partition to their recorded generation outputs. We analyze conditioning interventions separately because each native carrier requires its own magnitude check.

Native counting endpoints. The additional DiMOO (Xin et al., 2025), Janus-Pro (Chen et al., 2025b), and Show-o2 (Xie et al., 2026) rows in Table 2 use the reference core operator at query/value projections and the same ordered counting supervision. Frozen and full states answer the same 1,024 held-out questions at a 128-token budget and generate the same 1,000 counting prompts with paired noise. DiMOO uses its native 512×512 masked decoder with 64 steps and guidance 4; Janus-Pro generates 384×384 images with 576 autoregressive image tokens and classifier-free guidance 5; Show-o2 uses its native 432×432 decoder with a 50-point Euler grid and guidance 7.5.

Native donor interfaces. Table 3 uses the same 128 registered prompt triples and scores each image against both attributes. Janus-Pro replaces the next-image-token logits, with every stream receiving the same sampled recipient token. DiMOO uses masked-image-token logits and copies the current recipient image-token state into each donor pass. Show-o2 replaces conditional velocity before classifier-free guidance at each of 49 Euler evaluations on its 50-point time grid. We match each donor change to the recipient change at its native carrier, with the recipient request, unconditional branch, and initial noise fixed. For these three dynamic interfaces, the native-cast Frobenius-norm ratio of realized to reference change must lie in $[0.8, 1.2]$ at each nonzero-reference evaluation. Zero-reference evaluations require zero realized change; zero donor energy fails a swap. A pair qualifies only if every evaluation passes. Magnitude qualification concerns the donor intervention and does not alter paired endpoint results.

UniWorld-V1 evaluation. Eight full-budget UniWorld-V1 cores cover counting and spatial understanding under the reference training recipe. Each core answers both 32-token and 128-token versions of the same test questions. Table 2 uses the native 128-token results. At 32 tokens, counting accuracy changes from 59.86% to 56.15%; spatial accuracy retains the same 82.81% to 83.74% comparison as at 128 tokens. Counting and spatial generation decrease in three and two of the four runs, respectively. The decoder uses the native joint Qwen/T5 path, 28 flow steps, guidance 3.5, and paired prompt-index noise at 512×512 resolution. Later runs reuse the identical frozen observations from the first run within each capability. The campaign evaluates 15,360 fresh answers and generates 10,000 images; four-run comparisons reference 24,576 answers and 16,000 images. The eight inference-time checks confirm answer and pixel equality between zero cores and bypassed branches on the first question and generation prompt.

UniWorld-V1 donor interface. The UniWorld-V1 donor study uses the same 128 prompt triples and four source cores per task as the other donor studies. We insert the Qwen conditioning change before the frozen FLUX decoder and keep recipient T5 and pooled embeddings fixed. We align token sequences with SequenceMatcher, disabling automatic junk filtering. Equal spans copy donor changes. For replacement spans of donor/recipient lengths m, n, recipient offset j copies donor offset $\lfloor jm/n \rfloor$. Recipient-only spans copy the next donor position, capped at sequence end. Internal donor-only spans leave no retained entry; terminal donor-only spans add their mean to the final recipient position. We rescale after alignment. All 128 pairs in every run pass the native-cast Frobenius-norm ratio gate $[0.8, 1.2]$. Frozen and full outputs reuse the native endpoint images, and all conditions use the recipient's prompt-index noise. Table 3 shows the signed contrasts; Figure 8 shows the spatial item-level response distributions.

JanusFlow spatial retention. In the JanusFlow spatial direction audit, full and selected states reduce understanding accuracy by 2.20 and 2.34 pp on average, respectively, with decreases in every run. Selected-minus-zero generation is near zero, and selected-minus-complement is negative in every run (Table 15). This experiment applies parameter-state retention in JanusFlow. The donor experiments instead replace conditioning changes (Table 3).

A.21 Analysis and Uncertainty

Training runs are the replicates. Repeated measurements include questions, generation prompts, and random directions within a run. Saved analyses retain crossed run/prompt bootstrap intervals and paired contrasts, with spatial resampling stratified by 2D/3D. Tables emphasize absolute outcomes and means.

With four training runs, a two-sided exact run sign-flip test has minimum $p = 0.125$. We interpret prompt-based intervals as uncertainty over the reused prompts and report training-run effects separately. Holm adjustment covers every comparison in each listed contrast/subgroup family. The instruction study is descriptive on an already inspected panel. The reserved-image study supplies new evaluation images and reuses checkpoints.

Replay results retain the checkpoint and inference-protocol identity of the experiment that generated them.

A.22 Qualitative Score Checks

Figure 4 shows intact outputs from the original core cohort under identical prompts and decoding within each pair. We fix training run 17 for each capability and exclude the previously reserved diagnostic prompt pool recorded in the selection manifest. Counting candidates use single-object prompts with counts two through four and a gain or loss of an official exact-count hit; spatial candidates separately use 2D and 3D prompts of at most nine words and a positive or negative score change. Within each of these six strata, we select the five pairs nearest the median absolute score change, breaking ties by prompt index, before viewing the images.

The selected pairs illustrate visible composition alongside automatic scores; they do not estimate the prevalence of visually correct changes. Figure 4 annotates the lowest-index pair from each stratum with its prompt and scores. Figure 10 retains all thirty selected pairs. Frozen and full-core outputs sit side by side within each prompt. For prevalence calculations we separately pool all 4,000 counting,

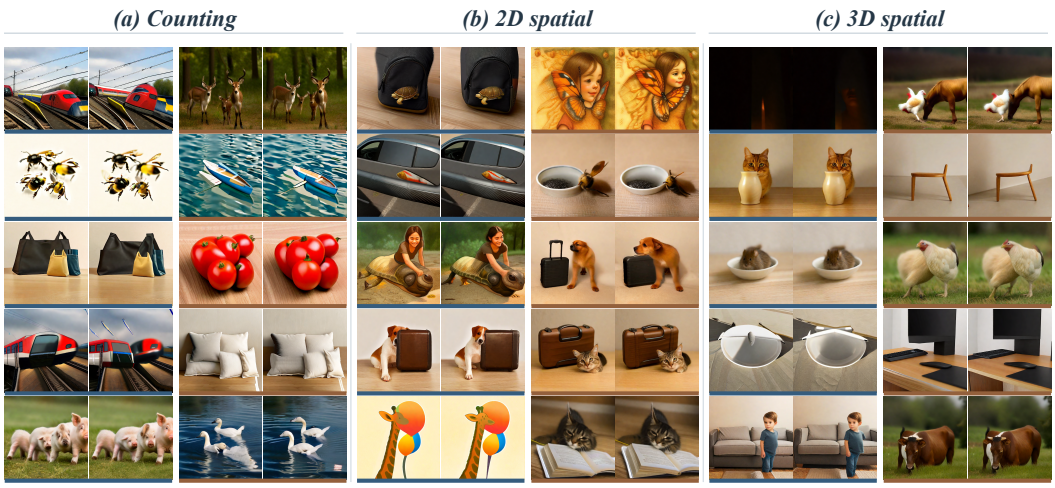

Figure 10: **Gallery of paired outputs.** Thirty frozen/full pairs: five score increases and five decreases per task. Blue/orange marks increases/decreases.

2,000 2D, and 2,000 3D prompt–checkpoint pairs; absolute score changes at most 10^{-12} count as numerical equality. The four trained cores are the independent replicates, with prompts supplying repeated measurements.

Method illustration. Figure 1 uses the first available source-training-pool image with an answer from one through four, with its recorded question and answer. Its generated images use the first recipient in the existing counting donor design, under the same checkpoint and initial noise as each other. We select these records before viewing their images or consulting their scores; the figure shows the recorded protocol execution, including the deformed semantic-donor output. All six images retain their original content and aspect ratio.

Table 10: **Branch, nuisance, and direction controls.** Changes from Own: exact-count hits (pp) and spatial scores ($\times 100$).

Target	Frozen	Full	Unconditional	Nuisance	Random sign
BLIP3-o / Counting					
Recipient	-0.59	-0.59	-1.95	-3.12	-0.39
Donor	$+1.95$	-1.56	$+1.95$	-1.56	-0.98
JanusFlow / Counting					
Recipient	$+0.78$	$+0.98$	$+1.56$	-1.56	$+1.37$
Donor	-0.39	-0.39	-3.32	$+1.56$	-0.78
UniWorld-V1 / Counting					
Recipient	-0.20	0.00	-0.20	$+0.98$	$+0.98$
Donor	$+1.37$	0.00	$+1.37$	$+0.39$	0.00
DiMOO / Counting					
Recipient	-3.12	$+5.47$	0.00	$+1.56$	$+0.78$
Donor	-1.56	-1.56	-0.78	-3.12	-0.78
Janus-Pro-1B / Counting					
Recipient	$+5.47$	-1.56	-2.34	-3.12	-2.34
Donor	$+4.69$	-1.56	$+7.81$	$+3.12$	$+1.56$
Show-o2-1.5B / Counting					
Recipient	-3.12	-1.56	-0.78	0.00	-2.34
Donor	$+0.78$	-2.34	-2.34	-1.56	0.00
BLIP3-o / Spatial relations					
Recipient	$+1.01$	-0.48	$+1.19$	$+0.73$	$+0.13$
Donor	-0.01	$+0.01$	$+1.20$	-0.43	$+0.30$
JanusFlow / Spatial relations					
Recipient	-0.57	-1.54	-0.12	-1.11	-0.72
Donor	-0.41	-0.79	-0.33	-0.20	$+0.29$
UniWorld-V1 / Spatial relations					
Recipient	$+1.23$	0.00	$+1.23$	$+0.13$	-0.51
Donor	$+1.16$	0.00	$+1.16$	$+1.64$	$+0.40$
DiMOO / Spatial relations					
Recipient	-1.97	-0.20	-2.69	-3.79	-1.96
Donor	$+0.72$	-0.57	$+0.42$	$+1.39$	$+0.30$
Janus-Pro-1B / Spatial relations					
Recipient	$+0.88$	$+4.65$	$+6.74$	$+3.35$	$+4.17$
Donor	-0.15	$+3.16$	$+5.30$	$+0.29$	$+3.27$
Show-o2-1.5B / Spatial relations					
Recipient	$+1.22$	-0.15	$+0.44$	$+0.02$	$+0.60$
Donor	$+0.61$	-0.04	$+1.05$	$+0.02$	$+0.66$

Table 11: **Counting transitions under donor replacement.** Joint percentages, normalized within each backbone. Rows: Own; columns: Swap. Recipient and Donor are exclusive exact-count hits; Other denotes another detected count; Missing denotes no detected object.

	After Swap				
Before Own	Recipient	Donor	Other	Missing	Total
BLIP3-o					
Recipient	10.35	0.78	4.88	0.20	16.21
Donor	1.95	5.27	2.54	0.00	9.77
Other	8.59	2.93	54.30	1.37	67.19
Missing	0.20	0.20	0.98	5.47	6.84
After total	21.09	9.18	62.70	7.03	100.00
JanusFlow					
Recipient	15.04	0.98	4.88	0.20	21.09
Donor	1.37	6.64	5.66	0.00	13.67
Other	7.42	3.12	48.83	0.98	60.35
Missing	0.00	0.20	1.56	3.12	4.88
After total	23.83	10.94	60.94	4.30	100.00
UniWorld-V1					
Recipient	49.61	0.39	2.54	0.00	52.54
Donor	0.98	8.59	0.00	0.00	9.57
Other	2.34	1.17	32.03	0.00	35.55
Missing	0.00	0.00	0.00	2.34	2.34
After total	52.93	10.16	34.57	2.34	100.00
DiMOO					
Recipient	45.31	0.78	7.81	0.00	53.91
Donor	1.56	2.34	7.03	0.78	11.72
Other	6.25	3.12	22.66	0.00	32.03
Missing	0.00	0.00	0.00	2.34	2.34
After total	53.12	6.25	37.50	3.12	100.00
Janus-Pro-1B					
Recipient	4.69	0.00	9.38	2.34	16.41
Donor	1.56	0.78	3.12	0.00	5.47
Other	7.03	7.81	29.69	7.81	52.34
Missing	1.56	0.78	8.59	14.84	25.78
After total	14.84	9.38	50.78	25.00	100.00
Show-o2-1.5B					
Recipient	43.75	0.78	2.34	0.00	46.88
Donor	1.56	5.47	0.78	0.00	7.81
Other	0.78	0.78	39.06	0.00	40.62
Missing	0.00	0.00	0.00	4.69	4.69
After total	46.09	7.03	42.19	4.69	100.00

Table 12: **Generation changes across requested counts.** Full minus frozen: score (×100), presence and exact hits (pp).

	Counts 1–4 (116 prompts)			*Counts 5–8 (124 prompts)*		
Backbone	Score	Presence	Exact hits	Score	Presence	Exact hits
BLIP3-o	−0.22	−0.22	**-0.22**	+0.81	−1.01	**+2.62**
JanusFlow	−1.08	−1.72	**-0.43**	+1.51	+0.81	**+2.22**
UniWorld-V1	0.00	0.00	0.00	−1.71	0.00	**-3.43**
DiMOO	0.00	−2.59	**+2.59**	+2.02	+1.61	**+2.42**
Janus-Pro-1B	0.00	+6.90	**-6.90**	−5.24	−10.48	0.00
Show-o2-1.5B	0.00	−0.86	**+0.86**	−1.61	0.00	**-3.23**

Table 13: **Donor responses across requested counts.** Exact-count hits (%); superscripts give Swap minus Own.

Requested counts			*Frozen*		*Own*		*Swap*	
Recipient	Donor	Pairs	Recip.	Donor	Recip.	Donor	Recip.	Donor
BLIP3-o								
1–4	1–4	26	19.23	26.92	20.19	24.04	**25.96**$^{+5.77}$	**14.42**$^{-9.62}$
1–4	5–6	52	21.15	5.77	22.12	1.92	**27.40**$^{+5.29}$	**2.40**$^{+0.48}$
5–6	1–4	29	13.79	13.79	12.07	15.52	**12.93**$^{+0.86}$	**18.10**$^{+2.59}$
5–6	5–6	21	0.00	4.76	2.38	3.57	**10.71**$^{+8.33}$	**7.14**$^{+3.57}$
JanusFlow								
1–4	1–4	26	30.77	15.38	25.96	18.27	**29.81**$^{+3.85}$	**15.38**$^{-2.88}$
1–4	5–6	52	26.92	7.69	28.85	7.69	**32.69**$^{+3.85}$	**6.25**$^{-1.44}$
5–6	1–4	29	13.79	17.24	8.62	16.38	**6.90**$^{-1.72}$	**12.93**$^{-3.45}$
5–6	5–6	21	9.52	19.05	13.10	19.05	**17.86**$^{+4.76}$	**14.29**$^{-4.76}$
UniWorld-V1								
1–4	1–4	26	65.38	7.69	65.38	6.73	**64.42**$^{-0.96}$	**3.85**$^{-2.88}$
1–4	5–6	52	65.38	1.92	65.87	0.00	**65.38**$^{-0.48}$	**0.96**$^{+0.96}$
5–6	1–4	29	41.38	10.34	41.38	13.79	**42.24**$^{+0.86}$	**15.52**$^{+1.72}$
5–6	5–6	21	19.05	38.10	19.05	30.95	**22.62**$^{+3.57}$	**33.33**$^{+2.38}$
DiMOO								
1–4	1–4	26	69.23	7.69	69.23	11.54	**69.23**$^{0.00}$	**3.85**$^{-7.69}$
1–4	5–6	52	71.15	1.92	80.77	3.85	**78.85**$^{-1.92}$	**0.00**$^{-3.85}$
5–6	1–4	29	17.24	20.69	17.24	27.59	**17.24**$^{0.00}$	**17.24**$^{-10.34}$
5–6	5–6	21	23.81	19.05	19.05	9.52	**19.05**$^{0.00}$	**9.52**$^{0.00}$
Janus-Pro-1B								
1–4	1–4	26	26.92	3.85	23.08	15.38	**30.77**$^{+7.69}$	**15.38**$^{0.00}$
1–4	5–6	52	32.69	5.77	17.31	1.92	**17.31**$^{0.00}$	**3.85**$^{+1.92}$
5–6	1–4	29	3.45	17.24	6.90	3.45	**6.90**$^{0.00}$	**10.34**$^{+6.90}$
5–6	5–6	21	14.29	19.05	19.05	4.76	**0.00**$^{-19.05}$	**14.29**$^{+9.52}$
Show-o2-1.5B								
1–4	1–4	26	57.69	11.54	61.54	11.54	**65.38**$^{+3.85}$	**11.54**$^{0.00}$
1–4	5–6	52	53.85	5.77	55.77	3.85	**55.77**$^{0.00}$	**3.85**$^{0.00}$
5–6	1–4	29	27.59	10.34	24.14	13.79	**27.59**$^{+3.45}$	**6.90**$^{-6.90}$
5–6	5–6	21	23.81	9.52	38.10	4.76	**23.81**$^{-14.29}$	**9.52**$^{+4.76}$

Table 14: **Core state interventions.** Answer accuracy (%) and generation score (×100); superscripts give changes from frozen. Avg. equally weights 2D and 3D. The original archive omits some optimizer settings.

	Counting		*Spatial relations*			
Intervention	Answer	Gen.	Answer	Gen. 2D	Gen. 3D	Gen. Avg.
Frozen	59.18	39.54	79.88	14.07	26.61	20.34
Full core	**64.28**$^{+5.10}$	**38.28**$^{-1.26}$	**81.45**$^{+1.56}$	**14.45**$^{+0.38}$	**26.68**$^{+0.07}$	**20.56**$^{+0.22}$
Depth-group retention						
Random groups (2/6)	59.67	39.03	79.88	14.19	26.71	20.45
Complement (4/6)	59.42	38.78	80.13	14.16	26.68	20.42
Source-selected (2/6)	**61.43**$^{+2.25}$	**39.04**$^{-0.50}$	**80.22**$^{+0.34}$	**14.41**$^{+0.34}$	**26.78**$^{+0.17}$	**20.59**$^{+0.25}$
Singular components						
Top-64 components	60.30	38.99	79.98	14.28	26.89	20.59
Remaining components	60.23	39.77	80.13	14.62	26.85	20.73
Spectrum-matched random	59.38	39.18	79.64	14.65	26.88	20.77
Update direction						
Reflected update	61.04	39.16	79.35	13.97	27.35	20.66
Random-sign update	59.64	38.96	79.72	14.47	26.94	20.71
Fixed $-Q$	61.23	39.56	79.30	13.96	26.97	20.47

Table 15: **Spatial responses under core retention.** Mean changes on 512 questions and 1,000 generation prompts per run.

Contrast	**Answer Δ (pp)**	**Generation Δ (×100)**
Full − frozen	**-2.20**	**-0.01**
Selected − frozen	**-2.34**	**+0.04**
Selected − complement	−2.64	−0.18
Selected − random mean	−2.28	−0.07