IDENTITY MEETS INTERACTION: DIAGNOSING THE JOINT IDENTITY–INTERACTION RELIABILITY GAP IN REFERENCE-GUIDED HANDHELD PRODUCT INSERTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in instruction and reference-guided image editing enable models to follow complex instructions, transfer the appearance of reference content, and generate compelling visual results. However, editing guided by a reference object image requires more than a plausible composite: the object's identity-defining details also need to be preserved. This requirement is especially important in e-commerce, where product text, logos, and other identity cues carry commercial value. Existing models and benchmarks primarily focus on instruction following and overall visual quality, leaving the joint satisfaction of product identity and plausible hand–object interaction largely outside quantitative evaluation. To fill this gap, we introduce **HandHeldBench**, a diagnostic benchmark comprising 1,000 controlled-synthetic samples and 1,000 real-captured samples. We further propose **HandHeldEval**, an evaluation protocol for product scale, structural and high-frequency detail, identity similarity, and hand–object interaction plausibility. Results across mainstream general-purpose and task-specific editors on HandHeldBench show that current models cannot consistently achieve both plausible hand–object interaction and faithful preservation of product identity details. We define this recurring limitation as the *Joint Identity–Interaction Reliability Gap*. Beyond benchmarking, we construct **HandHeld-55k**, a training dataset of 54,923 validated editing quartets, and study reference-guided product-detail restoration as a strong post-processing baseline. Together, these components provide a systematic foundation for analyzing identity preservation and physical interaction in handheld product insertion.

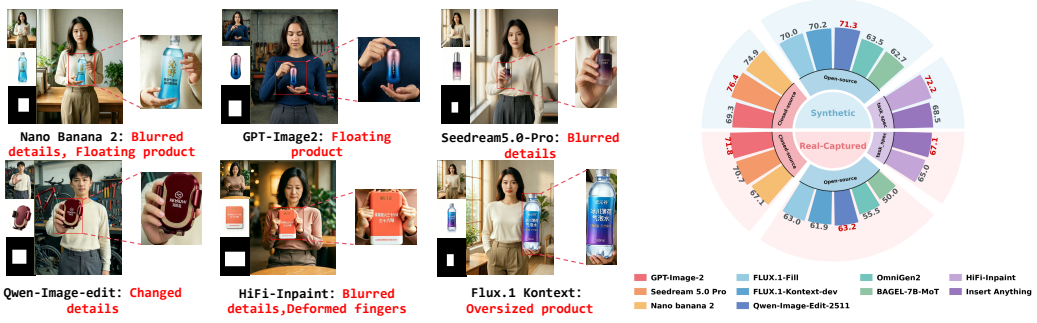

Figure 1: **Representative failures and overall task performance. Left:** typical failures in product identity, hand geometry, contact, and occlusion. **Right:** model performance on HandHeldBench; even strong general-purpose and task-specific editors remain weak when identity preservation and interaction plausibility are evaluated jointly.

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

1 INTRODUCTION

Recent advances in instruction- and reference-guided image editing have enabled complex instruction following and reference transfer (Brooks et al., 2023; Zhang et al., 2023a; Sheynin et al., 2024; Zhao et al., 2024; Yang et al., 2023; Chen et al., 2024b). Unified multimodal editors and proprietary systems further extend these capabilities to multi-image conditioning, text rendering, and iterative editing (Tian et al., 2025; Wu et al., 2025a; Black Forest Labs et al., 2025; OpenAI, 2026; Google, 2026; ByteDance Seed Team, 2026).

However, when reference-guided editing is used to place a product in a person's hand, visual plausibility alone is insufficient: the edited image should preserve the reference product's identity-defining details while producing a physically plausible hand–object interaction. In e-commerce applications, placing a reference product into a person's empty hand may appear convincing at full-image scale while corrupting text or logos, changing product geometry or scale, or producing implausible hand–object contact. These objectives are compatible in principle but operationally coupled because grasp synthesis modifies the same local geometry, boundaries, and occlusions that carry product identity. We define the *Joint Identity–Interaction Reliability Gap* as the inability of current editors to satisfy both requirements consistently in the same output. General-purpose benchmarks primarily assess instruction compliance, perceptual quality, or global similarity (Ku et al., 2024b; Ma et al., 2024; Krojer et al., 2024; Jia et al., 2026). Recent diagnostic benchmarks isolate physical realism, structured-content factuality, localized errors, or identity consistency (Pu et al., 2026; Zhuo et al., 2026; Sani et al., 2026; Xu et al., 2026; Luo et al., 2026), but the ability to preserve fine-grained product identity and realize plausible hand–object interaction jointly remains poorly understood. Representative failure modes and the corresponding task-level performance are summarized in Fig. 1.

To address this gap, we introduce **HandHeldBench**, a diagnostic benchmark for evaluating the Joint Identity–Interaction Reliability Gap. Figure 2 summarizes its complementary tracks. Both cover diverse product categories and scales, but follow deliberately different distributions: the controlled-synthetic track balances category coverage and generation conditions, whereas the real-captured track preserves natural variation in products and interactions. As Fig. 2 shows, real captures contain richer identity cues—including fine text, logos, packaging structures, and surface textures—together with more varied product forms and more complex hand–object contact, grasp, and occlusion. The two tracks therefore examine identity preservation and interaction plausibility under complementary controlled and real-world conditions.

Evaluating this joint capability reliably remains challenging. We therefore introduce **HandHeldEval**, which operationalizes the two axes through six verifiable sub-dimensions: hand plausibility, finger-count correctness, physical interaction consistency, product scale, shape, and fine-grained details such as text. It combines specialized quantitative metrics with instance-level human questions, avoiding the localized-error blindness and identity–quality conflation of general-purpose perceptual and vision-language-model evaluators (Ku et al., 2024a; Luo et al., 2026).

Beyond evaluation, we develop a product-detail restoration pipeline as a strong baseline and introduce **HandHeld-55k**, a synthetic training dataset containing 54,923 validated reference-guided editing quartets. Small products are particularly vulnerable to blurred packaging, corrupted logos, and missing text because they occupy few pixels.Our automated pipeline generates product pairs, reconstructs empty-hand sources, and optionally restores localized product detail while preserving the surrounding interaction.

We benchmark 10 open- and closed-source image editing models spanning diverse architectures and conditioning interfaces, together with two task-adapted variants. Their failures suggest mismatched inductive biases: whole-image editors may lose small identity cues under limited spatial and attention capacity, while grasp synthesis resynthesizes the same product–hand boundary and can overwrite text or texture. Conversely, localized or identity-focused editors may protect appearance but lack the geometric and contact priors needed for coherent finger geometry and occlusion. In summary, our contributions are as follows:

- We formalize the *Joint Identity–Interaction Reliability Gap* and introduce **HandHeldBench**, comprising complementary 1K controlled-synthetic and 1K real-captured test sets.

- We propose **HandHeldEval**, which reports six diagnostic metrics separately and uses fine-grained human questions to assess their alignment with human judgments.
- We construct **HandHeld-55k**, a large-scale quartet dataset with optional product-detail restoration, and adapt two representative backbones to improve scale control and joint identity–interaction reliability.

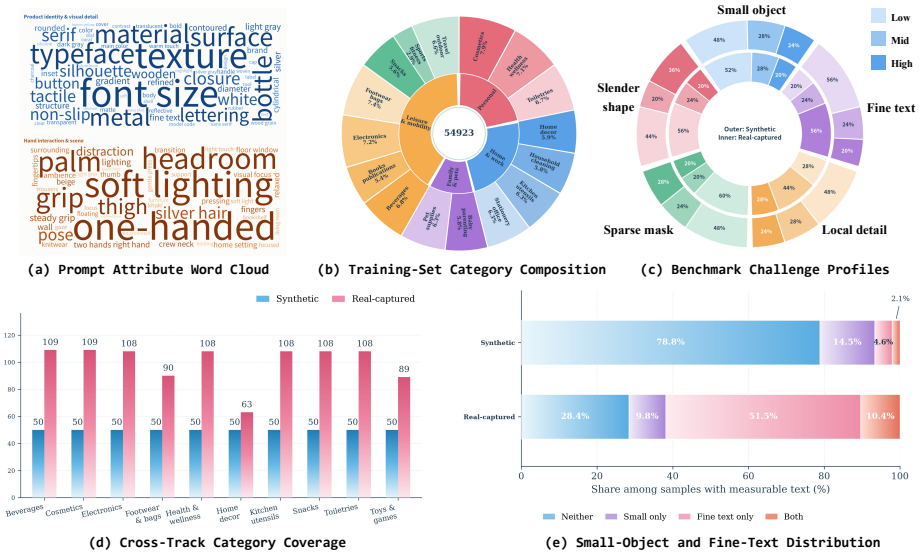

Figure 2: **Data profile of HandHeld-55k(a-b) and HandHeldBench(c-e).**

2 Related Work

Identity-preserving reference-guided editing. AnyDoor combines semantic and local features for zero-shot object transfer, HiFi-Inpaint adds high-frequency conditioning for human–product inpainting, and FlowFixer restores details lost under scale or viewpoint changes (Chen et al., 2024b; Liu et al., 2026b; Jun et al., 2026). These methods emphasize transfer fidelity under spatial placement or restoration, but typically assess identity apart from interaction geometry. A product may remain recognizable in a local crop while its grasp, contact, or occlusion is physically invalid; identity improvement alone therefore does not characterize joint reliability.

Interaction-aware generation and composition. Affordance Diffusion and GenHeld model hand layout and contact from object- or hand-conditioned cues, while HOComp combines pose guidance with appearance preservation for human-centric composition (Ye et al., 2023; Min & Sridhar, 2024; Liang et al., 2025). Their primary concern is plausible pose, affordance, and contact. Existing evaluations consequently capture interaction realism or coarse appearance, but rarely test whether small text, logos, and packaging structure survive at the final handheld scale. Better interaction can therefore coexist with unnoticed instance-level identity drift.

Benchmarks for image editing. General editing benchmarks cover broad tasks and quality dimensions (ImagenHub, I2EBench), physical or localized diagnostics (PICABench, ImagenWorld), and reasoning or spatial control (UnicBench, Inter-Edit) (Ku et al., 2024b; Ma et al., 2024; Pu et al., 2026; Sani et al., 2026; Ye et al., 2026; Liu et al., 2026a). However, their identity and physical-consistency axes remain largely disjoint: none uses one quartet to measure scale-aware reference identity and hand–object interaction on the same output, or evaluates reference-guided detail restoration (Table 1). They therefore cannot distinguish joint success from a model that merely shifts failure between identity, interaction, scale, and preservation. This missing coupled diagnosis motivates HandHeldBench and HandHeldEval.

Table 1: **Benchmark capability comparison.** Size counts evaluation samples only; training corpora are excluded. Check marks denote explicit capabilities. *Interaction* evaluates hand–object contact, grasp, and occlusion; *Detail post-proc.* denotes evaluated reference-guided restoration of product detail.

Benchmark	Size	Real data	Reference-guided	Identity	Scale	Interaction	Non-edit preservation	Detail post-proc.
ImagenHub (Ku et al., 2024b)	7×(100–200)	✓	✓	✓	×	×	×	×
I2EBench (Ma et al., 2024)	2K+	✓	×	×	✓	×	✓	×
PICABench (Pu et al., 2026)	900	✓	×	×	×	✓	×	×
ImagenWorld (Sani et al., 2026)	3.6K	✓	✓	×	×	×	×	×
UnicBench (Ye et al., 2026)	1.1K	✓	×	×	×	×	✓	×
Inter-Edit (Liu et al., 2026a)	6.25K	✓	×	×	×	×	✓	×
HandHeldBench	**2K**	✓	✓	✓	✓	✓	✓	✓

3 Benchmark, Evaluation, and Data Construction

This section presents HandHeldBench and its two test tracks, HandHeldEval, and HandHeld-55k with localized detail restoration and two backbone adaptations.

3.1 HandHeldBench

HandHeldBench comprises two separately reported 1K tracks (Fig. 2): controlled synthesis balances 20 categories and known construction variables, while 10 real-captured categories preserve natural framing, fine text, and appearance variation.

Evaluation task. Given an empty-hand source I_s, product reference I_r, interaction region M, and instruction p, an editor predicts $\widehat{I} = f(I_s, I_r, M, p)$. It should adapt pose, scale, illumination, and occlusion while preserving product identity, plausible contact, and the surrounding scene. We score these requirements separately; M is supplied only to models with native spatial conditioning.

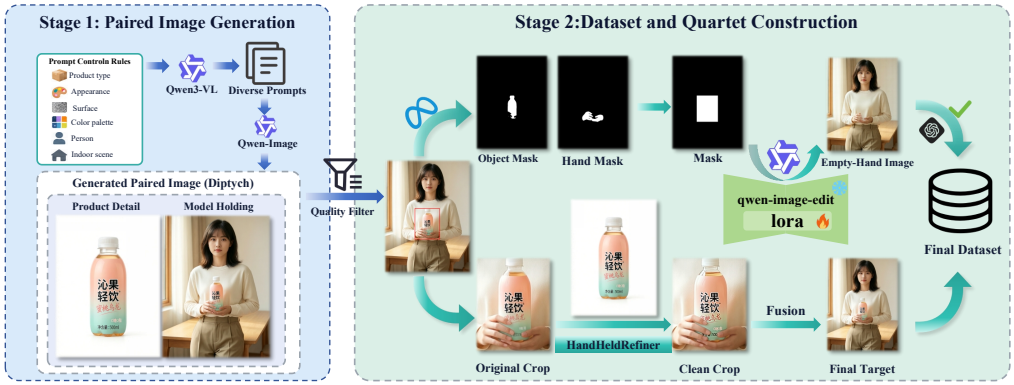

Figure 3: **Synthetic training-quartet construction.** Structured product, person, grasp, and scene controls are expanded into prompts and paired product-reference/handheld candidates. Quality and interaction checks retain consistent pairs; SAM 3 supplies the interaction mask, while the LoRA-trained Qwen-Image-Edit-2511 module reconstructs the empty-hand source (see Appendix C.2). The validated source–reference–mask–target contract retains the target and records any enhanced target separately.

3.2 Data Curation

Controlled synthetic track. Qwen3-VL expands structured product, person, grasp, and scene controls; Qwen-Image-2512 then renders paired reference/handheld images (Bai et al., 2025; Wu et al., 2025a). Pairing establishes product identity and plausible interaction before SAM 3 localization and empty-hand reconstruction with adapted Qwen-Image-Edit-2511 (Carion et al., 2025). Hard filters remove incomplete content, implausible contact, duplicates, identity drift, and artifacts

before deterministic selection of 50 localizable samples per category. Reconstructor details appear in Appendix C.2.

Real-captured track. From more than 5,000 captured product and handheld photographs (Fig. 4), we normalize framing, reconstruct only the empty-hand source, and reject product residue, invalid hands, incomplete people, contaminated references, and scene inconsistency. The resulting 1,000 quartets contain 918 unique references; references and targets remain captured, while repeated product views are grouped during resampling.

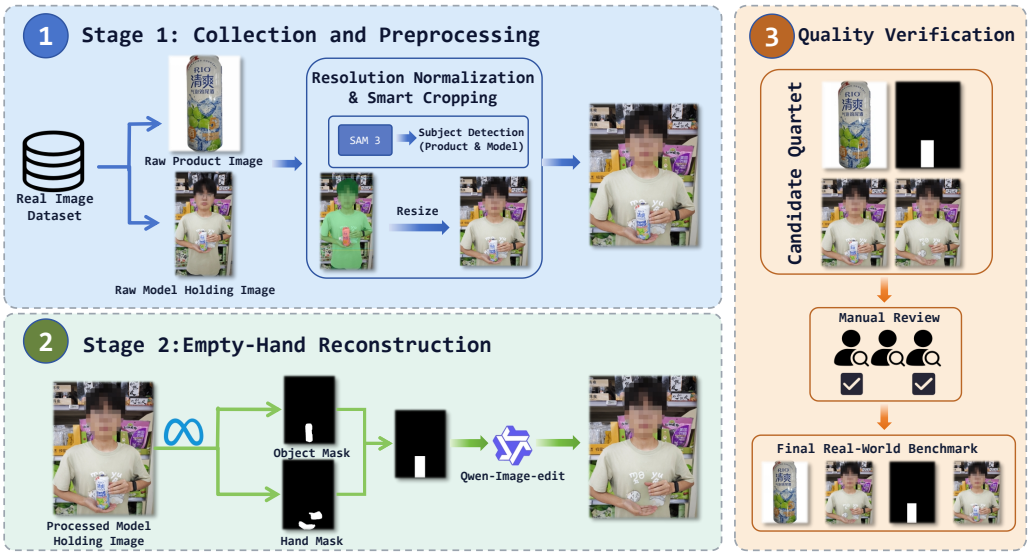

Figure 4: **Construction of the real-captured benchmark track.** Captured product references and handheld targets are normalized, SAM 3 localizes the held product and interaction region, and Qwen-Image-Edit-2511 reconstructs only the empty-hand source. Automated validity checks and final review select 1,000 quartets from more than 5,000 captured images while keeping the references and targets real.

3.3 HANDHELDEVAL

HandHeldEval follows the two axes of the reliability gap and the sequence in which an edit can fail. The metric suite jointly reflects the six sub-dimensions introduced above rather than assigning one metric to each. On the product-identity axis, we first ask whether the inserted product has a physically appropriate *scale*; **RSE** is the predicted-to-target product-area ratio, where 1 denotes exact scale agreement. We then remove the advantage of simply drawing a larger object by resizing the reference to the target scale. At that scale, **TS-KR**$= 100\sqrt{rc}$ measures *shape and structural identity* through retained local matches r and their spatial coverage c, whereas **A-HFSSIM** measures *fine-grained detail* by aligning the prediction with the same correspondences and comparing high-frequency residuals only on visible, non-hand-occluded product support (Jiang et al., 2024; Wang et al., 2004). **DINO** additionally records full-image semantic similarity to `gt_full`, but is not treated as exact logo or text verification (Siméoni et al., 2025).

On the interaction axis, a successful result requires a plausible hand, correct finger count, and physically coherent contact, support, non-penetration, and occlusion. **HOIS** turns these requirements into seven region-grounded atomic questions; for question i, the Qwen3-VL logits over *Plausible*, *Minor*, and *Implausible* give $S_i = 100[p_i(\text{Plausible}) + 0.5p_i(\text{Minor})]$, and HOIS averages the seven scores (Bai et al., 2025). Finally, because a locally successful insertion should not damage the person or scene, **NEPS** evaluates the non-edit region as the geometric mean of pixel, SSIM, and LPIPS preservation outside a safety-dilated mask.

Each metric retains its own valid denominator and status; not-matchable, not-applicable, and evaluation-error cases are never converted to zero. Scores are first averaged within category and then

macro-averaged across categories, with category-stratified bootstrap intervals and product-grouped resampling for repeated real references. Full formulas, thresholds, and failure rules are provided in Appendix B; human validation is specified with the experiments and Appendix E.1.

3.4 HANDHELD-55K

Closing the training gap. Controlled synthesis offers balance but limited natural packaging detail; real captures offer richer products and interactions but not scalable task supervision. HandHeld-55k fills this gap with quartets in which identity and interaction are both trustworthy.

Reverse construction for scalable supervision. Forward construction would force the data generator to solve the same unreliable joint task. We instead render a paired product reference and handheld target, filter identity and interaction failures, localize the region with SAM 3, and reconstruct the empty-hand source using Qwen-Image-Edit-2511 adapted on 672 removal pairs. Establishing interaction first and applying the more constrained removal step enables scalable, auditable source–reference–mask–target quartets (Appendix C.2).

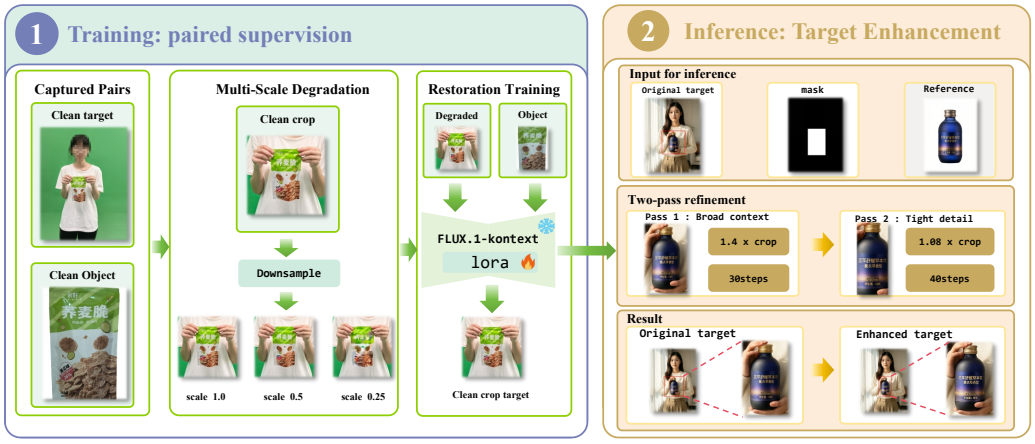

Figure 5: **Reference-guided product-detail restoration. Training:** SAM 3-localized target and reference crops are normalized, and three-scale SDXL degradation creates pseudo-paired [degraded, reference] $\rightarrow$ clean supervision for a rank-128 FLUX.1-Kontext LoRA. **Inference:** broad-context and tight-detail passes restore packaging structure and texture, after which mask-constrained blending maps the refined crop back only within product support. Original and enhanced targets remain separately addressable.

Localized restoration for identity detail. Because handheld products occupy few pixels, full-image generation may blur identity cues, whereas full-target repair may disturb a valid grasp. Inspired by FlowFixer (Jun et al., 2026), multi-scale crops retain context and emphasize packaging detail, then mask-constrained blending replaces only product support (Fig. 5). Thus interaction is established first and identity restored locally; original and enhanced targets remain separate (Appendix C.3).

HandHeld-55k contains 54,923 filtered quartets across 16 categories, with optional restored supervision and strict benchmark exclusion (Brooks et al., 2023; Zhang et al., 2023a; Chen et al., 2024b; Liang et al., 2025; Liu et al., 2026b). Rank-64 LoRA adaptation of FLUX.1-Kontext-dev and Qwen-Image-Edit-2511 (with spatial control) improves scale control and joint reliability (Black Forest Labs et al., 2025; Wu et al., 2025a; Hu et al., 2022a; Zhang et al., 2023b); Appendix C gives optimization details.

4 EXPERIMENTS

4.1 EVALUATION DETAILS

We evaluate ten native editors—GPT-Image-2, Nano banana 2, Seedream 5.0 Pro, FLUX.1-Fill, FLUX.1-Kontext-dev, BAGEL-7B-MoT, OmniGen2, Qwen-Image-Edit-2511, HiFi-Inpaint, and Insert Anything—plus LoRA-adapted Kontext and Qwen variants (OpenAI, 2026; Google, 2026; ByteDance Seed Team, 2026; Black Forest Labs et al., 2025; Deng et al., 2025; Wu et al., 2025b;a; Zhang et al., 2023b; Liu et al., 2026b; Song et al., 2026). All receive the same source, reference, and instruction; only native interfaces receive masks. We generate one normalized 780×1024 output without reranking.

We macro-average the six HandHeldEval metrics by category while preserving invalid states. Overall is only a ranking aid: $0.20S_{\text{scale}} + 0.15\text{TS-KR} + 0.15\text{A-HFSSIM} + 0.30\text{HOIS} + 0.10\text{NEPS} + 0.10S_{\text{DINO}}$, where $S_{\text{scale}} = 100e^{-|\ln \text{RSE}|/2}$ and $S_{\text{DINO}} = 100\text{DINO}$. Human evaluation uses a fixed stratified 10% subset with two ratings averaged per image. Appendices D and E.1 give full protocols.

4.2 BENCHMARK RESULTS

Tables 2 and 3 reveal the Joint Identity–Interaction Reliability Gap as a family of distinct failure profiles rather than a uniform quality deficit. No model dominates scale, structure, high-frequency detail, interaction, and preservation simultaneously; similar aggregate performance can conceal fundamentally different weaknesses.

Table 2: Results on the 1,000-sample controlled-synthetic track. For RSE, values closer to 1 are better; all other metrics are higher-is-better. The best and second-best results are highlighted in red and blue, respectively. For Δ Improvement, positive values denote improvement (reduced absolute log deviation from 1 for RSE; adapted minus base otherwise).

Model	NEPS↑	RSE→ 1	TS-KR↑	A-HFSSIM↑	HOIS↑	DINO↑	Overall↑
FLUX.1-Fill	98.47	0.972	51.31	67.34	47.44	0.8421	70.02
FLUX.1-Kontext-dev	72.41	1.974	72.41	72.51	62.73	0.8161	70.19
HiFi-Inpaint	99.30	1.009	48.32	62.99	58.10	0.8256	72.22
Qwen-Image-Edit-2511	99.01	0.876	51.15	67.38	54.77	0.8467	71.30
Insert Anything	99.37	1.488	59.36	68.60	47.52	0.8753	68.54
OmniGen2	60.57	1.980	60.09	63.28	56.57	0.7782	63.53
BAGEL-7B-MoT	67.40	1.785	51.50	60.84	54.42	0.7770	62.66
GPT-Image-2	94.39	0.440	51.74	71.74	66.60	0.8073	69.28
Seedream 5.0 Pro	93.07	0.828	63.14	68.65	67.28	0.8935	76.39
Nano banana 2	87.67	0.845	61.67	78.00	60.02	0.8755	74.86
FLUX.1-Kontext-dev + LoRA	97.27	0.959	57.20	62.27	54.32	0.9249	72.78
Δ Improvement	+24.86	+0.638	-15.21	-10.24	-8.41	+0.1088	+2.58
Qwen-Image-Edit-2511 + LoRA	99.13	0.927	56.33	70.12	57.12	0.8956	74.23
Δ Improvement	+0.12	+0.057	+5.18	+2.74	+2.35	+0.0489	+2.93

Scale can masquerade as identity. FLUX.1-Kontext yields strong local structure and detail readings while exhibiting the largest product-size error on both tracks. Oversized products expose more matchable pixels and visible high-frequency evidence, making local identity scores appear deceptively strong. This profile motivates target-scale normalization, but normalization does not make the generated size correct; TS-KR and A-HFSSIM must therefore be interpreted jointly with RSE.

Identity strength does not imply plausible interaction. Insert Anything exhibits one of the strongest real-track identity profiles among open models, yet its interaction score remains low. Its insertion-specific bias preserves recognizable product evidence without reliably resolving grasp, contact, and occlusion. The result confirms that success on the identity axis cannot stand in for joint task reliability.

Table 3: Results on the 1,000-sample real-captured track. For RSE, values closer to 1 are better; all other metrics are higher-is-better. The best and second-best results are highlighted in red and blue, respectively. For Δ Improvement, positive values denote improvement (reduced absolute log deviation from 1 for RSE; adapted minus base otherwise).

Model	NEPS↑	RSE→ 1	TS-KR↑	A-HFSSIM↑	HOIS↑	DINO↑	Overall↑
FLUX.1-Fill	97.02	0.909	46.35	50.81	39.13	0.7930	63.01
FLUX.1-Kontext-dev	72.67	2.242	74.33	61.08	45.37	0.7357	61.90
HiFi-Inpaint	97.84	0.990	43.72	48.64	46.80	0.7456	65.03
Qwen-Image-Edit-2511	96.44	1.031	51.04	57.38	32.01	0.8035	63.24
Insert Anything	98.85	1.231	65.21	61.39	38.70	0.8602	67.11
OmniGen2	64.62	1.881	57.96	51.26	38.76	0.6454	55.51
BAGEL-7B-MoT	63.94	1.777	39.20	42.06	32.96	0.6503	49.98
GPT-Image-2	90.41	0.901	68.72	59.15	54.56	0.8198	71.77
Seedream 5.0 Pro	88.32	1.268	69.04	58.17	55.17	0.8497	70.72
Nano banana 2	80.01	1.452	65.26	61.11	51.39	0.8132	67.10
FLUX.1-Kontext-dev + LoRA	94.49	1.271	60.07	47.58	41.46	0.8075	63.85
Δ Improvement	+21.82	+0.568	-14.26	-13.50	-3.91	+0.0718	+1.95
Qwen-Image-Edit-2511 + LoRA	97.96	0.996	56.28	59.77	35.62	0.8443	66.29
Δ Improvement	+1.52	+0.027	+5.24	+2.39	+3.61	+0.0408	+3.05

Proprietary editors have different strengths. The proprietary systems are stronger overall, but their advantages are not interchangeable. GPT-Image-2 combines strong scale control with interaction quality, Seedream is strongest on interaction while retaining competitive identity, and Gemini is especially strong in aligned high-frequency detail.

Conditioning and adaptation shape distinct reliability profiles. Whole-image editors reconstruct products flexibly but may alter unrelated content, whereas mask-conditioned editors better preserve context without guaranteeing identity or plausible contact. Task adaptation likewise remains backbone-dependent: Kontext mainly corrects oversizing and non-edit changes, with reduced local identity scores at realistic scale, while Qwen improves structure, detail, and interaction jointly. Thus, HandHeld-55k narrows the joint gap without eliminating each backbone's conditioning and generation biases.

4.3 Human Evaluation

We score a fixed category- and scale-stratified 10% subset, with two annotators per image. Ratings cover non-edit preservation, scale, product detail, hand–object contact, and overall quality; complete model-by-axis results are reported in Appendix E.2.

Human judgments favor balanced reliability. Proprietary editors remain strongest overall, with Seedream, GPT-Image-2, and Nano banana 2 leading through different combinations of contact, detail, and scale. Open models remain uneven: Qwen is preferred on controlled synthesis, Insert Anything on real captures, and OmniGen2 and BAGEL are weakest. Qwen adaptation improves every human axis on both tracks, most clearly contact and quality, while Kontext adaptation mainly repairs scale and preservation. Yet Insert Anything retains stronger real-product detail despite weaker contact, showing that adaptation narrows rather than removes the joint gap.

For model-level metric validation, we pair each automatic score with the corresponding human-axis mean over the same 12 models. Table 4 reports Spearman ρ and Kendall τ_b with model-bootstrap 95% intervals; Fig. 6 visualizes ρ.

Metrics reflect human preference within their scopes. All associations retain the expected direction after Qwen adaptation: RSE closeness aligns with human scale; and NEPS most strongly on real captures. HOIS remains weaker because crop questions isolate anatomy and contact, whereas raters judge the full scene; conversely, DINO cannot verify exact text or contact. The metrics are therefore complementary, while broad model-bootstrap intervals limit claims to model ordering.

Table 4: Model-level alignment between automatic metrics and human judgments. Brackets contain 95% bootstrap confidence intervals over models. Stars on Spearman ρ denote two-sided permutation significance: $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.

Metric	Synthetic ρ [95% CI]	Synthetic τ_b [95% CI]	Real ρ [95% CI]	Real τ_b [95% CI]
NEPS	$+0.245$ $[-0.683, +0.772]$	$+0.107$ $[-0.528, +0.594]$	$+0.781^{**}$ $[+0.226, +0.979]$	$+0.626$ $[+0.143, +0.933]$
RSE	$+0.361$ $[-0.339, +0.782]$	$+0.198$ $[-0.317, +0.639]$	$+0.557$ $[-0.108, +0.848]$	$+0.382$ $[-0.081, +0.700]$
TS-KR	$+0.601^{*}$ $[+0.032, +0.893]$	$+0.424$ $[-0.016, +0.793]$	$+0.853^{***}$ $[+0.524, +1.000]$	$+0.727$ $[+0.390, +1.000]$
A-HFSSIM	$+0.678^{*}$ $[+0.140, +0.993]$	$+0.576$ $[+0.138, +0.966]$	$+0.699^{*}$ $[+0.226, +0.940]$	$+0.515$ $[+0.129, +0.862]$
HOIS	$+0.441$ $[-0.217, +0.804]$	$+0.273$ $[-0.186, +0.621]$	$+0.517$ $[-0.200, +0.883]$	$+0.364$ $[-0.167, +0.759]$
DINO	$+0.650^{*}$ $[+0.029, +0.972]$	$+0.545$ $[+0.082, +0.903]$	$+0.909^{***}$ $[+0.616, +1.000]$	$+0.788$ $[+0.508, +1.000]$

Qwen's joint gains further show that identity and interaction can improve together, yet persistent detail–contact imbalances across models support a reliability gap rather than an intrinsic trade-off.

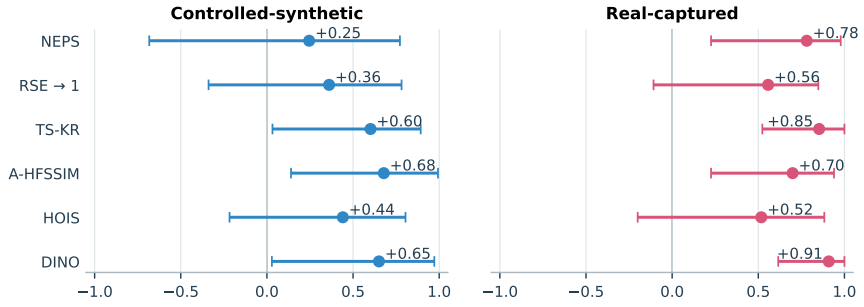

Figure 6: Model-level Spearman alignment between automatic metrics and their corresponding human axes. Points are correlations across 12 models and bars are model-bootstrap 95% confidence intervals. RSE is represented as negative absolute log deviation from 1, so positive association means that closer-to-one values agree with higher human scale ratings.

5 Limitations and Future Directions

HandHeldBench exposes three limitations of current editors. First, product identity remains entangled with scale: enlarging the inserted product can preserve more text and texture while violating scene-consistent size. Future models should combine scale-aware reference encoding with explicit layout constraints. Second, identity-focused and mask-conditioned editors may preserve products and surrounding content yet still fail finger geometry, contact, grasp, and occlusion. Interaction-aware architectures should therefore reason jointly about anatomy, affordance, and physical support. Third, whole-image editors often improve holistic realism by resynthesizing the scene, but may alter the person, clothing, illumination, or background. Future systems should couple global reasoning with localized controls that protect non-edit regions. Overall, progress requires joint optimization of identity, interaction, scale, and context rather than improving any dimension in isolation.

6 Conclusion

In this work,we introduce HandHeldBench and HandHeldEval to diagnose the Joint Identity–Interaction Reliability Gap in reference-guided handheld product insertion. Across both tracks, models exhibit distinct failures: oversized products inflate identity measures, identity-preserving insertions fail contact, and preservation depends strongly on different conditioning strategies. Human judgments favor joint reliability across identity, interaction, scale, and context, while metric–human alignment supports decomposed evaluation. HandHeld-55k turns this diagnosis into task-specific supervision through reverse quartet construction and localized detail restoration; adapting two representative backbones improves scale control and joint reliability. Together, these resources advance identity-faithful, physically credible product insertion.

AI USE STATEMENT

Generative AI was used solely to aid and polish the writing of this paper. It was not used for research design, data construction, experimentation, evaluation, analysis, or human annotation. The authors reviewed all AI-assisted edits and assume full responsibility for the final submission.

ETHICS STATEMENT

This work involves images of people and branded products, raising privacy, consent, intellectual-property, and representation concerns. Release is restricted to assets with compatible collection terms and licenses; identifying metadata is removed and access controls are applied where appropriate. Human evaluation follows informed participation, minimal necessary exposure, and aggregate reporting.

REPRODUCIBILITY STATEMENT

The main paper specifies the benchmark tracks, quartet contract, metrics, model roster, and human-evaluation design. The appendix documents preprocessing, failure handling, inference settings, metric validity, and training configurations. The release package links results to frozen manifests, immutable sample identifiers, and audited run records.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

ByteDance Seed Team. Beyond generation, it understands design: Introducing seedream 5.0 pro. https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, 2026. Released July 8, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, et al. Sam 3: Segment anything with concepts, 2025.

Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Virtualmodel: Generating object-id-retentive human-object interaction image by diffusion model for e-commerce marketing, 2024a. URL https://arxiv.org/abs/2405.09985.

Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17448–17460, 2025.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602, 2024b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026. Gemini 3.1 Flash Image; released February 26, 2026.

Jiun Tian Hoe, Weipeng Hu, Wei Zhou, Chao Xie, Ziwei Wang, Xudong Jiang, Yap-Peng Tan, and Chee Seng Chan. Reference-free human-object interaction editing. *Neurocomputing*, pp. 134869, 2026. ISSN 0925-2312. doi: 10.1016/j.neucom.2026.134869. URL https://doi.org/10.1016/j.neucom.2026.134869.

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. In *The Tenth International Conference on Learning Representations*, 2022a.

Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, Lei Bai, Wanli Ouyang, Lin Chen, Fei Zhao, Yao Hu, Zihan Wang, Yuan Xie, and Shaohui Lin. Compbench: Benchmarking complex instruction-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19865–19875, 2024.

Jinyoung Jun, Won-Dong Jang, Wenbin Ouyang, Raghudeep Gadde, and Jungbeom Lee. Flowfixer: Towards detail-preserving subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22049–22058, 2026.

Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1202.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.663.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Dong Liang, Jinyuan Jia, Yuhao Liu, and Rynson W. H. Lau. Hocomp: Interaction-aware human-object composition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 108272–108308, 2025. doi: 10.52202/085713-3263.

Delong Liu, Haotian Hou, Zhaohui Hou, Zhiyuan Huang, Shihao Han, Mingjie Zhan, Zhicheng Zhao, and Fei Su. Inter-edit: First benchmark for interactive instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37290–37300, 2026a.

Yichen Liu, Donghao Zhou, Jie Wang, Xin Gao, Guisheng Liu, Jiatong Li, Quanwei Zhang, Qiang Lyu, Lanqing Guo, Shilei Wen, Weiqiang Wang, and Pheng-Ann Heng. Hifi-inpaint: Towards high-fidelity reference-based inpainting for generating detail-preserving human-product images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1994–2004, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online RL for image editing via high-fidelity reward modeling. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41494–41516, 2024. doi: 10.52202/079017-1313.

Chaerin Min and Srinath Sridhar. Genheld: Generating and editing handheld objects, 2024.

OpenAI. System card: Chatgpt images 2.0 and thinking mode. https://deploymentsafety.openai.com/chatgpt-images-2-0/chatgpt-images-2-0.pdf, 2026. Released April 21, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, Kaiwen Zhu, Shuo Cao, Bin Fu, Si Liu, Hongsheng Li, Yu Qiao, Wenlong Zhang, Xi Chen, and Yihao Liu. PICABench: How far are we from physically realistic image editing? In *The Fourteenth International Conference on Learning Representations*, 2026.

Samin Mahdizadeh Sani, Max Ku, Nima Jamali, Matina Mahdizadeh Sani, Paria Khoshtab, Wei-Chieh Sun, Parnian Fazel, Zhi Rui Tam, Thomas Chong, Edisy Kin Wai Chan, Donald Wai Tong Tsang, Chiao-Wei Hsu, Ting Wai Lam, Ho Yin Sam Ng, Chiafeng Chu, Chak-Wing Mak, Keming Wu, Hiu Tung Wong, Yik Chun Ho, Chi Ruan, Zhuofeng Li, I-Sheng Fang, Shih-Ying Yeh, Ho Kei Cheng, Ping Nie, and Wenhu Chen. Imagenworld: Stress-testing image generation models with explainable human evaluation on open-ended real-world tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8871–8879. IEEE, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. DINOv3, 2025.

Wensong Song, Hong Jiang, Zongxin Yang, Zheqiao Cheng, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9097–9105, 2026. doi: 10.1609/aaai.v40i11.37866.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18310–18319, 2023.

Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. Mige: Mutually enhanced multimodal instruction-based image generation and editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10622–10631, 2025. doi: 10.1145/3746027.3755811.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, et al. Qwen-image technical report, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation, 2025b.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Withanyone: Toward controllable and id consistent image generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhu Xu, Qingchao Chen, Yuxin Peng, and Yang Liu. Semantic-aware human object interaction image generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54918–54932. PMLR, 2024.

Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. Hoi-swap: Swapping objects in videos with hand-object interaction awareness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2454.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Keming Ye, Zhipeng Huang, Canmiao Fu, Qingyang Liu, Jiani Cai, Zheqi Lv, Chen Li, Jing Lyu, Zhou Zhao, and Shengyu Zhang. UnicEdit-10M: A dataset and benchmark breaking the scale–quality barrier via unified verification for reasoning-enriched edits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37279–37289, 2026.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22479–22489, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023a. doi: 10.52202/075280-1365.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. doi: 10.52202/079017-0100.

Hang Zhou, Xinxin Zuo, Sen Wang, and Li Cheng. Pics: Pairwise image compositing with spatial interactions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=zNCNEOhKps`.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, and Hongsheng Li. Factuality matters: When image generation and editing meet structured visuals. In *The Fourteenth International Conference on Learning Representations*, 2026.



Figure 7: **Representative HandHeldBench quartets.** Each row shows, from left to right, the interaction mask, product reference, reconstructed empty-hand source, and paired handheld target. The examples illustrate variation in product geometry, scene, and grasp while following the shared quartet contract.

Table 5: Roles in a HandHeldBench quartet. The target is available to the evaluator but is never supplied to the editing model.

Role	Definition
Source	A 780×1024 RGB image of the person without the product. It is reconstructed from the paired handheld image while preserving identity, clothing, viewpoint, and background.
Reference	A product image used as the identity condition. Real-track references are captured product photographs; synthetic references are the product panel paired with the rendered handheld target.
Mask	A binary interaction region in source coordinates. White denotes the region in which the product and necessary hand contact may be generated. It is also used to define the non-edit diagnostic region, even for models that do not accept a mask.
Target	A handheld-product image used for scale anchoring, product localization, and evaluation. It is captured in the real track and rendered in the controlled-synthetic track.
Prompt	A task instruction that requests natural contact, plausible hand anatomy, product-detail preservation, realistic scale, and preservation outside the edited region.

A Benchmark Details

A.1 Frozen tracks and sample contract

HandHeldBench contains independently frozen 1K controlled-synthetic and real-captured tracks. The former has 50 samples in each of 20 categories; the latter spans ten shared categories and 918 unique product references. Repeated references correspond to distinct captured handheld views and are grouped during resampling.

Table 6: Provenance and freezing rules for the two HandHeldBench tracks. *Generated* denotes controlled rendering, whereas *reconstructed* denotes product removal from the paired target.

Track	Source	Reference	Target	Mask
Controlled synthetic	Product-removed reconstruction of the paired rendered target	Generated isolated product panel	Generated handheld scene under structured controls	Product–hand interaction box derived from SAM 3 regions
Real captured	Product-removed reconstruction of the captured handheld target	Captured product photograph	Captured handheld photograph	Product-centered editing box expanded to cover contact boundaries

Each manifest row stores an immutable identifier, category, role paths, dimensions, and hashes. Training audits exclude both benchmark parent pools and require zero overlap; missing localization or evaluator failures remain explicit states rather than zero scores.

A.2 Construction and provenance

Controlled samples are generated as paired isolated-product and handheld panels under structured product, person, grasp, and scene controls. Hard gates check completeness, contact, hand anatomy, cross-panel identity, localizability, and source reconstruction before selecting 50 samples per category. Real samples undergo EXIF correction and person-preserving $780{:}1024$ cropping; references and targets remain captured, while only the product region is reconstructed to form the empty-hand source. Freezing records manifest and file hashes plus benchmark-parent exclusion lists.

A.3 Composition and quality control

Synthetic balancing isolates category effects; the real track preserves naturally uneven product size, text, viewpoint, and capture variation across beverages, cosmetics, electronics, footwear and bags, health and wellness, home decor, kitchen utensils, snacks, toiletries, and toys and games. The tracks are therefore reported separately.

Candidates are rejected for incomplete content, duplicate products, malformed people or hands, implausible contact or scale, identity drift, contaminated references, residual product content, or source–target inconsistency. Final audits verify unique identifiers, paths, dimensions, modes, hashes, and inherited training exclusions.

B HandHeldEval Definitions and Implementation

B.1 Inputs and metric states

For each sample, the evaluator receives source S, reference R, edit mask M, target T, and prediction P, normalized to 780×1024. Task, product, and hand masks retain distinct roles. Every metric records its value, denominator, status, and reason: `success` enters the mean; `not_matchable` and `not_applicable` remain outside it; generation failures and malformed evaluator assets are reported separately. Missing states are never imputed as zero.

B.2 Non-edit preservation and product scale

Let Ω_r be the complement of the edit mask dilated by $r = \max(5, \mathrm{round}(0.01 \min(H, W)))$; LPIPS uses $r_{\mathrm{deep}} = \max(32, \mathrm{round}(0.04 \min(H, W)))$. On images normalized to $[0, 1]$,

$$P_{\mathrm{pix}} = \mathrm{clip}(1 - \mathrm{mean}_{\Omega_r} |S - P|, 0, 1)\,, \tag{1}$$

$$P_{\mathrm{ssim}} = \mathrm{clip}(\mathrm{mean}_{\Omega_r} \mathrm{SSIM}(S, P), 0, 1)\,, \tag{2}$$

$$P_{\mathrm{deep}} = 1 - \mathrm{clip}\Big(\mathrm{LPIPS}_{\Omega_{r_{\mathrm{deep}}}}(S, P), 0, 1\Big)\,. \tag{3}$$

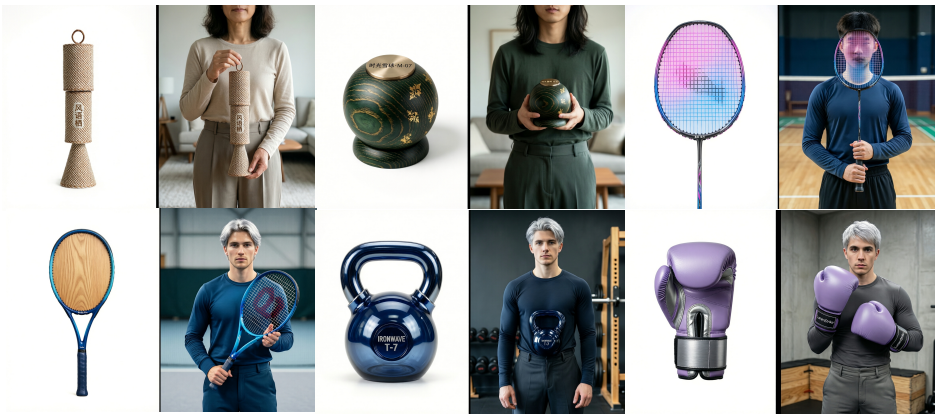

Figure 8: **Examples rejected during benchmark quality control.** The reference–candidate pairs illustrate failures such as product-identity drift, implausible scale or grasp, inconsistent geometry, or insufficient agreement between the isolated product and handheld target. Rejected candidates are excluded before track freezing.

SSIM uses complete 11×11 Gaussian windows ($\sigma = 1.5$), and LPIPS uses v0.1 AlexNet weights (Zhang et al., 2018). Their geometric mean is

$$\mathrm{NEPS} = 100\,(P_{\mathrm{pix}} P_{\mathrm{ssim}} P_{\mathrm{deep}})^{1/3}. \tag{4}$$

Non-finite components, empty masks, or insufficient safe support are evaluation errors.

For target and predicted product-area fractions A_{gt} and A_{pred},

$$\mathrm{RSE} = \frac{A_{\mathrm{pred}} + \epsilon}{A_{\mathrm{gt}} + \epsilon}, \tag{5}$$

where ϵ is one pixel. RSE$= 1$ denotes equal area, values below 1 indicate undersizing, and values above 1 indicate oversizing.

B.3 Target-scale identity

The reference crop is resized to the target-product scale before OmniGlue matching (Jiang et al., 2024). With retained-match ratio r and 4×4 spatial coverage c,

$$r = \min\left(\frac{K_{\mathrm{pred}}}{K_{\mathrm{self}}}, 1\right), \qquad c = \min\left(\frac{C_{\mathrm{pred}}}{C_{\mathrm{self}}}, 1\right), \qquad \text{TS-KR} = 100\sqrt{rc}. \tag{6}$$

OmniGlue uses confidence 0.02; fewer than 20 stable reference self-matches is *not matchable*.

A-HFSSIM reuses these correspondences to estimate a RANSAC similarity transform with a 3-pixel threshold. It compares signed residuals $H(I) = I - G_{\sigma=1.5}(I)$ over support shared by the reference product, aligned prediction, and non-hand-occluded region:

$$\text{A-HFSSIM} = 100 \max(\mathrm{mean}_{w \subset \Omega_{\mathrm{vis}}}\, \mathrm{SSIM}(H(R), H(P_{\mathrm{aligned}})), 0). \tag{7}$$

Fewer than 20 inliers, less than 1% visible support, or no complete 11×11 window is *not matchable*.

B.4 Hand–object interaction

A Qwen3-VL-32B judge answers seven region-grounded questions covering anatomy, proportion, contact, grasp, support, penetration, and occlusion (Bai et al., 2025). For question i,

$$S_i = 100\,[p_i(\text{Plausible}) + 0.5 p_i(\text{Minor})], \qquad \text{HOIS} = \frac{1}{7}\sum_{i=1}^{7} S_i. \tag{8}$$

The interaction crop is expanded by 20%; absence of a visible hand or product makes HOIS not applicable.

Table 7: Atomic HOIS questions and scoring anchors. Minor denotes a visible but non-catastrophic defect; Implausible requires clear visual evidence.

Q	Axis	Plausible	Minor	Implausible
1	Anatomy	Possible finger count and articulation	Slightly stiff or locally ambiguous finger	Extra/fused fingers or impossible joint direction
2	Proportion	Hand scale and palm/finger proportion fit perspective	Mild stretch or proportion error	Severe stretch, compression, or scale mismatch
3	Contact	Visible contact surface without gap	Weak or locally loose contact	Clear gap between hand and product
4	Grasp	Grasp suits product shape and weight	Unusual but feasible grasp	Physically impossible grasp point or pose
5	Support	Hand or another object visibly supports product	Support is ambiguous	Product visibly floats without support
6	Penetration	No impossible interpenetration	Boundary ambiguity only	Hand and product clearly pass through each other
7	Occlusion	Front/back ordering is geometrically coherent	Local ordering is ambiguous	Clear occlusion reversal

B.5 Full-image semantics and Overall

DINO is the cosine similarity between L2-normalized DINOv3 ViT-L/16 features of the prediction and `gt_full`; it measures full-image semantic agreement rather than exact text or localized identity (Oquab et al., 2024; Siméoni et al., 2025). The ranking-only Overall first maps scale and DINO to 0–100:

$$S_{\text{scale}} = 100 \exp\left(-\frac{|\ln \text{RSE}|}{2}\right), \qquad S_{\text{DINO}} = 100\,\text{DINO}, \tag{9}$$

then applies

$$\begin{aligned}\text{Overall} = 0.20 S_{\text{scale}} + 0.15\,\text{TS-KR} + 0.15\,\text{A-HFSSIM} \\ + 0.30\,\text{HOIS} + 0.10\,\text{NEPS} + 0.10 S_{\text{DINO}}.\end{aligned} \tag{10}$$

Rows missing any component receive no Overall value.

B.6 Aggregation

Metrics are computed in the order NEPS/RSE, TS-KR, A-HFSSIM, DINO, and HOIS, with states recorded independently. Scores are averaged within category and then macro-averaged. Confidence intervals use 2,000 category-stratified bootstrap repetitions; real-track resampling groups repeated product references, and paired model comparisons reuse identical indices.

C Data Pipeline and Training Details

C.1 HandHeld-55k construction

HandHeld-55k follows a shared source–reference–mask–target contract informed by object compositing and HOI generation work (Song et al., 2023; Zhou et al., 2026; Hu et al., 2022b; Xu et al., 2024; Chen et al., 2025; 2024a; Xue et al., 2024; Hoe et al., 2026). Structured product, person, grasp, clothing, and scene controls generate paired product-reference and handheld images. Samples pass completeness, anatomy, contact, identity, and scene-quality gates before SAM 3 localization, empty-hand reconstruction, optional detail restoration, and final integrity checks.

The release contains 54,923 quartets across 16 categories: 37,900 use enhanced targets and 17,023 retain accepted originals. It merges 41,692 primary-pool and 13,231 category-completion samples; the manifest records provenance and target type. Audits find 54,923 unique IDs, 219,692 decodable role files, binary 780×1024 masks, and zero missing files, duplicate IDs, empty/full masks, or benchmark-parent overlaps. Category counts and mask statistics are summarized in Fig. 2.

C.2 Empty-hand source reconstruction

Qwen-Image-Edit-2511 with Blockwise-ControlNet-Inpaint is adapted on 672 curated product-removal pairs. Only a rank-32 DiT LoRA is trained; the base editor, text encoder, VAE, and



Figure 9: **HandHeldEval diagnostics for one representative prediction.** The visualization exposes the task and localized product/interaction regions, target-scale crops used for structural and high-frequency comparison, and the seven atomic HOIS judgments. These intermediate views make scale, identity, and interaction errors auditable rather than reducing them to a single opaque score.

Table 8: Empty-hand reconstruction configuration used in both benchmark and training-quartet construction.

Item	Setting
Training pairs	672 curated product-removal pairs
Trainable parameters	Qwen-Image-Edit-2511 DiT LoRA, rank 32
Frozen components	Base editor, text encoder, VAE, and Blockwise ControlNet
Training schedule	60 epochs; selected checkpoint `step-5800`
Inference canvas	780×1024 padded to 784×1024, then cropped back
Sampling	40 steps; seed 1; denoising strength 1.0; mask blur 1
Mask	Product box expanded by 20 pixels; white is editable
Preservation	Exact copy from input outside the generation mask

ControlNet remain frozen. Reconstruction is limited to a product box expanded by 20 pixels, after which outside-mask pixels are copied exactly. The selected `step-5800` checkpoint is screened for residue, malformed fingers, seams, and scene changes.

C.3 Reference-conditioned target restoration

HandHeldRefiner uses 5,687 captured reference–target views from 3,531 product groups. SAM 3 crops are expanded by $1.6\times$, normalized to 768×768, and degraded with SDXL at timestep 50 and scales 1.0, 0.5, and 0.25. A rank-128 FLUX.1-Kontext-dev LoRA is trained with learning rate 10^{-4}, BF16, two data-parallel ranks, and the complete `step-28000` checkpoint. Broad and tight inference passes preserve interaction context while restoring text, logos, seams, and texture; fusion is restricted to product support, and original targets remain separately addressable.

C.4 Adaptation settings

Both task adaptations use the same 54,923 quartets and rank-64 LoRA. FLUX.1-Kontext-dev uses three ranks, one epoch, AdamW (Loshchilov & Hutter, 2019) at 10^{-4}, and checkpoint

Table 9: Task-adaptation settings. The two rows share data and LoRA rank but preserve their native conditioning interfaces.

Backbone	LoRA rank	GPUs	Micro / accum.	Global batch	LR
FLUX.1-Kontext-dev	64	3	1 / 1	3	10^{-4}
Qwen-Image-Edit-2511	64	5	1 / 4	20	5×10^{-5}



Figure 10: **Reference-guided product-detail restoration examples.** Each row presents the product reference, handheld target and localized region, original product crop, restored crop, and locally fused enhanced target. HandHeldRefiner restores text, logos, and texture while restricting fusion to product support, leaving the surrounding hand and scene unchanged.

`step-18308`. Qwen-Image-Edit-2511 keeps its base modules and ControlNet frozen, uses five ranks with global batch 20, AdamW at 5×10^{-5}, and the first-epoch `step-2747` checkpoint. Each adapted model otherwise retains its native conditioning and evaluation pipeline.

D Baseline Inference Setup

D.1 Prompts and input assembly

The common English instruction is:

> The model naturally holds the product in hand. The hand and the product are in natural contact; the hand is anatomically plausible and free of distortion. The product's details remain unchanged, and the product's size and position conform to reality and physics. Everything outside the edited region remains unchanged. Keep the person's face, body, clothing, background, lighting, camera view, and composition unchanged.

Composite-input models prepend "The left half is an empty-handed model photo, and the right half is the product reference." Separate-image models prepend "Image 1 is an empty-handed model photo, and Image 2 is the product reference." Unless stated otherwise, no negative prompt is used.

Table 10: Inference configurations used for the reported results. API denotes provider-managed sampling whose step, scheduler, or guidance is unavailable. CN denotes Blockwise ControlNet.

Model	Input assembly	Sampling / checkpoint	Native render and normalization
FLUX.1-Fill-dev	One left-source/right-reference composite plus a composite binary mask	35 steps; guidance 30; seed 1018	Composite resized to approximately 1M pixels; generated left half resized and cropped to the source canvas
FLUX.1-Kontext-dev	One left-source/right-reference composite; no mask input	50 steps for the reported run; embedded guidance 2.5; seed 1018	Condition long side at most 1536; native 784×1024 canvas cropped to 780×1024
HiFi-Inpaint	Separate reference, masked source, and binary mask	28 steps; guidance 3.5; seed 1018	Right-pad to 784×1024, then crop to 780×1024
Qwen-Image-Edit-2511	Separate source and reference; source and mask also condition ControlNet	40 steps; true CFG 4.0; ControlNet scale 1.25; denoising strength 1.0; seed 1018	Right-pad to 784×1024, then crop to 780×1024
Insert Anything	Reference, source, and binary mask; official mask-prompt path	Official FLUX.1-Fill/Redux release and mask-prompt LoRA checkpoint	Native mask-prompt output normalized once to the benchmark canvas
OmniGen2	Separate source/reference; no mask	35 steps; text/image guidance 5.0/2.0; seed 1018	Native source resolution, converted to 780×1024
BAGEL-7B-MoT	Reference, source, and instruction; no mask	35 steps; text/image CFG 4.0/2.0; shift 3.0; seed 1018	Native source resolution, converted to 780×1024
GPT-Image-2	Separate source/reference; no mask	API model `gpt-image-2`; medium quality; seed unavailable	API request 1168×1536; center-fit to 780×1024
Seedream 5.0 Pro, `doubao-seedream-5-0-pro-260628`	Separate source and reference; no mask input	API-managed; seed 1018; watermark disabled	API 2K output, center-fit with Lanczos to 780×1024
Nano banana 2, `gemini-3.1-flash-image`	Separate source and reference; no mask input	API-managed; seed unavailable	Gemini 3:4, 1K output, center-fit with Lanczos to 780×1024
FLUX.1-Kontext-dev + LoRA	Same as native Kontext	Native settings plus rank-64 `step-18308`; LoRA alpha 1.0	Identical to native Kontext
Qwen-Image-Edit-2511 + LoRA	Same as native Qwen-Image-Edit-2511	Native settings plus rank-64 `step-2747`; LoRA alpha 1.0	Identical to native Qwen-Image-Edit-2511

HiFi-Inpaint uses its native caption-style prompt: "A realistic photo of the same person naturally holding the product shown in the reference image. The product's shape, color, printed text, logo and texture match the reference. The person's face, clothing and background remain consistent." Qwen-Image-Edit-2511 uses a Chinese content-description prompt that asks for a complete, clearly rendered product, legible text and logo, realistic placement, and a normal hand grasping the product. Its Chinese negative list suppresses oversized products or hands, duplicated or malformed anatomy, penetration or floating, edits outside the mask, and changes to the person, clothing, background, composition, or viewpoint. The exact UTF-8 strings are released with the run configuration; this setting is a disclosed protocol deviation rather than a hidden attempt to make prompts appear identical.

D.2 Evaluated model configurations

Table 10 lists all native and adapted systems in the quantitative tables. Mask availability follows the model's native interface rather than being forced across architectures. Every successful output is converted to a 780×1024 RGB PNG with the deterministic path-specific rule shown below; no model output is reranked.

Open-weight systems use fixed local releases whose model and adapter checksums are recorded in the release manifest. Closed APIs record exact identifiers and invocation dates; unavailable scheduler, step, guidance, or seed fields are explicitly null.

Table 11: Human-evaluation rubric and normalization. Each dimension is mapped independently to 0–100; a raw score cannot be compared across rows before normalization.

Axis	Question	Range	Anchors
Non-edit	Are person, clothing, lighting, and background preserved outside the edit region?	1–5	1: major unrelated changes; 3: visible but limited changes; 5: preserved.
Scale	Is product size physically reasonable relative to the target scene?	1–3	1: clearly too large/small; 2: mild deviation; 3: reasonable.
Detail	Are shape, color, logo, text, texture, and packaging structure preserved?	1–4	1: identity evidence corrupted; 2: only coarse detail; 3: medium/large detail retained; 4: fine detail retained.
Contact	Are anatomy, finger count, contact, grasp, support, penetration, and occlusion plausible?	1–4	One point for each of plausible interaction, no deformation, correct finger count, and coherent overall hand.
Quality	Overall realism, composition, completeness, and visible artifacts.	1–5	1: very poor; 2: poor; 3: acceptable; 4: good; 5: very good.

The comparison preserves each system's recorded native configuration and spans 28-, 35-, 40-, and 50-step paths plus provider-managed APIs. It is a capability comparison, not an equal-compute ablation. The two adaptation comparisons are stricter: each adapted model retains its base model's input assembly, prompt, sampler, seed, output transform, and evaluation assets, with the LoRA checkpoint as the intended change.

E Human Evaluation

E.1 Protocol

For every evaluated model and track, we draw a fixed 10% subset with seed 20260921. The controlled-synthetic subset contains five samples from each of 20 categories; the real-captured subset contains ten from each of ten categories. This yields 100 outputs per model and track and preserves the same sample IDs across models. Sampling requires a decodable full prediction and the two displayed crops for every included model, so no model is evaluated on an easier denominator.

Ten raters participate. Each sampled output is assigned to two different raters, model identity and adaptation status are hidden, and task order is stably shuffled. An item displays source, product reference, full prediction, a product-detail crop, and a hand–object crop. Automatic scores and targets are not displayed as numeric hints. The two ratings are kept independently, and their arithmetic mean is the image-level score for each dimension.

For an axis with range $[m_k, M_k]$, a raw mean x_k is normalized as

$$h_k = 100 \frac{x_k - m_k}{M_k - m_k}. \tag{11}$$

Human Overall is the unweighted mean of the five normalized axes. Because the sampled category counts are exactly balanced within each track, the sample mean equals the category-macro mean. The detailed table includes all 12 evaluated models.

Automatic–human alignment is evaluated per axis: NEPS with non-edit, RSE with scale, TS-KR and A-HFSSIM with detail, HOIS with contact, and DINO with overall quality. Both sides are reduced to one value per model, giving $n = 12$ paired model points per track. We report Spearman ρ and Kendall τ_b (which accounts for ties), with 20,000 model-bootstrap repetitions and seed 20260924; permutation tests use 100,000 shuffles. RSE is represented as $-|\ln \mathrm{RSE}|$ for correlation, so positive association denotes agreement with human scale. These correlations validate model-level ordering, not per-image prediction, and the multiple axis tests are not interpreted as independent discoveries.

E.2 Detailed human-evaluation results

Tables 12 and 13 report all five normalized human axes and their unweighted mean. Each entry is the mean of two annotators over 100 images for the corresponding model and track.

Table 12: Human evaluation on the fixed controlled-synthetic 10% subset. Values are normalized to 0–100 and averaged over two annotators; Human Overall is their unweighted mean. The best and second-best Overall values are highlighted in red and blue.

Model	Non-edit	Scale	Detail	Contact	Quality	Human Overall
FLUX.1-Fill	92.50	72.00	49.33	68.33	55.25	67.48
FLUX.1-Kontext-dev	50.75	35.50	64.00	53.00	37.75	48.20
HiFi-Inpaint	88.00	48.50	45.00	60.67	43.25	57.08
Qwen-Image-Edit-2511	88.75	80.50	58.00	87.33	64.00	75.72
Insert Anything	89.75	57.00	62.33	52.33	48.25	61.93
OmniGen2	25.00	53.50	48.00	42.33	24.00	38.57
BAGEL-7B-MoT	51.00	45.50	39.00	51.33	31.25	43.62
GPT-Image-2	93.50	65.00	82.67	94.00	77.25	82.48
Seedream 5.0 Pro	95.75	87.50	92.33	92.33	86.50	90.88
Nano banana 2	95.00	86.50	80.00	86.33	76.75	84.92
FLUX.1-Kontext-dev + LoRA	95.75	86.50	64.33	87.67	71.50	81.15
Qwen-Image-Edit-2511 + LoRA	93.25	85.50	61.00	90.33	70.75	80.17

Table 13: Human evaluation on the fixed real-captured 10% subset. Values are normalized to 0–100 and averaged over two annotators; Human Overall is their unweighted mean. The best and second-best Overall values are highlighted in red and blue.

Model	Non-edit	Scale	Detail	Contact	Quality	Human Overall
FLUX.1-Fill	92.00	71.50	43.67	64.00	48.50	63.93
FLUX.1-Kontext-dev	61.50	36.50	66.67	66.33	41.75	54.55
HiFi-Inpaint	87.50	62.00	43.00	63.00	45.00	60.10
Qwen-Image-Edit-2511	95.00	75.00	53.00	72.33	57.50	70.57
Insert Anything	95.75	78.50	67.00	73.00	66.25	76.10
OmniGen2	33.75	33.00	41.33	26.33	17.25	30.33
BAGEL-7B-MoT	38.75	23.50	22.67	31.00	13.50	25.88
GPT-Image-2	91.75	81.50	79.00	87.00	76.25	83.10
Seedream 5.0 Pro	92.25	78.50	88.00	89.67	80.50	85.78
Nano banana 2	88.25	66.50	80.00	89.00	73.00	79.35
FLUX.1-Kontext-dev + LoRA	91.50	65.50	55.33	74.33	56.50	68.63
Qwen-Image-Edit-2511 + LoRA	95.75	76.00	54.00	76.33	62.25	72.87

The base-versus-adapted rows use the same sampled instances and scoring protocol. Both comparisons evaluate the complete adaptation recipe; isolating dataset size, LoRA rank, or target restoration requires separate one-factor-at-a-time runs.