DECOEDIT: BEYOND FLAT CONTROL FOR 3D-AWARE UNIFIED IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent advances in multimodal image editing have achieved impressive results, current frameworks remain largely confined to "flat" 2D controls (e.g., text prompts, bounding boxes, or masks). Precise 3D-aware geometric manipulation—such as coordinating an object's location, physical scale, and 3D orientation while strictly preserving its fine-grained identity—remains an elusive challenge, typically suffering from severe detail degradation or requiring expensive per-instance fine-tuning. In this paper, we present **DecoEdit**, an all-in-one image editing framework that goes beyond flat control by seamlessly integrating 1D textual, 2D spatial, and 3D volumetric inputs under a unified architecture. First, to power this unified model, we establish a data curation pipeline and curate a large-scale dataset, namely **DecoDataset**, containing 1.6 million samples, encompassing diverse multi-dimensional (1D/2D/3D) tasks and their complex compositions. Furthermore, we establish a comprehensive image editing benchmark named **DecoBench** to systematically evaluate multi-task editing performance. Finally, to address the fundamental trade-off between complex spatial transformations and identity preservation, we propose a novel decoupled pipeline: **DecoEdit** factorizes the editing process by utilizing structural guidance from SAM-3D to anchor the target geometry, scale, and spatial orientation, while concurrently leveraging a detail-preservation path to lock in high-fidelity textures. Extensive experiments demonstrate that **DecoEdit** achieves state-of-the-art results. Notably, on the challenging 3D spatial pose and orientation control tasks, **DecoEdit** significantly outperforms leading closed-source models such as Seedream 5.0 in both geometric accuracy and identity consistency.

1 INTRODUCTION

Multimodal image editing has witnessed rapid progress, empowering models to synthesize visually compelling content guided by natural language instructions (Ye et al., 2026; Luo et al., 2026; Ju et al., 2026; Yin et al., 2026; Han et al., 2026; Mao et al., 2026; Huang et al., 2025; Feng et al., 2025; Yu et al., 2025; Barrett & Cheney, 2002; Hui et al., 2024). However, translating user intent into reliable edits in real-world scenarios remains fundamentally bounded by controllability. Real-world editing requirements rarely exist in flat isolation: consider an instruction such as "relocate this wooden chair to the left desk and rotate it 45° toward the camera". This instruction inherently entangles a 1D semantic relocation operation, 2D spatial destination, and a 3D volumetric novel-view rotation and constraints. While humans conceive editing effortlessly across these dimensional layers, contemporary image editors (Cao et al., 2025; Deng et al., 2025; Labs et al., 2025; Brooks et al., 2023) remain predominantly confined to "flat" 2D interactions, struggling to faithfully execute 3D-aware manipulations.

Existing literature largely tackles these requirements along fragmented trajectories. Mainstream instruction-guided and subject-driven editors operate on 2D planes, utilizing text prompts, bounding boxes, or binary masks to insert or restyle objects (Brooks et al., 2023; Zhao et al., 2024; Labs et al., 2025; Zhang et al., 2023a). When requested to perform out-of-plane geometric modifications, these models either fail to perceive spatial depth, distort physical scale, or break visual integrity. Conversely, 3D-aware editing and novel-view synthesis paradigms either rely on expensive per-instance optimization, struggle to seamlessly composite objects into complex natural scenes, or suffer from severe detail degradation (Michel et al., 2023; Chen et al., 2025; Yenphraphai et al.,

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.

Dataset	Size	Tasks	MS	2D Spatial	2D Object	3D View	3D Object	Joint 1D/2D/3D
InstructPix2Pix	313K	-	–	–	–	–	–	–
HQ-Edit	200K	-	–	–	–	–	–	–
Pico-Banana-400K	400K	35	–	–	–	–	–	–
Emu Edit	10M	16	–	–	–	–	–	–
OmniEdit	1.2M	7	✓	–	–	–	–	–
MagicBrush	10K	-	–	✓	–	–	–	–
UltraEdit	4.1M	9	✓	✓	–	–	–	–
ImgEdit	1.2M	13	✓	–	✓	–	–	–
AnyEdit	2.5M	25	✓	✓	✓	✓	–	–
DecoDataset (Ours)	**1.63M**	**42**	✓	✓	✓	✓	✓	✓

Table 1: **Comparison of image editing datasets**. Our dataset combines 42 control subtasks, 650k 3D training samples, and joint 1D/2D/3D control within a unified set. *MS*: multiple data sources. *2D Spatial*: mask, box or layout. *2D Object*: object image input for reference. *3D View*: explicit rotation/viewpoint. *3D Object*: inserting objects from images into 3D tasks.

2024; Zhang et al., 2025; Qin et al., 2026; Zhu et al., 2025). These performance bottlenecks stem from two critical factors. On the data side, there is a severe deficiency of large-scale, unified datasets that co-annotate 1D semantic instructions, 2D spatial regions, and 3D geometric poses in natural environments (Tab. 1). Existing training corpora either consist purely of flat 2D image pairs or isolated single-object 3D scans. Lacking joint multi-dimensional supervision, existing models are unable to learn generalized cross-dimensional priors, forcing many 3D-aware approaches to regress into expensive per-instance fine-tuning. On the architectural side, a critical root cause lies in the entanglement of appearance and geometry within a single 2D reference: directly rotating or warping an object in pixel space often obliterates fine-grained instance identities—such as intricate textures, embossed logos, and non-rigid structural nuances.

To overcome these limitations, our work systematically explores unified image editing across three core pillars: **Data**, **Evaluation**, and **Methodology**. First, we establish a unified data curation pipeline that seamlessly synthesizes 1D textual instructions, 2D signals, and 3D volumetric inputs derived from 3D models. This pipeline scales to over 1.6 million instances and forms **DecoDataset**. It systematically pairs standard textual edit families with fine-grained spatial and geometric conditions, covering both individual controls and their combinatorial intersections under rigorous quality verification. Second, we introduce **DecoBench**, a comprehensive evaluation suite designed to assess multi-dimensional image editing. Moving beyond flat 2D metrics, **DecoBench** comprehensively gauges semantic adherence, spatial placement, and 3D angle precision. Third, we present **DecoEdit**, an all-in-one editing framework built upon a Mixture-of-Transformers (MoT) (Deng et al., 2025; Liang et al., 2024) paradigm that natively ingests heterogeneous conditional tokens. To resolve the trade-off between structural reorientation and identity preservation, **DecoEdit** adopts an appearance-geometry decoupled pipeline: it leverages normalized structural renderings from SAM-3D Chen et al. (2026) to explicitly anchor target geometry, scale, and 3D orientation, while simultaneously routing the uncorrupted source object image to lock in high-frequency instance textures. Experiments confirm that **DecoEdit** achieves state-of-the-art performance. Crucially, on the challenging 3D orientation control task, **DecoEdit** markedly outperforms top proprietary models such as Seedream 5.0 in both rotation accuracy and identity consistency.

2 Related Work

Image Editing. Instruction-guided image editing has evolved from synthetic pair generation (Brooks et al., 2023) and human-annotated triplets (Zhang et al., 2023a) to expressive MLLM-directed editing (Fu et al., 2024) and large-scale synthetic pipelines (Zhao et al., 2024). Concurrently, subject-driven and spatially-grounded editing paradigms leverage reference images or spatial priors (e.g., bounding boxes, masks, and sketches) to insert or replace visual concepts while enforcing localization (Li et al., 2023; Zhang et al., 2023b). More recently, unified editing frameworks and generalist models have emerged to handle diverse tasks within integrated architectures (Yu et al., 2025; Liu et al., 2025; Wu et al., 2026; Zhao et al., 2026). While these models broaden semantic and multimodal capabilities, they remain fundamentally anchored in flat 2D operations.

3D-Aware Image Manipulation. To transcend planar constraints, a growing line of research explores 3D-aware image editing and novel-view manipulation. Early methods lift 2D contents into 3D proxy geometries for manual editing followed by generative refinement (Yenphraphai et al., 2024), or inject projected geometric keypoints and camera priors (Chen et al., 2024; Shi et al., 2023). Other works formulate 3D editing via procedural synthetic training (Michel et al., 2023), specialized graphic engines like Blender (Chen et al., 2025), multi-object 9-DoF coordinate maps (Qin et al., 2026), or multi-stage inpainting and composition cascades (Zhu et al., 2025; Zhang et al., 2025). Despite impressive spatial manipulations, existing 3D editing solutions often rely on expensive per-instance optimization, multi-step heuristic pipelines, or fail to composite objects seamlessly into complex backgrounds. More critically, they suffer from appearance-geometry entanglement, which degrades high-frequency instance details (e.g., textures, logos) during rotation.

Image Editing Datasets and Benchmarks. The development of controllable editing is profoundly steered by training corpora and evaluation benchmarks. Early datasets focused on paired instruction-following pairs (Brooks et al., 2023; Zhang et al., 2023a), which were later expanded to million-scale synthesized sets covering diverse textual tasks and spatial regions (Zhao et al., 2024; Sheynin et al., 2024). On the evaluation front, benchmarks such as GEdit-Bench (Liu et al., 2025) and AnyEdit (Yu et al., 2025) have systematized multi-task visual evaluation. Nevertheless, an evident gap exists in multi-dimensional supervision: current datasets either curate flat 2D modifications or isolate 3D object assets devoid of contextual environments, lacking joint annotations spanning 1D instructions, 2D boxes, and 3D poses. This data scarcity prevents models from learning cross-dimensional priors.

3 Dataset, Benchmark and Method

We first describe the large-scale data curation and rigorous filtering pipeline of **DecoDataset** (Sec. 3.1), followed by the standardized evaluation protocol of **DecoBench** (Sec. 3.2). Finally, we elaborate on the architecture and training pipeline of our unified editor, **DecoEdit** (Sec. 3.3).

3.1 DecoDataset: Scalable Multi-Dimensional Data Engine

Training a unified multi-dimensional editor requires balanced supervision across text, 2D coordinates, and 3D geometric views. To achieve this at scale, we establish an automated data curation and cascading verification pipeline (Fig. 1). Sourced from open-domain corpora (*ImgEdit* (Ye et al., 2026), *Banana-400k* (Qian et al., 2026)) and our in-house collection, **DecoDataset** scales ordinary single-turn 2D edits into a multi-control training corpus exceeding 1.6 million instances. Fig. 7 and Fig. 8 in Appendix show training set statistics and samples. Tab.5 and Tab. 6 in Appendix detail task composition and controls.

3.1.1 Data Construction Pipeline

Task Inversion and Semantic Grounding. We first identify clean single-turn edits and symmetrically invert them to double data diversity (e.g., inverting *addition* into *removal*, and augmenting bi-directional *replacements*). For each edit pair $(I_{\text{in}}, T_{\text{raw}}, I_{\text{out}})$, Gemini-3.7-Flash automatically detects the targeted entity label. Using this label, we segment the object and ground its 2D bounding box B^*, decoupling spatial localization from pixel synthesis.

Metric 3D Geometry and Pose Alignment. To extract rigid 3D target orientations without manual annotations, we design an automated *reconstruct-render-align* loop: *(1) 3D Reconstruction:* Using the extracted foreground mask, SAM-3D (Chen et al., 2026) predicts initial geometry, 3D translation, and rotation, while MoGe (Wang et al., 2025) estimates monocular metric depth and camera intrinsics K. *(2) Differentiable Pose Optimization:* To ensure exact alignment with target pixels, we enable layout post-processing: translation and scale are optimized for up to 5 steps, followed by 25 steps of joint pose-scale-translation optimization against target RGB and mask cues. If the resulting mask IoU drops, the optimization safely falls back to the initial pose. *(3) Normalized Pose Scaffold (I_g^*):* After transforming into a canonical coordinate space, the reconstructed object is rendered via original camera intrinsics. The object is then tightly cropped and scaled to occupy $\approx 85\%$ of a standardized 1024×1024 white canvas, producing a normalized, geometry-oriented pose reference with coarse appearance I_g^*.

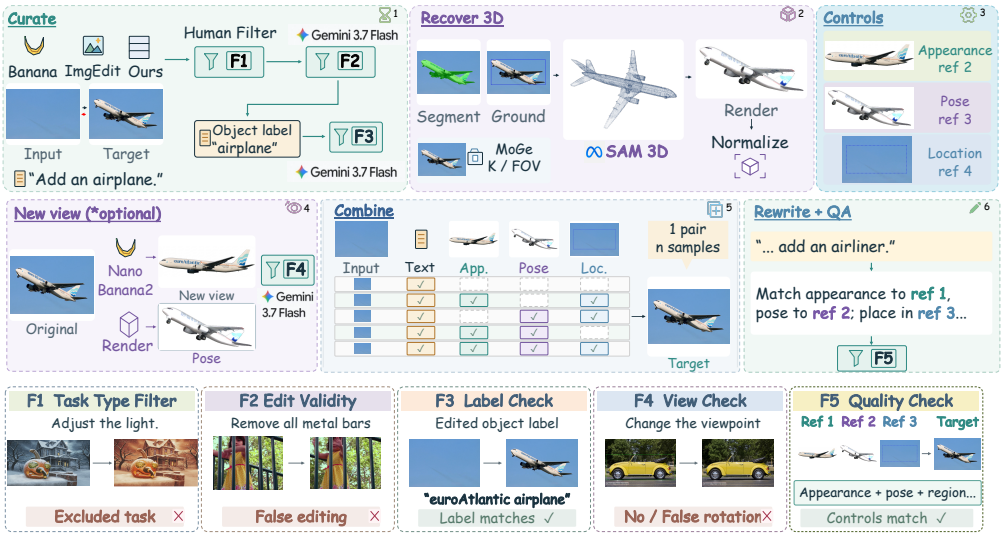

Figure 1: **Data construction pipeline.** It involves five filters to select high-quality samples.

View Augmentation for Decoupled Training. To explicitly empower the model's appearance–geometry decoupling, we synthesize novel-view variations using Gemini-3.1-Image-Preview. Specifically, we generate: (i) *in-the-wild reference edits* with adapted lighting, pose, and background, and (ii) *appearance-isolated novel views* where foreground instances rotate across arbitrary angles while strictly preserving high-frequency textures and background context.

Combinatorial Multi-Control Synthesis and Instruction Rewriting. A single visual edit pair naturally scales into an entire family of training instances ("1 pair $\rightarrow$ n samples"). By independently toggling conditions, we assemble combinations: $\{B^*\}$, $\{I_g^*\}$, $\{B^*+I_g^*\}$, with or without appearance references I_a. Finally, Gemini-3.7-Flash rewrites the raw edit texts into role-specific instructions (e.g., specifying that the target must strictly adopt the view of reference pose I_g^* while locking the texture of reference I_a into box B^*).

3.1.2 Five-Stage Cascaded Data Filtering

To ensure high visual fidelity and precise dimensional alignment, the raw candidates undergo a five-stage automated filtering cascade. **F1: Task Type Filter.** We explicitly discard ill-posed manipulation types, such as extreme artistic style transfer (e.g., anime, Van Gogh), blind outpainting, and zoom-in modifications that violate geometric consistency. **F2: Edit Validity.** Gemini-3.7-Flash filters out editing failures, multi-object entanglements, and unintentional background modifications to enforce clean, single-instance supervision. **F3: Label Check.** We verify that the extracted object label strictly matches both the visual pixels and textual instructions, eliminating tracking drift. **F4: View Check.** For augmented novel-view samples, Gemini-3.7-Flash discards cases where rotations failed, background drifted, or physical unnaturalness occurred. **F5: Quality Check.** After full combinatorial assembly and instruction rewriting, all samples undergo final verification evaluated by Gemini-3.7-Flash across 9 rigorous criteria (covering image validities, spatial box compliance, orientation correctness, reference detail fidelity, and unintended human modifications; see Appendix for full filtering, view construction, and evaluation prompts). Only samples achieving a perfect score (5/5) across all nine dimensions are retained, guaranteeing a production-grade training corpus.

3.2 DecoBench: Multi-Dimensional Evaluation Suite

To rigorously assess multi-dimensional editing, we construct **DecoBench**, focusing on real-world user simulation.

Cross-Task Sampling. For each 3D-related manipulation task, we construct several representative cases. To enable fair and fine-grained cross-task comparisons, we prioritize instances sharing the identical object ID across distinct task branches, extracting all multi-task occurrences of that identity. When shared IDs are exhausted, the remaining quota is filled via uniform random sampling.

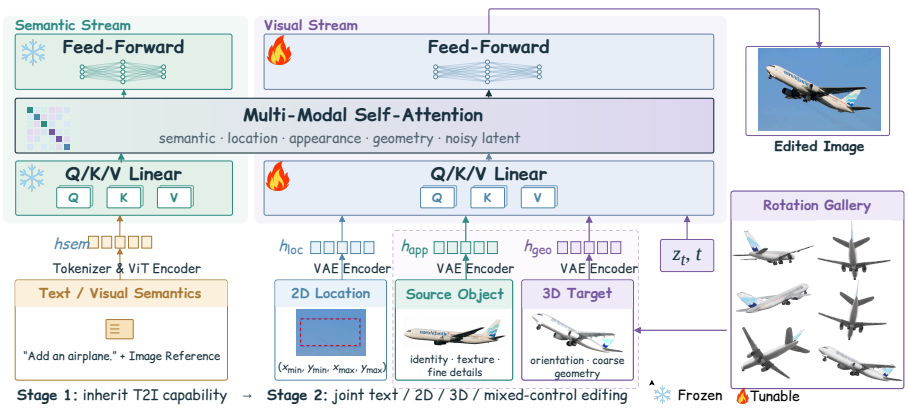

Figure 2: **The MoT architecture of DecoEdit**. It comprises a semantic stream and a visual stream.

Strict Zero-Leakage. To ensure strict evaluation integrity, we enforce complete data decontamination: once an instance ID is chosen for **DecoBench**, all corresponding instances in the training set are thoroughly purged—regardless of whether that identity appears in the source canvas, target output, or conditional controls (B^*, I_a, I_g^*). This guarantees zero test-set memorization or data leakage.

Expert Curation. The initial candidates are **manually constructed** by human experts to ensure unambiguous instructions and balanced visual diversity. To simulate realistic open-domain user interactions, we augment the benchmark along two lines: (1) We directly manipulate object 3D meshes to render custom, unconstrained viewing angles, closely reflecting arbitrary interactive camera controls from real users, resulting in 256 testing cases. (2) We intentionally include 5 physically violating edge cases (e.g., an object defying gravity or unusual structural orientations) to probe the model's controllability boundaries and constraint adherence under adversarial prompts. This results in a clean, decontaminated benchmark containing 261 testing cases.

3.3 DecoEdit: Architecture and Decoupled Methodology

3.3.1 Overall Framework

DecoEdit unifies heterogeneous conditions into a dual-stream backbone consisting of a Semantic Stream and a Visual Stream, interconnected via a shared Multi-Modal Self-Attention module as illustrated in Fig. 2.

Input Tokenization. The model ingests four role-identifiable conditional inputs: **Semantic Tokens (h_{sem}):** Text instructions and semantic image references T are mapped into semantic embeddings via a Tokenizer and ViT encoder: $h_{\text{sem}} = \text{Encoder}_{\text{sem}}(T)$, respectively. **2D Spatial Tokens (h_{loc}):** Target coordinates $B^* = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ are converted into visual mask layouts and encoded by the VAE encoder $\mathcal{E}_{\text{vae}}$ into spatial tokens h_{loc}. **Source Appearance Tokens (h_{app}):** The cropped source instance I_a is encoded as $h_{\text{app}} = \mathcal{E}_{\text{vae}}(I_a)$, serving to lock in identity, logos, and high-frequency textures. **3D Target Pose Tokens (h_{geo}):** The SAM-3D geometric rendering I_g^* is encoded into latent geometry tokens $h_{\text{geo}} = \mathcal{E}_{\text{vae}}(I_g^*)$. **Noisy Latent ($z_t$):** The diffusion target latent at timestep t is denoted as z_t.

MoT Structure. The semantic tokens h_{sem} are routed into dedicated Semantic Q/K/V projections and feed-forward networks (FFN), preserving high-level semantic intent. In parallel, the spatial and conditional visual tokens $H_{\text{vis}} = [h_{\text{loc}} \,\|\, h_{\text{app}} \,\|\, h_{\text{geo}} \,\|\, z_t]$ are projected via a shared Visual Q/K/V linear layer. To enable dynamic cross-dimensional reasoning, the token streams are concatenated along the sequence dimension and interact through Multi-Modal Self-Attention module for information interaction.

3.3.2 Appearance-Geometry Decoupled Strategy

Limitations of Coupled 3D Conditioning. An intuitive approach for 3D-aware image editing is to condition the diffusion model directly on a 3D structural rendering to simultaneously dictate the target orientation and instance details, akin to the paradigm adopted by ControlNet (Zhang et al., 2023b) and its 3D variants. However, our empirical investigations reveal that existing open-domain

3D reconstruction proxies—including state-of-the-art formulations such as SAM-3D (Chen et al., 2026)—are fundamentally bounded in representational resolution: while they provide coarse volumetric geometry and rigid camera poses, they inevitably discard high-frequency instance textures, fine edge boundaries, and micro-structures (e.g., logos, complex patterns, and materials). Consequently, models following this coupled strategy can satisfy the requested rotation angle, but invariably suffer from severe texture blurring and catastrophic detail degradation.

To resolve the classical trade-off where rotating an object destroys its fine-grained textures, **DecoEdit** establishes an appearance-geometry decoupled editing pipeline: (1) Orientation Guidance via 3D Proxy: rather than forcing the network to infer unseen poses from 2D pixels, we query a pre-rendered *Rotation Gallery* generated via SAM-3D to obtain I_g^*. This rendering acts as a pure orientation image, supplying rigid 3D viewpoint and elevation guidance. (2) Detail Preservation via Original Instance Reference: concurrently, the uncorrupted source reference I_a is routed strictly to anchor high-frequency details. Because I_a is relieved from the burden of dictating orientation, the model transfers intricate textures, non-rigid patterns, and text markings from I_a directly onto the geometric framework provided by I_g^*.

To achieve balanced multi-task mastery without catastrophic interference, **DecoEdit** is trained in two distinct stages. First, we train our model on broad text-to-image pairs to align the cross-stream projection spaces and inherit open-domain visual priors. In the second stage, we train **DecoEdit** on our curated **DecoDataset** for unified editing.

4 Experiments

4.1 Experimental Setup

Implementation Details. Our method instantiates the dual-stream design as a MoT backbone with two parameter-disjoint streams of 7B parameters each a semantic stream and a visual DiT stream for a total of 14B parameters. We initialize our model from a pretrained 7B Qwen multimodal backbone (Bai et al., 2025) and fine-tune its visual generation branch, while keeping the language branch, vision encoder, and VAE frozen. Training uses AdamW with a learning rate of 2.5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.02.

Evaluation Metrics. Instruction-driven editing is intrinsically hard to measure with conventional low-level metrics such as PSNR as documented explicitly in ImgEdit (Ye et al., 2026) and motivates the judge-based protocols of recent editing benchmarks (Liu et al., 2025; Yu et al., 2025; Luo et al., 2026). We therefore evaluate with an MLLM judge (Gemini-3.7-Flash) under a rubric of seven criteria, and report the seven per-aspect scores numerically. This principle conditions the seven scored dimensions, each graded on a 0–10 scale: **IA** (Instruction Adherence), **SCA** (Spatial Control Accuracy), **PVA** (Pose and View Alignment), **IAF** (Identity and Appearance Fidelity), **NCP** (Non-target Content Preservation), **IEQ** (Image-Editing Quality), and **DCP** (Detail Completion Plausibility). In parallel, we conduct human evaluation: four annotators, blind to model identity, rank the outputs of five sampled models per case with ties allowed.

Competitors. We compare against several open-source methods covering general generation/editing and unified multimodal families: HiDream-O1-Image(-Dev) (Cai et al., 2026), BAGEL (Deng et al., 2025), OmniGen (Xiao et al., 2025), OmniGen2 (Wu et al., 2026), FireRed-Image-Edit-1.0/1.1 (Team et al., 2026), DreamOmni2 (Xia et al., 2025), HunyuanImage 3.0 (Cao et al., 2025), Qwen-Image-Edit-2511 (Zhao et al., 2026), Qwen-Image-2.1 (Zhao et al., 2026), FLUX.2 [klein] 4B/9B (Labs et al., 2025), FLUX.2 [dev] and SenseNova-U1.5-8B-MoT (Diao et al., 2026). The closed-source group covers commercial APIs: Gemini 2.5 Flash Image, Gemini 3.1 Flash Image, Wan 2.7 Image (Mao et al., 2026), Seedream 5.0, FLUX.2 [max], and GPT-Image-1.5.

4.2 Main Results

Quantitative Comparison. Tab. 2 summarizes the comparison. **DecoEdit** attains a macro score of 8.97, the impressive result among all comparable models, surpassing the strongest open-source baseline SenseNova-U1.5-8B-MoT (8.01) by $+0.96$ and outperforming every leading closed-source editor: FLUX.2 [max] (8.84), GPT-Image-1.5 (8.71), Seedream 5.0 (8.58), Gemini 3.1 Flash Im-

Model	IA	SCA	PVA	IAF	NCP	IEQ	DCP	Macro	DiT Param.
Open-source models									
HiDream-O1-Image-Dev	1.17	1.57	2.21	2.14	1.17	2.04	0.27	2.34	8B
HiDream-O1-Image	2.08	3.15	3.51	2.88	3.11	2.85	1.08	3.52	8B
BAGEL	2.54	3.35	4.24	3.20	3.10	2.61	1.69	3.55	7B
OmniGen	0.53	0.78	0.85	0.71	0.97	1.48	0.29	1.26	3.8B
OmniGen2	2.75	3.46	4.70	4.53	5.06	5.01	2.47	4.82	4B
FireRed-Image-Edit-1.0	3.28	4.20	5.17	5.32	4.78	5.55	4.13	4.68	20B
FireRed-Image-Edit-1.1	4.41	5.51	5.91	6.02	7.31	6.86	4.92	5.72	20B
DreamOmni2	4.72	6.07	5.87	5.91	7.63	7.26	5.37	6.25	12B
HunyuanImage 3.0 (API)	5.68	6.31	7.25	6.95	7.19	6.66	6.10	6.92	80A13B
Qwen-Image-Edit-2511	3.81	5.29	5.41	4.68	8.11	7.52	3.35	5.99	20B
Qwen-Image-2.1	6.15	6.79	7.30	7.07	8.05	6.98	5.45	6.94	7B
FLUX.2 [klein] 4B	5.05	5.94	6.29	6.02	9.21	8.46	4.98	6.81	4B
FLUX.2 [dev]	5.26	6.25	7.02	6.58	9.12	7.54	5.69	7.02	32B
FLUX.2 [klein] 9B	6.74	7.36	7.56	7.66	9.27	8.35	6.47	7.55	9B
SenseNova-U1.5-8B-MoT	7.75	8.16	7.92	8.30	8.36	7.65	7.75	8.01	8B
Closed-source models									
Gemini 2.5 Flash Image	6.19	7.02	7.10	7.28	9.47	8.36	6.71	7.29	–
Gemini 3.1 Flash Image	8.60	8.78	8.68	8.73	*9.44*	8.22	8.58	8.58	–
Wan 2.7 Image	8.45	8.79	8.59	8.68	9.42	8.36	8.50	8.53	–
Seedream 5.0	*8.69*	*8.86*	8.75	8.99	*9.44*	8.67	8.77	8.58	–
FLUX.2 [max]	8.87	8.87	9.02	*8.96*	**9.52**	**9.15**	*8.78*	8.84	–
GPT-Image-1.5	8.52	8.46	*8.87*	**9.14**	8.75	9.06	**9.18**	*8.71*	–
Ours									
Baseline	5.74	6.59	6.42	6.07	9.05	7.76	5.41	6.98	7B
DecoEdit	**8.98** +3.24	**9.07** +2.48	**9.05** +2.63	8.70 +2.63	9.39 +0.34	*8.92* +1.16	8.90 +3.50	**8.97** +1.99	7B

Table 2: **Comparison on DecoBench**. Refer to Sec. 4.1 for evaluation metric abbreviations. Baseline denotes the initialization checkpoint before fine-tuning. The w/o Decoupling and w/o Check variants are separately trained ablations.

Rater	**Ours**	Seedream 5.0	HY-Image 3.0	Gemini 2.5 Flash	Qwen-Image-Edit	Spearman $\rho \uparrow$	p-value	Kendall $\tau_b \uparrow$
Rater 1	**72.50**	62.75	44.50	*44.75*	25.50	0.644	< 0.001	0.547
Rater 2	61.25	**71.25**	*48.25*	41.25	28.00	0.451	< 0.001	0.383
Rater 3	**74.25**	66.00	40.00	*42.25*	27.50	0.676	< 0.001	0.592
Rater 4	**73.00**	65.25	*44.50*	41.00	26.25	0.732	< 0.001	0.655
Average	**70.25**	66.31	*44.31*	42.31	26.81	0.733	< 0.001	0.648

Table 3: Human evaluation and agreement with the VLM judge on 3D editing tasks. Model columns report normalized human ranking scores (0–100; higher is better). Correlation columns report mean within-case correlations with Gemini-3.7-Flash

age (8.58) and Wan 2.7 Image (8.53), while using far fewer parameters than the large open-source competitor (HunyuanImage 3.0, 80B total / 13B active).

Human Agreement with The Judge. We verify that the judge is a faithful proxy for human preference. Four raters independently rank five models (ours, Seedream 5.0, HunyuanImage 3.0, Gemini 2.5 Flash Image, Qwen-Image-Edit-2511) on 50 randomly sampled **DecoBench** cases, allowing ties; ranks are normalized to a 0–100 score (Tab. 3). Human consensus orders the models as $70.25 > 66.31 > 44.31 > 42.31 > 26.81$, i.e., ours first and Seedream 5.0 close behind. Agreement between the two is strong and statistically significant: per-rater Spearman ρ with Gemini-3.7-Flash ranges from 0.451 to 0.732 (all adjusted $p < 0.001$), and the per-case mean-rank consensus reaches $\rho = 0.733$ with Kendall $\tau_b = 0.648$, using 100,000 within-case two-sided permutations and Holm correction across raters. Fig. 9 and Fig. 10 in Appendix provide further visualizations and human evaluation data.

Qualitative Analysis. Fig. 3 and Fig. 4 illustrate the failure modes that **DecoEdit** removes. Existing models and baselines conditioned on a single coupled reference exhibit one of two behaviors: they copy the flat-shaded geometry of the pose scaffold, smearing away textures, embossed logos and printed text, or they honor the appearance reference and drift back toward the original viewpoint, ignoring the requested rotation. **DecoEdit** adopts the target orientation while transferring high-frequency detail from the source instance, and it respects the target box in **DecoBench** cases where

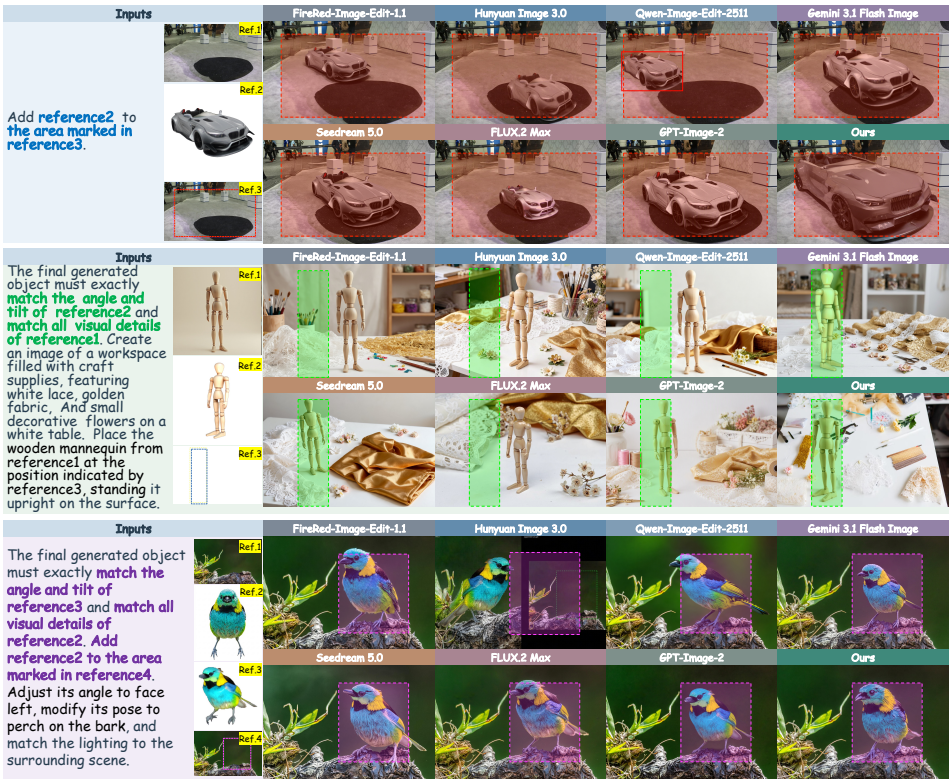

Figure 3: **Qualitative comparison with existing state-of-the-art method** on 3D-aware tasks.



Figure 4: **Qualitative ablation study.** Our method demonstrates excellent 3D control performance.

several baselines place the object outside the requested region. Results in Tab. 4 further indicate the effectiveness of the appearance-geometry decoupling strategy.

4.3 Ablation Study

Necessity of The Fifth-stage Data Check. As shown in Tab. 4, we retrain the full model on the corpus produced *before* the F5 nine-rule screening (*w/o Check*), keeping the training steps matched. Adding the F5 improves every task category without exception: object addition $8.85 \rightarrow 9.18$, appearance editing $8.54 \rightarrow 8.99$, pose/view editing $7.86 \rightarrow 8.57$, object replacement $8.44 \rightarrow 8.84$, and conditional scene generation $8.17 \rightarrow 9.27$, costing 0.60 macro points.

Data Scaling. We additionally train on 10%, 20% and 50% of **DecoDataset** under the same schedule and observe a monotone improvement in editing quality as the corpus grows (Fig. 5), supporting the scaling potential of the data engine.

Joint 1D/2D/3D training. Fig. 6 ablates the training mixture along the control dimensions: removing 1D instructions (2D+3D), removing 2D spatial control (1D+ 3D), and reducing the 3D stream to one third of its volume (1D+2D+$\frac{1}{3}$3D). All three fall short of the jointly trained model on the

Task	Configuration	IA↑	SCA↑	PVA↑	IAF↑	NCP↑	IEQ↑	DCP↑	Avg.↑
Object addition	Baseline	5.73 ↓3.52	6.00 ↓3.07	6.35 ↓2.81	6.22 ↓2.59	8.91 ↓0.64	6.77 ↓2.52	5.08 ↓3.99	6.60 ↓2.58
	w/o Decoupling	8.53 ↓0.72	8.66 ↓0.41	8.47 ↓0.69	8.16 ↓0.65	9.32 ↓0.23	8.87 ↓0.42	8.57 ↓0.50	8.65 ↓0.53
	w/o Check	8.80 ↓0.45	8.98 ↓0.09	8.86 ↓0.30	8.44 ↓0.37	9.23 ↓0.32	8.84 ↓0.45	8.88 ↓0.19	8.85 ↓0.33
	Ours	**9.25**	**9.07**	**9.16**	**8.81**	**9.55**	**9.29**	**9.07**	**9.18**
Appearance editing	Baseline	5.58 ↓3.12	8.76 ↓0.76	8.70 ↓0.75	5.39 ↓3.03	**9.61**	**8.61**	–	7.77 ↓1.22
	w/o Decoupling	8.55 ↓0.15	9.45 ↓0.07	**9.48**	8.18 ↓0.24	9.30 ↓0.22	8.21 ↓0.12	–	8.86 ↓0.13
	w/o Check	8.18 ↓0.52	9.00 ↓0.52	9.15 ↓0.30	8.03 ↓0.39	9.21 ↓0.31	7.70 ↓0.63	–	8.54 ↓0.45
	Ours	**8.70**	**9.52**	9.45	**8.42**	9.52	8.33	–	**8.99**
Pose/view editing	Baseline	4.50 ↓3.94	7.44 ↓1.12	4.50 ↓3.94	6.69 ↓1.69	8.56 ↓0.69	7.44 ↓1.00	5.64 ↓3.00	6.43 ↓2.14
	w/o Decoupling	3.50 ↓4.94	6.69 ↓1.87	3.44 ↓5.00	5.88 ↓2.50	8.38 ↓0.87	6.31 ↓2.13	4.07 ↓4.57	5.50 ↓3.07
	w/o Check	7.38 ↓1.06	7.88 ↓0.68	7.56 ↓0.88	8.00 ↓0.38	8.25 ↓1.00	7.94 ↓0.50	8.07 ↓0.57	7.86 ↓0.71
	Ours	**8.44**	**8.56**	**8.44**	**8.38**	**9.25**	**8.44**	**8.64**	**8.57**
Object replacement	Baseline	5.80 ↓2.98	6.38 ↓2.59	6.00 ↓2.84	5.51 ↓3.12	9.08 ↓0.12	8.31 ↓0.41	5.58 ↓3.21	6.75 ↓2.09
	w/o Decoupling	**8.84**	8.87 ↓0.10	8.78 ↓0.06	8.54 ↓0.09	**9.30**	**8.87**	**8.81**	**8.85**
	w/o Check	8.40 ↓0.38	8.64 ↓0.33	8.56 ↓0.28	8.03 ↓0.60	8.92 ↓0.28	8.14 ↓0.58	8.51 ↓0.28	8.44 ↓0.40
	Ours	8.78	**8.97**	**8.84**	**8.63**	9.20	8.72	8.79	8.84
Conditional scene generation	Baseline	6.80 ↓2.70	6.10 ↓3.05	6.45 ↓2.75	8.45 ↓0.65	–	8.90 ↓0.50	–	7.34 ↓1.93
	w/o Decoupling	8.50 ↓1.00	8.90 ↓0.25	7.75 ↓1.45	8.95 ↓0.15	–	8.95 ↓0.45	–	8.61 ↓0.66
	w/o Check	8.05 ↓1.45	8.30 ↓0.85	7.80 ↓1.40	7.95 ↓1.15	–	8.75 ↓0.65	–	8.17 ↓1.10
	Ours	**9.50**	**9.15**	**9.20**	**9.10**	–	**9.40**	–	**9.27**

Table 4: **Per-task results on DecoBench**. Baseline denotes the initialization checkpoint before fine-tuning. The w/o Decoupling and w/o Check variants are separately trained ablations.



Figure 5: **Ablation results** on the impact of varying data scales on image editing performance.

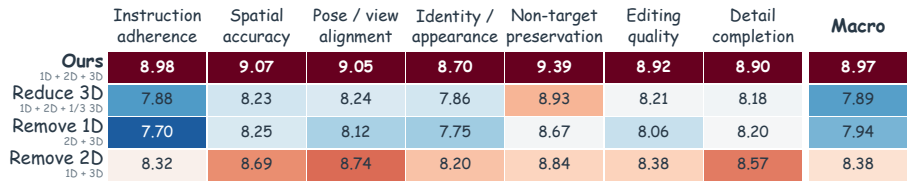

Figure 6: **Quantitative ablation study** on the effectiveness of joint 1D/2D/3D training paradigm.

macro score (8.97 vs. 7.94, 8.38, and 7.89). The pattern is informative rather than uniform: dropping 1D supervision hurts instruction adherence and non-target preservation most, and dropping 2D supervision degrades spatial compliance. Removing any single dimension therefore damages a different capability, which validates the unified-corpus design over per-task specialists.

5 Conclusion

In this work, we explored unified, multi-dimensional image editing across three foundational pillars: Data, Evaluation, and Methodology. First, we established a dedicated data curation pipeline and constructed **DecoDataset**, a large 1.6 million dataset bridging 1D textual, 2D spatial, and 3D geometric controls. Furthermore, we established **DecoBench** to systematically evaluate multi-factor editing capabilities. Moreover, we introduced **DecoEdit**, an MoT-based all-in-one editor featuring an appearance-geometry decoupled pipeline that leverages SAM-3D for orientation guidance while preserving high-frequency instance details from source references. Extensive evaluations show that **DecoEdit** sets a new state-of-the-art, even significantly outperforming top proprietary models such as Seedream 5.0 on challenging 3D orientation tasks.

AI USE STATEMENT

In this work, we used generative AI tools for synthetic-data generation, data annotation and filtering, instruction rewriting and automated evaluation. Specifically, Gemini-3.7-Flash was used to extract and verify target-object labels, rewrite editing instructions, assess sample quality and condition consistency, and evaluate generated editing outputs. Gemini-3.1-Image-Preview was used to synthesize novel-view image variants within the data-construction pipeline. We also used ChatGPT to check the wording. These related uses supported the presentation of the reported experiments rather than providing additional experimental measurements.

AI-assisted data construction was checked through the filtering procedures described in Section 3.1, and benchmark cases were manually curated as described in Section 3.2. We further assessed automated evaluation by comparing model-generated scores with independent rankings from four human annotators, as reported in Section 4.2. Human rankings are reported separately from automated scores. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies controllable image editing through DecoDataset, DecoBench, and DecoEdit. We do not endorse applications involving impersonation, harassment, misinformation, or violations of privacy and intellectual-property rights. Our data-filtering procedures include checks for unintended changes to human-related content and non-target regions; these checks support data quality. Any reuse or distribution of source images, derived data, or model artifacts should respect the corresponding licenses, usage terms, and privacy constraints, and edited images should not be represented as unmodified records of real events.

REPRODUCIBILITY STATEMENT

The main paper and appendix document the procedures underlying our results. Section 3.1 describes data construction and filtering, while Section 3.2 explains benchmark construction and the identity-based train–test separation procedure. Sections 3.4–3.5 describe the model architecture and appearance–geometry decoupling strategy, and Section 4.1 reports model initialization, trainable components, optimizer settings, and evaluation setup. Appendix A.1 provides training-data statistics and the complete task and control taxonomy. Appendices A.2 and A.5 detail benchmark composition, metric applicability, score aggregation, and human-evaluation procedures. Appendix A.4 documents the additional ablation configurations, including checkpoint differences relevant to interpreting the comparisons. Appendix A.6 reports construction-stage sample counts and API-cost estimation assumptions. Finally, Appendix A.7 provides the prompts used for data filtering, novel-view generation, and automated evaluation, including input-role definitions, scoring criteria, and output formats.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

William A Barrett and Alan S Cheney. Object-based image editing. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 777–784, 2002.

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.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv e-prints*, pp. arXiv–2509, 2025.

Jiacheng Chen, Ramin Mehran, Xuhui Jia, Saining Xie, and Sanghyun Woo. Blenderfusion: 3d-grounded visual editing and generative compositing. *arXiv preprint arXiv:2506.17450*, 2025.

Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. In *International Conference on Learning Representations*, volume 2024, pp. 8979–9001, 2024.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026.

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.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. Sensenova-u1. 5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026.

Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2969–2977, 2025.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54820–54833, 2024.

Feng Han, Yibin Wang, Chenglin Li, Zheming Liang, Dianyi Wang, Yang Jiao, Zhipeng Wei, Chao Gong, Cheng Jin, and Jiaqi Wang. Unireditbench: A unified reasoning-based image editing benchmark. In *European Conference on Computer Vision*, pp. 287–304. Springer, 2026.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4409–4437, 2025.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Cherry Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. Editverse: Unifying image and video editing and generation with in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 137234–137255, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

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 *International Conference on Learning Representations*, volume 2026, pp. 33027–33056, 2026.

Chaojie Mao, Chen-Wei Xie, Chongyang Zhong, Haoyou Deng, Jiaxing Zhao, Jie Xiao, Jinbo Xing, Jingfeng Zhang, Jingren Zhou, Jingyi Zhang, et al. Wan-image: Pushing the boundaries of generative visual intelligence. *arXiv preprint arXiv:2604.19858*, 2026.

Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36:3497–3516, 2023.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Zhenyuan Qin, Xincheng Shuai, and Henghui Ding. Scenedesigner: Controllable multi-object image generation with 9-dof pose manipulation. *Advances in Neural Information Processing Systems*, 38:133376–133400, 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 *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 8871–8879. IEEE, 2024.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271. IEEE, 2025.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. Image sculpting: Precise object editing with 3d geometry control. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4241–4251. IEEE, 2024.

Fukun Yin, Shiyu Liu, Yucheng Han, Zhibo Wang, Peng Xing, Rui Wang, Wei Cheng, Yingming Wang, Aojie Li, Zixin Yin, et al. Reasonedit: Towards reasoning-enhanced image editing models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23283–23293, 2026.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135. IEEE, 2025.

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.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023b.

Yuhong Zhang, Han Wang, Yiwen Wang, Rong Xie, and Li Song. Freeinsert: Personalized object insertion with geometric and style control. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10361–10369, 2025.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

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.

Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19130–19140. IEEE, 2025.

A Appendix

This appendix provides additional dataset statistics, the complete task and control taxonomy, supplementary benchmark results, ablation analyses, human-evaluation details, data-construction and API-cost accounting, and the prompts used for data filtering and evaluation.

A.1 Dataset Statistics and Control Taxonomy

Training-data composition. Figure 7 and Table 5 summarize **DecoDataset** by task family and control dimension. The dataset contains 1,627,974 training samples: 415,299 in 1D, 562,209 in 2D, and 650,466 in 3D. These samples span ten task families and 42 control subtasks, comprising 9, 18, and 15 subtasks in the three dimensions, respectively. Object insertion and replacement are the largest families, with 762,833 and 413,799 samples. The remaining families cover removal, attribute editing, resizing, relocation, pose/view adjustment, background editing, composite editing, and object-conditioned scene generation. The counts describe constructed training samples rather than unique source images, since one editing pair can yield multiple control-conditioned variants.

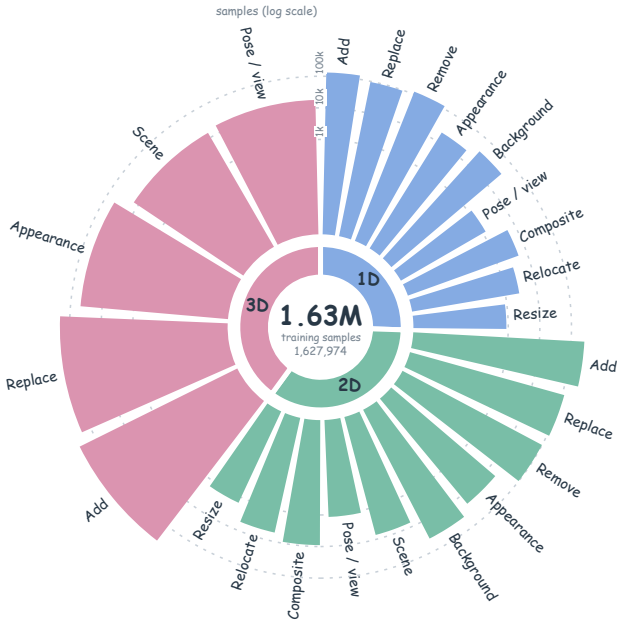

Figure 7: Training-data distribution across task families and control dimensions.

Notation and subtask counting. In Table 5, S denotes the source scene; A and P denote appearance and pose references; R denotes an appearance, pose, or joint appearance–pose reference; B denotes spatial control; and L denotes the target layout. B_{src} and B_{tgt} denote source and target controls. Each symbol represents one input image, excluding the target output. Each numeric cell reports the sample count above the subtask count in parentheses; a dash denotes no training samples and zero subtasks. A subtask is an observed combination of dimension, task family, and control configuration. Counts distinguish reference roles and spatial-control roles, but do not treat data sources or forward/inverse directions as additional subtasks. The displayed configurations pool roles and dimensions within each task and are not a one-to-one enumeration of subtasks; identical formulas use the same text color.

Input roles and dimensional categories. All configurations receive a text instruction. A 1D configuration uses a source image without additional reference or spatial-control images. A 2D configuration introduces spatial controls and/or an appearance reference without a dedicated target-view control image. A 3D configuration includes a target-view reference, which can jointly specify appearance or be supplied separately from an appearance reference. Consequently, a textual rotation

Task family	Input configurations	1D	2D	3D	Total
		Samples; (control subtasks)			
Insertion	S; S + R; S + B; S + A + P; S + R + B; S + A + P + B	132,163 (1)	263,193 (3)	367,477 (5)	762,833 (9)
Removal	S; S + B	120,716 (1)	98,756 (1)	– (0)	219,472 (2)
Replacement	S; S + R; S + B; S + A + P; S + R + B; S + A + P + B	95,033 (1)	116,473 (3)	202,293 (4)	413,799 (8)
Attribute editing	S; S + R; S + B; S + R + B	20,713 (1)	21,051 (1)	48,184 (2)	89,948 (4)
Resizing	S; S + B; S + B_{src} + B_{tgt}	872 (1)	1,551 (2)	– (0)	2,423 (3)
Relocation	S; S + B; S + B_{src} + B_{tgt}	2,748 (1)	5,229 (2)	– (0)	7,977 (3)
Pose / view adjustment	S; S + R; S + B; S + R + B	1,248 (1)	1,171 (1)	17,849 (2)	20,268 (4)
Background editing	S; S + B	36,516 (1)	38,534 (1)	– (0)	75,050 (2)
Composite editing	S; S + B; S + B_{src} + B_{tgt}	5,290 (1)	9,001 (3)	– (0)	14,291 (4)
Object-conditioned scene generation	R + L; A + P + L	– (0)	7,250 (1)	14,663 (2)	21,913 (3)
Total		**415,299 (9)**	**562,209 (18)**	**650,466 (15)**	**1,627,974 (42)**

Table 5: Training samples and control subtasks by task family and dimension.

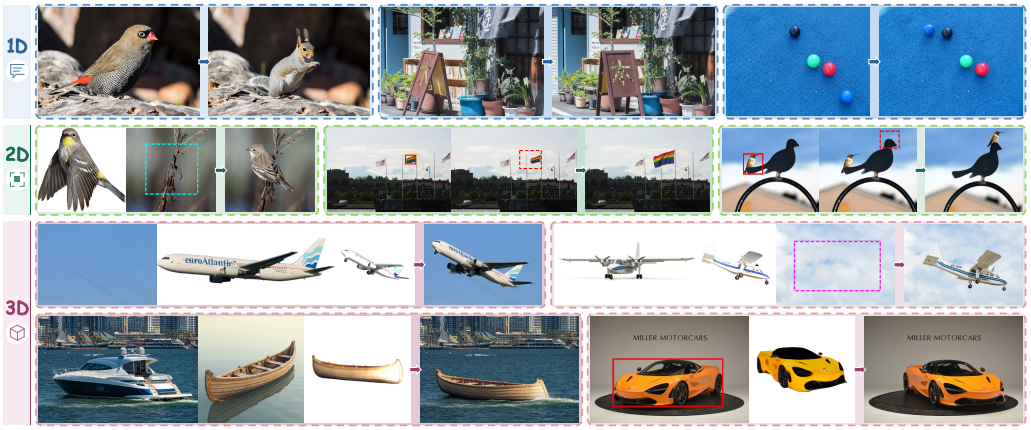

Figure 8: Representative editing samples with 1D, 2D, and 3D controls.

request alone does not make a sample a 3D configuration. In the decoupled configuration, appearance evidence specifies identity and visual details, the pose reference specifies the target view, and spatial controls specify the relevant object or region. Figure 8 illustrates representative samples across these configurations.

Detailed and pooled task counts. Table 6 enumerates the 42 detailed variants, which cover 24 dimension–task categories and map to 35 pooled entries within task families. There are nine distinct pooled input configurations across all families. Pooling merges appearance, pose, and joint references into R, and spatial-control roles into B; it does not make their detailed supervision roles equivalent. For example, a joint appearance–pose reference is one image, whereas separate appearance and pose references are two images. Spatial inputs distinguish object selection, preservation, edit regions, destination position or extent, and blank-canvas layouts. Their roles are determined by the instruction and annotations rather than line style alone. Input-image counts exclude the target image. Repeated C-identifiers indicate the same pooled entry within a task family. Physics-violation cases are benchmark-only and are not included in the 42 training variants.

Reference and spatial-control roles. In Table 6, J denotes a joint appearance–pose reference, whereas $A + P$ denotes two separate reference images. The detailed spatial roles are B_{s} for object selection, B_{k} for preservation, B_{e} for an edit region, and B_{t} for a target position or size; L specifies a target layout on a blank canvas. Box-annotated scene copies count as separate inputs. Pooling merges A, P, and J into R, and spatial roles into B, while $B_{\mathrm{src}} + B_{\mathrm{tgt}}$ retains the two-image selection/preservation and destination pair.

ID	Task family	Dim.	Images	Detailed input controls	Pooled controls
T01	Object addition	1D	1	Scene	C01 S
T02		2D	2	Scene + appearance ref.	C02 $S+R$
T03		2D	2	Scene + dashed target box	C03 $S+B$
T04		2D	3	Scene + appearance ref. + dashed target box	C04 $S+R+B$
T05		3D	2	Scene + joint appearance/pose ref.	C02 $S+R$
T06		3D	3	Scene + appearance ref. + pose ref.	C05 $S+A+P$
T07		3D	3	Scene + joint appearance/pose ref. + solid edit region	C04 $S+R+B$
T08		3D	3	Scene + joint appearance/pose ref. + dashed target box	C04 $S+R+B$
T09		3D	4	Scene + appearance ref. + pose ref. + dashed target box	C06 $S+A+P+B$
T10	Object removal	1D	1	Scene	C07 S
T11		2D	2	Scene + solid selection	C08 $S+B$
T12	Object replacement	1D	1	Scene	C09 S
T13		2D	2	Scene + appearance ref.	C10 $S+R$
T14		2D	2	Scene + solid selection	C11 $S+B$
T15		2D	3	Scene + appearance ref. + solid selection	C12 $S+R+B$
T16		3D	2	Scene + joint appearance/pose ref.	C10 $S+R$
T17		3D	3	Scene + appearance ref. + pose ref.	C13 $S+A+P$
T18		3D	3	Scene + joint appearance/pose ref. + solid selection	C12 $S+R+B$
T19		3D	4	Scene + appearance ref. + pose ref. + solid selection	C14 $S+A+P+B$
T20	Appearance editing	1D	1	Scene	C15 S
T21		2D	2	Scene + solid selection	C16 $S+B$
T22		3D	2	Scene + joint appearance/pose ref.	C17 $S+R$
T23		3D	3	Scene + joint appearance/pose ref. + solid selection	C18 $S+R+B$
T24	Object resizing	1D	1	Scene	C19 S
T25		2D	2	Scene + solid selection	C20 $S+B$
T26		2D	3	Scene + solid selection + dashed target box	C21 $S+B_{\mathrm{src}}+B_{\mathrm{tgt}}$
T27	Object relocation	1D	1	Scene	C22 S
T28		2D	2	Scene + solid selection	C23 $S+B$
T29		2D	3	Scene + solid selection + dashed target box	C24 $S+B_{\mathrm{src}}+B_{\mathrm{tgt}}$
T30	Pose / view editing	1D	1	Scene	C25 S
T31		2D	2	Scene + solid selection	C26 $S+B$
T32		3D	2	Scene + pose ref.	C27 $S+R$
T33		3D	3	Scene + solid selection + pose ref.	C28 $S+R+B$
T34	Background editing	1D	1	Scene	C29 S
T35		2D	2	Scene + solid preservation	C30 $S+B$
T36	Composite editing	1D	1	Scene	C31 S
T37		2D	2	Scene + solid preservation	C32 $S+B$
T38		2D	2	Scene + dashed target box	C32 $S+B$
T39		2D	3	Scene + solid preservation + dashed target box	C33 $S+B_{\mathrm{src}}+B_{\mathrm{tgt}}$

Continued on the next page.

Table 6 (continued)

ID	Task family	Dim.	Images	Detailed input controls	Pooled controls
T40	Object-conditioned scene generation	2D	2	appearance ref. + blank target layout	C34 $R+L$
T41		3D	2	joint appearance/pose ref. + blank target layout	C34 $R+L$
T42		3D	3	appearance ref. + pose ref. + blank target layout	C35 $A+P+L$
Physics violation		Special		No training samples.	Outside the 42 variants.

Table 6: Detailed and pooled controls for all 42 training variants.

A.2 Benchmark Composition and Evaluation Details

Evaluation categories. The standard 3D evaluation contains 256 cases: 97 object-addition, 33 appearance-editing, 16 pose/view-editing, 90 object-replacement, and 20 conditional-scene-generation cases, as listed in Table 9. Five additional physics-violation cases form a separate diagnostic group, bringing the benchmark to 261 cases. Tables 7, 8, and 9 exclude this special group from their standard task-level results. The benchmark construction follows the identity-based isolation procedure described in the main text, removing related training instances when the selected identity appears in source, target, or conditional inputs.

Diagnostic scores and aggregation. The evaluation reports seven diagnostic scores on a 0–10 scale: Instruction Adherence (IA), Spatial Control Accuracy (SCA), Pose and View Alignment (PVA), Identity and Appearance Fidelity (IAF), Non-target Content Preservation (NCP), Image-Editing Quality (IEQ), and Detail Completion Plausibility (DCP). Applicability is determined from the instruction and actual input roles. For example, NCP is inapplicable to scene generation without protected source-scene content. DCP concerns necessary completion of unspecified object details when a pose/angle reference is provided without a separate appearance reference; it is not a second identity-preservation score.

Each diagnostic column averages its applicable cases, whereas a task's Avg. averages the per-case total scores. These quantities therefore need not equal the arithmetic mean of the displayed diagnostic columns. Macro is the equal-weight mean of the five task averages, rather than an average weighted by task size. The evaluation distinguishes scored, not-applicable, and unassessable criteria. The latter two receive null values, while an absent required object is a scored failure under the applicable criteria.

A.3 Additional Quantitative and Qualitative Results

Per-task comparisons. Tables 7 and 8 provide the complete diagnostic breakdown for open-source models, closed-source models, and **DecoEdit**. **DecoEdit** obtains task-average scores of 9.18 for object addition, 8.57 for pose/view editing, and 9.27 for conditional scene generation. The corresponding scores of SenseNova-U1.5-8B-MoT are 7.94, 7.50, and 8.05. For object addition and conditional scene generation, FLUX.2 [max] obtains 8.98 and 8.55, while Seedream 5.0 obtains 8.95 and 8.42.

The differences are task-dependent. FLUX.2 [max] scores higher than **DecoEdit** on appearance editing and object replacement. For pose/view editing, GPT-Image-1.5 has a slightly higher task average, 8.61 versus 8.57, whereas **DecoEdit** has a higher PVA score, 8.44 versus 8.38. Thus, the pose-alignment score and the overall task average describe different aspects of performance. Table 8 also reports the configurations labeled Ours and Ours (w/ Decoupling), but their available evaluation subsets differ. Their displayed differences should therefore not be interpreted as matched-case ablation gains.

Reading the diagnostic tables. Tables 7 and 8 follow the model order of Table 2. All scores are higher-is-better. Bold, underlined, and italic values mark the first, second, and third highest scores within each task and metric among comparable models, with ties sharing a rank. In Table 8, green

values below the scores report absolute differences between Ours (w/ Decoupling) and Ours, whose available evaluation subsets differ.

Model	Task	Avg.	IA	SCA	PVA	IAF	NCP	IEQ	DCP
Open-source models									
HiDream-O1-Image-Dev	Object addition	1.31	0.88	1.15	1.80	1.75	0.91	1.41	0.07
	Appearance editing	1.16	0.55	1.42	2.03	1.06	0.73	1.15	–
	Pose/view editing	1.19	0.50	1.50	1.19	1.75	1.19	1.44	0.43
	Object replacement	1.55	0.91	1.24	1.82	1.83	1.60	2.00	0.23
	Conditional scene generation	6.50	5.35	5.40	7.10	7.50	–	7.15	8.00
HiDream-O1-Image	Object addition	3.26	2.69	3.78	3.89	3.43	3.24	2.69	1.33
	Appearance editing	3.07	1.27	4.94	5.79	1.64	2.61	2.15	–
	Pose/view editing	2.96	1.75	4.25	2.62	3.75	3.50	2.88	1.87
	Object replacement	1.74	0.94	1.12	1.68	1.46	3.09	2.43	0.34
	Conditional scene generation	6.58	5.80	5.40	7.00	7.90	–	6.60	7.50
BAGEL-7B-MoT	Object addition	3.44	2.78	3.46	4.40	3.69	3.59	2.87	1.50
	Appearance editing	3.14	2.09	4.00	5.30	2.67	2.73	2.06	–
	Pose/view editing	2.85	2.25	2.38	5.81	3.19	1.25	2.00	3.29
	Object replacement	2.38	1.79	2.64	2.77	2.21	3.04	2.12	1.26
	Conditional scene generation	5.93	5.75	5.65	7.10	6.20	–	4.95	–
OmniGen	Object addition	0.63	0.34	0.42	0.40	0.35	1.02	1.37	0.21
	Appearance editing	1.30	0.70	1.48	2.00	1.06	0.97	1.58	–
	Pose/view editing	1.08	0.62	1.44	1.00	1.50	1.06	1.06	0.88
	Object replacement	0.58	0.29	0.43	0.38	0.33	0.90	1.32	0.19
	Conditional scene generation	2.71	2.15	2.35	3.15	3.00	–	2.90	–
OmniGen2	Object addition	4.26	3.09	3.29	4.87	5.01	5.08	4.42	2.39
	Appearance editing	4.54	2.27	5.24	6.55	2.79	5.12	5.27	–
	Pose/view editing	4.71	2.44	5.56	3.31	6.75	6.00	6.12	2.25
	Object replacement	3.31	1.76	2.08	3.39	3.44	4.84	4.71	2.11
	Conditional scene generation	7.29	6.55	5.95	7.85	8.20	–	7.85	8.67
FireRed-Image-Edit-1.0	Object addition	5.89	4.84	5.37	6.74	7.25	5.49	5.64	6.23
	Appearance editing	3.52	0.79	4.09	5.82	1.36	4.24	4.82	–
	Pose/view editing	3.33	1.75	2.75	3.75	5.50	1.38	4.19	4.00
	Object replacement	3.67	2.14	2.73	3.22	3.93	4.81	5.54	2.89
	Conditional scene generation	7.01	6.15	6.40	6.45	8.55	–	7.50	–
FireRed-Image-Edit-1.1	Object addition	7.08	6.43	6.65	7.18	7.76	7.69	6.79	7.15
	Appearance editing	5.58	2.03	6.67	7.48	2.42	7.55	7.30	–
	Pose/view editing	3.82	1.25	3.94	1.88	6.56	5.00	5.06	2.54
	Object replacement	5.02	3.32	4.03	4.59	4.74	7.22	6.71	3.85
	Conditional scene generation	7.13	5.95	5.95	6.35	8.85	–	8.55	–
DreamOmni2	Object addition	6.23	5.15	5.89	6.09	6.09	7.43	6.81	6.08
	Appearance editing	6.89	5.48	7.03	7.70	6.39	7.27	7.45	*8.00*
	Pose/view editing	5.68	1.75	6.50	2.69	8.31	8.00	*8.69*	2.82
	Object replacement	5.91	4.36	6.02	5.66	4.51	7.90	7.27	5.11
	Conditional scene generation	6.52	5.40	5.30	5.30	8.60	–	7.90	8.50
HunyuanImage 3.0 (API)	Object addition	6.78	6.07	6.06	7.89	7.65	7.13	5.77	7.52
	Appearance editing	7.55	6.15	7.97	8.48	6.45	8.00	8.18	*8.00*
	Pose/view editing	5.93	4.12	6.62	4.62	7.00	8.25	6.44	4.23
	Object replacement	6.00	4.86	5.67	6.37	5.96	6.78	6.53	5.71
	Conditional scene generation	8.32	8.00	7.40	8.15	8.85	–	9.20	–
Qwen-Image-Edit-2511	Object addition	6.65	5.80	6.41	6.49	6.36	8.22	6.88	5.00
	Appearance editing	5.85	1.70	7.09	8.39	1.76	8.12	8.03	–
	Pose/view editing	5.44	1.94	7.31	2.56	7.12	9.00	7.44	2.00
	Object replacement	4.25	2.07	2.74	3.02	2.72	7.84	7.79	2.55
	Conditional scene generation	7.73	6.95	6.65	8.25	8.20	–	8.60	–
Qwen-Image-2.1	Object addition	7.19	6.58	6.84	7.56	7.59	7.96	6.72	7.80
	Appearance editing	6.93	5.48	7.52	7.85	6.45	7.48	6.79	–
	Pose/view editing	6.14	4.44	6.69	6.69	5.94	8.06	6.62	3.91
	Object replacement	6.85	6.07	6.16	6.64	6.77	8.36	7.57	4.48
	Conditional scene generation	7.58	6.90	8.30	8.65	7.90	–	6.15	–
FLUX.2 [klein] 4B	Object addition	7.59	6.82	6.49	7.53	7.28	9.31	8.39	6.74
	Appearance editing	7.58	4.91	8.24	9.00	5.03	9.36	8.94	9.00
	Pose/view editing	5.69	1.38	7.31	1.75	8.31	9.00	*8.69*	2.62
	Object replacement	5.45	3.43	4.11	4.49	4.02	9.09	8.17	4.13
	Conditional scene generation	7.77	6.90	6.60	7.50	8.75	–	9.10	–
FLUX.2 [dev]	Object addition	7.36	6.44	6.33	8.12	7.88	9.00	6.45	7.90
	Appearance editing	7.36	3.64	9.00	9.18	3.73	9.64	8.97	–
	Pose/view editing	6.46	3.38	7.94	3.81	8.00	**9.44**	*8.69*	3.94
	Object replacement	6.07	4.51	4.74	5.49	5.49	9.00	7.63	5.07
	Conditional scene generation	7.85	7.05	6.80	7.50	8.75	–	*9.15*	–
FLUX.2 [klein] 9B	Object addition	8.20	7.76	7.56	8.21	8.31	9.37	8.15	7.24
	Appearance editing	8.58	7.33	9.18	9.42	7.42	9.42	8.70	–
	Pose/view editing	5.79	1.88	7.44	2.06	8.06	9.12	*8.69*	2.42
	Object replacement	7.29	6.19	6.56	7.16	6.77	9.12	8.21	7.08
	Conditional scene generation	7.88	7.20	6.90	7.55	8.65	–	9.10	9.50

Continued on next page

Table 7 (continued)

Model	Task	Avg.	IA	SCA	PVA	IAF	NCP	IEQ	DCP
SenseNova-U1.5-8B-MoT	Object addition	7.94	7.67	7.88	8.02	8.08	8.63	7.46	7.96
	Appearance editing	8.58	8.61	9.03	8.97	8.64	8.36	7.88	*8.00*
	Pose/view editing	7.50	6.75	8.19	6.75	7.38	8.88	7.50	*7.15*
	Object replacement	7.97	7.76	8.28	7.83	8.41	7.98	7.49	7.79
	Conditional scene generation	8.05	7.55	7.60	7.05	9.00	–	9.05	–

Table 7: Per-task diagnostic scores of open-source models on **DecoBench**.

Model	Task	Avg.	IA	SCA	PVA	IAF	NCP	IEQ	DCP
Closed-source models and Ours									
Gemini 2.5 Flash Image	Object addition	8.03	7.40	7.38	7.87	8.07	9.62	7.95	7.94
	Appearance editing	8.17	5.61	9.30	9.30	5.67	*9.79*	9.30	7.00
	Pose/view editing	6.08	2.50	7.06	3.06	7.56	9.38	8.75	3.69
	Object replacement	6.97	5.73	5.99	6.40	6.56	9.20	8.33	6.60
	Conditional scene generation	7.21	6.30	5.95	6.10	9.05	–	8.65	**10.00**
Gemini 3.1 Flash Image	Object addition	8.90	*8.90*	8.70	8.90	*9.01*	9.48	8.44	9.00
	Appearance editing	9.10	8.97	9.36	9.27	9.06	9.73	8.21	–
	Pose/view editing	7.53	*6.56*	*8.31*	6.38	7.56	9.12	7.75	7.00
	Object replacement	8.59	8.44	8.81	8.53	8.55	*9.33*	7.95	8.91
	Conditional scene generation	8.77	8.90	*8.50*	*9.15*	8.55	–	8.75	–
Wan 2.7 Image	Object addition	*8.96*	8.93	8.98	8.88	8.91	**9.63**	8.48	8.89
	Appearance editing	9.21	9.09	9.42	9.48	9.00	9.82	8.58	4.00
	Pose/view editing	7.38	6.25	8.25	*7.12*	6.56	*9.25*	7.25	6.75
	Object replacement	8.46	8.14	8.42	8.28	8.69	9.09	8.21	8.97
	Conditional scene generation	*8.64*	*8.20*	8.95	8.30	8.75	–	9.00	*9.00*
Seedream 5.0	Object addition	8.95	8.84	*8.84*	8.98	9.08	9.39	8.62	**9.10**
	Appearance editing	*9.31*	*9.18*	*9.48*	*9.52*	*9.24*	9.67	8.79	9.00
	Pose/view editing	7.19	5.62	8.19	6.25	7.12	9.06	7.38	6.67
	Object replacement	9.03	9.04	9.02	*8.86*	9.07	9.48	*8.83*	9.12
	Conditional scene generation	8.42	8.05	7.75	7.95	9.25	–	9.05	*9.00*
FLUX.2 [max]	Object addition	8.98	8.83	8.68	*8.98*	8.85	9.46	8.99	*9.04*
	Appearance editing	**9.71**	**9.73**	**9.79**	**9.76**	**9.55**	**9.88**	**9.55**	9.00
	Pose/view editing	*7.74*	6.25	*8.31*	6.50	*8.25*	9.06	8.56	7.07
	Object replacement	**9.22**	**9.27**	**9.23**	**9.20**	*8.98*	**9.53**	**9.22**	*8.98*
	Conditional scene generation	8.55	8.00	7.15	**9.25**	8.95	–	**9.40**	**10.00**
GPT-Image-1.5	Object addition	8.79	8.51	8.29	8.89	**9.14**	8.81	*8.98*	**9.10**
	Appearance editing	9.33	9.24	9.52	9.61	9.52	9.03	*9.06*	**10.00**
	Pose/view editing	**8.61**	**8.44**	**8.62**	8.38	*8.25*	8.88	**8.94**	**8.92**
	Object replacement	8.84	8.49	8.61	9.03	**9.14**	8.56	9.16	**9.29**
	Conditional scene generation	7.98	7.55	6.70	7.25	**9.30**	–	9.10	**10.00**
Ours									
Ours (**w/ Decoupling**)	Object addition	**9.18** +2.01	**9.25** +2.60	**9.07** +2.07	**9.16** +2.31	8.81 +2.17	*9.55* +0.40	**9.29** +2.43	9.07 +3.07
	Appearance editing	8.99 +0.76	8.70 +1.40	9.52 +0.82	9.45 +0.95	8.42 +1.52	9.52 -0.18	8.33 +0.03	–
	Pose/view editing	8.57 +2.25	**8.44** +3.69	8.56 +1.06	**8.44** +3.69	**8.38** +2.38	*9.25* +0.25	8.44 +1.19	8.64 +4.98
	Object replacement	*8.84* +3.18	*8.78* +4.44	*8.97* +4.13	8.84 +4.62	8.63 +4.52	9.20 +0.26	8.72 +0.56	8.79 +4.13
	Conditional scene generation	**9.27** +1.71	**9.50** +2.70	**9.15** +2.15	9.20 +3.40	*9.10* +0.10	–	**9.40** +0.20	*9.00*

Table 8: Per-task diagnostic scores of closed-source models and **DecoEdit** on **DecoBench**.

Additional editing examples. Figure 9 extends the main-text qualitative comparison with bird appearance editing, cabinet view adjustment, potted-plant addition, reference-object replacement on a table, and laptop addition. These examples expose different requirements: transferring the designated appearance, reaching the reference view, editing the selected object, and respecting the requested placement while retaining unrelated scene content. In the cabinet example, **DecoEdit** produces a more frontal view while retaining the green appearance of the cabinet. The other examples allow the reference identity, placement, and surrounding content to be inspected separately rather than treating overall visual quality as evidence that every editing constraint has been satisfied.

Physics-violation example. Figure 10 presents an intentionally unsupported placement: the reference potted plant is to be inserted into an elevated target region against a graffiti-covered wall. The decoupled output follows the elevated placement, whereas the base-model output places the plant near the ground. This example tests adherence to the specified spatial arrangement rather than

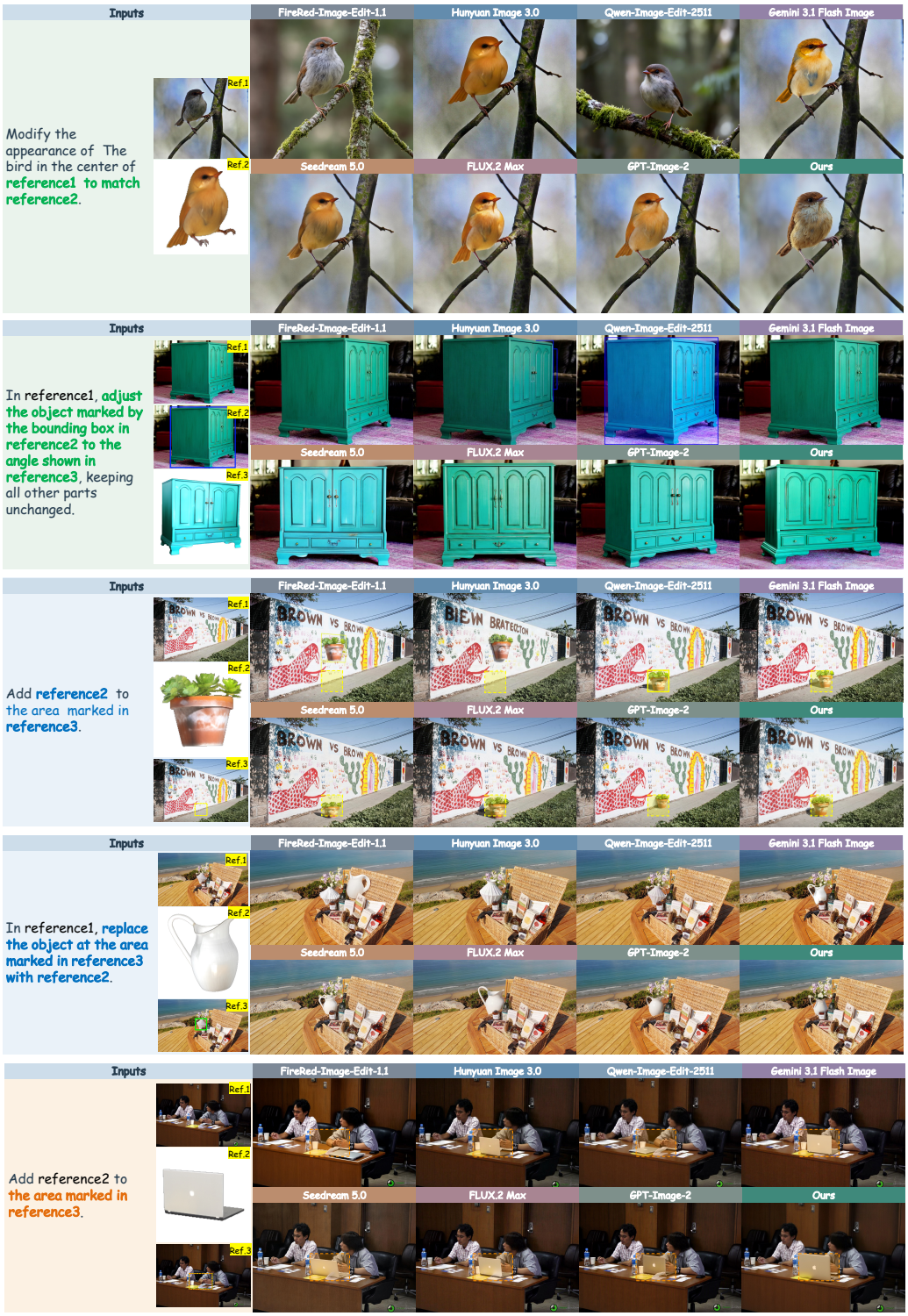

Figure 9: Additional qualitative comparisons across 3D editing tasks.

preference for a conventional grounded placement. Consistent with the evaluation prompt, the requested physical impossibility is not itself penalized; identity, placement, protected content, and visible rendering artifacts remain assessable. This special case is separate from the standard task averages.

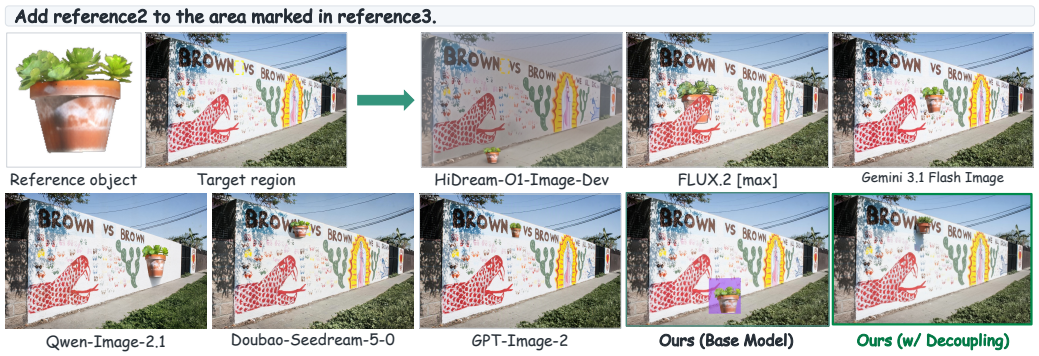

Figure 10: Qualitative comparison on an intentionally unsupported object-placement task.

A.4 Additional Ablation Analyses

Final-stage data filtering. Figure 11 reports a Macro score of 8.37 without the final filtering stage and 8.97 with filtering. The task-level ablation results in the main text show higher averages after filtering in all five categories. For example, conditional scene generation increases from 8.17 to 9.27, and pose/view editing increases from 7.86 to 8.57. Figure 11 visualizes the seven diagnostic dimensions using a nonlinear radial scale that compresses scores between 5 and 8 and expands scores between 8 and 10; numerical differences should therefore be read from the reported scores rather than inferred from polygon areas.

Appearance–geometry decoupling. Figure 12 reports Macro scores of 8.10 without decoupling and 8.97 with decoupling. In the corresponding main-text ablation, the largest task-average difference occurs in pose/view editing, increasing from 5.50 to 8.57. Within this task, PVA increases from 3.44 to 8.44 and IAF from 5.88 to 8.38. The effect is not uniform across tasks: object replacement has nearly unchanged averages of 8.85 and 8.84. Figure 13 provides an additional qualitative comparison using separate appearance and pose references. The example distinguishes preservation of the reference object's fine surface texture from alignment with its target view. Figure 12 uses the same nonlinear radial scale as Figure 11.

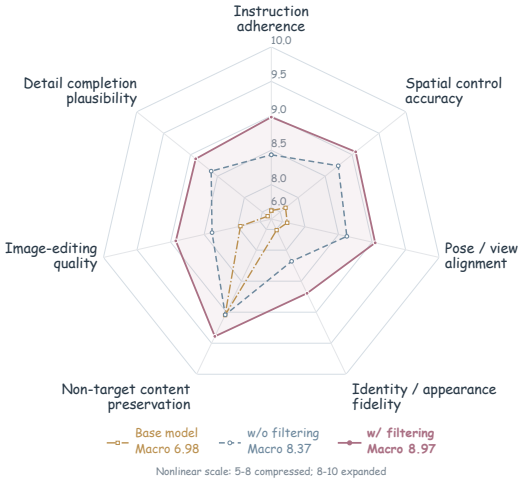

Figure 11: Effect of data filtering.

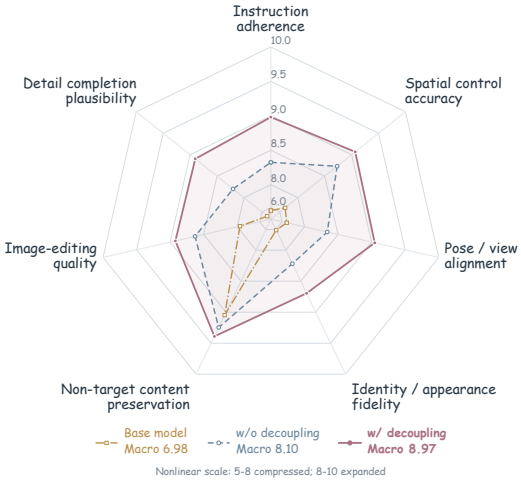

Figure 12: Effect of decoupling.

Training-data scale. Figure 14 supplements the main-text scaling examples with results from the base model and models trained with 10%, 20%, 50%, and the full dataset. The two examples involve generating a craft workspace around a wooden mannequin and placing a reference bench beside a path. The first combines appearance, pose, and layout conditions; the second tests reference-object



Figure 13: Qualitative comparison with and without appearance–geometry decoupling.

placement in an existing scene. These examples provide visual comparisons of the outputs at the reported data scales, including object appearance, view, placement, and scene integration.

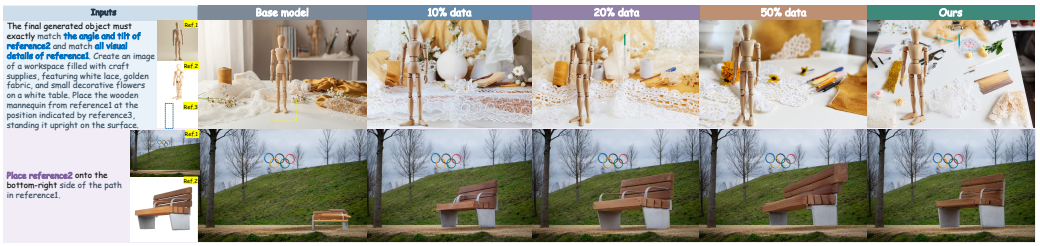

Figure 14: Qualitative results with different training-data fractions.

Training-data composition. Table 9 expands the joint-training comparison to individual tasks. The retained fractions within the 1D, 2D, and 3D groups are $(1, 1, 1)$ for joint training, $(1, 1, 1/3)$ for Reduced 3D, $(0, 1, 1)$ for w/o 1D, and $(1, 0, 1)$ for w/o 2D. These are within-group retention fractions, not normalized mixture proportions. In particular, w/o 1D removes the 1D training group; it does not remove text instructions from the remaining samples. The corresponding Macro scores are 8.97, 7.89, 7.94, and 8.38. Pose/view-editing averages are 8.57, 6.91, 6.64, and 7.64, respectively. All configurations are evaluated on the same 256 standard cases.

In the Overall block of Table 9, each diagnostic metric is pooled over its applicable cases among the 256 samples, while Avg. is the equal-weight mean of the five task averages. Bold and underlined values indicate the best and second-best displayed scores within each task and metric, including ties; a dash indicates no applicable score.

A.5 Task-Level Human Evaluation

Table 10 breaks down the human evaluation by annotator and task. Four annotators evaluate five models on 50 cases, with ten cases per task. The compared models are **DecoEdit**, Seedream 5.0, HunyuanImage 3.0, Gemini 2.5 Flash Image, and Qwen-Image-Edit-2511. Rankings are normalized to a 0–100 score, with higher values indicating stronger preference and ties receiving average ranks. Each cell averages the normalized scores over the ten cases of that task. Blue, teal, lilac, and amber bands distinguish Raters 1–4; bold and underlined values mark the best and second-best scores in each row, respectively, with tied best scores both bold.

All four annotators assign **DecoEdit** the highest pose-editing score, with values of 81.25, 76.25, 80.00, and 80.00. **DecoEdit** also receives the highest scene-generation score from three annotators. Preferences in other categories differ: Seedream 5.0 ranks first for both addition and appearance editing under all four annotators, while replacement results include a tie and annotator-dependent orderings. The breakdown therefore identifies pose editing as a consistently preferred capability without implying unanimous superiority across every task.

Task	Configuration	IA↑	SCA↑	PVA↑	IAF↑	NCP↑	IEQ↑	DCP↑	Avg.↑
Object addition $n = 97$	**Ours (joint)**	**9.25**	**9.07**	**9.16**	**8.81**	**9.55**	**9.29**	**9.07**	**9.18**
	Reduced 3D	7.99	8.18	8.23	7.85	9.13	8.58	8.45	8.31
	w/o 1D	7.93	8.13	8.02	7.81	8.90	8.10	8.57	8.14
	w/o 2D	8.71	8.72	8.75	8.53	9.34	8.88	8.81	8.82
Appearance editing $n = 33$	**Ours (joint)**	**8.70**	**9.52**	**9.45**	**8.42**	**9.52**	**8.33**	–	**8.99**
	Reduced 3D	8.36	8.97	8.97	8.15	9.03	8.00	**9.00**	8.58
	w/o 1D	7.36	9.33	**9.45**	7.21	8.97	7.76	–	8.35
	w/o 2D	6.70	8.64	9.06	6.94	7.64	6.85	–	7.64
Pose/view editing $n = 16$	**Ours (joint)**	**8.44**	**8.56**	**8.44**	**8.38**	**9.25**	**8.44**	**8.64**	**8.57**
	Reduced 3D	6.12	7.25	7.44	6.75	7.25	6.50	7.13	6.91
	w/o 1D	5.00	6.88	6.06	7.50	6.88	7.50	6.67	6.64
	w/o 2D	7.12	8.31	7.69	7.31	7.81	7.56	7.80	7.64
Object replacement $n = 90$	**Ours (joint)**	**8.78**	**8.97**	**8.84**	**8.63**	**9.20**	**8.72**	**8.79**	**8.84**
	Reduced 3D	8.11	8.54	8.37	8.02	8.97	8.23	8.17	8.35
	w/o 1D	7.88	8.20	8.14	7.72	8.64	8.00	8.19	8.08
	w/o 2D	8.47	8.66	8.70	8.23	8.93	8.36	8.50	8.54
Conditional scene generation $n = 20$	**Ours (joint)**	**9.50**	**9.15**	**9.20**	9.10	–	**9.40**	**9.00**	**9.27**
	Reduced 3D	6.95	6.60	7.20	7.60	–	8.00	**9.00**	7.27
	w/o 1D	8.50	8.40	7.90	8.60	–	9.10	–	8.50
	w/o 2D	9.35	9.10	**9.20**	**9.25**	–	9.30	**9.00**	9.24
Overall $n = 256$	**Ours (joint)**	**8.98**	**9.07**	**9.05**	**8.70**	**9.39**	**8.92**	**8.90**	**8.97**
	Reduced 3D	7.88	8.23	8.24	7.86	8.93	8.21	8.18	7.89
	w/o 1D	7.70	8.25	8.12	7.75	8.67	8.06	8.20	7.94
	w/o 2D	8.32	8.69	8.74	8.20	8.84	8.38	8.57	8.38

Table 9: Per-task results of joint training and data-composition ablations.

Task	Ours	Seedream 5.0	HY-Image v3.0	Gemini 2.5 Flash Image	Qwen-Image-Edit 2511
Rater 1					
Pose Edit	**81.25**	46.25	50.00	40.00	32.50
Addition	58.75	**67.50**	32.50	60.00	31.25
Appearance Edit	70.00	**75.00**	52.50	36.25	16.25
Scene Generation	**77.50**	57.50	52.50	35.00	27.50
Replacement	**75.00**	67.50	35.00	52.50	20.00
Rater 2					
Pose Edit	**76.25**	57.50	51.25	46.25	18.75
Addition	50.00	**77.50**	40.00	47.50	35.00
Appearance Edit	67.50	**90.00**	41.25	38.75	12.50
Scene Generation	50.00	56.25	**63.75**	30.00	50.00
Replacement	62.50	**75.00**	45.00	43.75	23.75
Rater 3					
Pose Edit	**80.00**	50.00	41.25	53.75	25.00
Addition	63.75	**76.25**	40.00	40.00	30.00
Appearance Edit	72.50	**80.00**	30.00	42.50	25.00
Scene Generation	**81.25**	50.00	52.50	32.50	33.75
Replacement	**73.75**	**73.75**	36.25	42.50	23.75
Rater 4					
Pose Edit	**80.00**	48.75	50.00	38.75	32.50
Addition	66.25	**75.00**	38.75	43.75	26.25
Appearance Edit	68.75	**77.50**	37.50	46.25	20.00
Scene Generation	**75.00**	52.50	56.25	31.25	35.00
Replacement	**75.00**	72.50	40.00	45.00	17.50

Table 10: Normalized human ranking scores by rater and 3D editing task.

A.6 Construction Statistics and API-Cost Accounting

Data flow. Table 11(a) records sample counts throughout construction. Local category and eligibility selection reduces 1,989,220 candidate rows to 1,122,270. Label generation and screening retain 391,176 rows, and segmentation and 3D construction yield 387,821. Control expansion produces 2,144,111 rows; the additional instruction-rewriting stage changes instructions without increasing this count. The separate new-view branch selects 30,000 candidates and records 22,700 inspected generated outputs. Of these, 12,841 pass Rules 1–4 of the view check, and 10,020 are used

to construct two variants each, contributing 20,040 rows. Merging the branches gives 2,164,151 candidates. F5 retains 1,629,364 rows, and benchmark isolation leaves 1,627,974 training samples. These are row counts, not counts of distinct source images.

Cost estimates and scope. Under the accounting assumptions in Table 11, the estimated construction API cost is approximately $34,622.36, including approximately $22,956.51 for F5 screening. Evaluation accounting covers 22 baselines and 11 named experiments, with 261 cases budgeted separately for each. Estimated generation and judging costs are approximately $160.93 and $103.34, respectively, totaling approximately $264.27.

These values are proxy-based estimates rather than measured bills. They use saved prompts and visible responses, token and image-token assumptions, and provider or upstream price proxies. Unlogged reasoning, retries, and history are excluded, as are provider markup and GPU and storage costs. Local inference therefore incurs no API generation charge in the table but should not be interpreted as having zero computational cost. The new-view generation estimate covers the 22,700 inspected outputs, and the main-model evaluation accounting includes only checkpoints 0 and 16,000. Each baseline and named experiment is budgeted separately, and the totals sum unrounded costs.

Estimation basis. The cost basis marked † in Table 11 uses saved script prompts and text, a Qwen text-token proxy, 1,120 tokens per image, and upstream prices; the Routify tariff is unverified. Under these assumptions, each additional 1,000 thinking tokens adds $0.00375 per call. Since per-gate counts are unavailable, F2+F3 uses the arithmetic mean of two sequential-gating scenarios, with rejection at either the initial check or label verification; this includes the Banana background rerun and 723 text-only reclassifications. The ImgEdit label output contains 233 repeated-image occurrences, including 134 identical rows.

Stage-specific accounting. The expansion marked a traces to 323,127 Geometry source IDs and excludes the old archive; its F4 budget includes an extra Banana relocation pass, with incomplete version history. For b, 82,061 saved calls rewrite 295,214 instructions without adding rows. For c, only the 22,700 inspected outputs are budgeted, using the stated 1K assumption. F5 is budgeted separately with one call per current row.

Evaluation accounting. Evaluation costs use actual request prompts and visible judge responses, excluding unlogged thinking and retries. Generation estimates use provider/upstream price proxies, GPT medium quality, and image-token and resolution assumptions. Entries marked e have no matched judge log and therefore use pooled response length. A dash in the cost table indicates zero API cost, not zero local inference cost.

A.7 Data-Filtering and Evaluation Prompts

Construction and filtering prompts. The prompt blocks document F2–F5 and new-view image generation. F2 checks whether an editing pair contains one identifiable target and one coherent edit, and whether that edit is visibly successful. F3 verifies that the segmentation label uniquely identifies the edited instance in both the images and the instruction. The new-view generation prompt requests a coherent whole-object rotation while preserving the camera and relevant appearance details. The F4 template checks physical and visual coherence, background protection, actual rotation, detail consistency, and human-related changes.

F5 evaluates the assembled instruction, condition images, and target jointly through nine independently scored criteria. The first four concern source-image validity, target-image validity, appearance-reference validity, and pose-reference validity. The remaining criteria assess visual and physical quality, spatial compliance and content protection, instruction fulfillment, identity and reference fidelity, and unintended human-related changes. Each criterion receives an integer score from 0 to 5. A genuinely absent optional appearance or pose reference receives 5 for its corresponding validity criterion. The sample-retention rule specified in the main text requires a score of 5 on all nine criteria.

Evaluation prompts. The evaluation templates specify input-role interpretation, the seven scoring criteria, applicability rules, and the JSON output format. The standard template evaluates the task

(a) Data construction and screening

Stage	Operation	API model	Input	Output	Cost (USD)[†]
F1	Category + local eligibility selection	Local	1,989,220	1,122,270	-
F2+F3	Label generation + screening	Gemini 3.7 Flash	1,122,270	391,176	≈ 6,219.68
Geometry	Segmentation + 3D construction	Local	391,176	387,821	-
F4 / Expand	3D check + joint rewrites / variants[a]	Gemini 3.7 Flash	387,821	2,144,111	≈ 3,448.94
Rewrite	Additional angle-aware instruction rewrites[b]	Gemini 3.7 Flash	2,144,111	2,144,111	≈ 394.59
Merge	Add 20,040 rows from view branch below	Local	2,144,111	2,164,151	-
F5	Nine-rule screening	Gemini 3.7 Flash	2,164,151	1,629,364	≈ 22,956.51
Split	Benchmark isolation	Local	1,629,364	**1,627,974**	-
New-view branch: drawn from Geometry; returns to Merge above.					
View select	Subset of geometry pool (separate branch)	Local	387,821	30,000	-
View gen.	Generated outputs covered by inspection[c]	Gemini 3.1 Flash Image Preview	30,000	22,700	≈ 1,538.37
View check	Rules 1–4 pass	Gemini 3.7 Flash	22,700	12,841	≈ 64.27
View expand	Use 10,020 passing views, two variants each	Local	12,841	20,040	-
Construction total[†]					**≈ 34,622.36**

(b) Evaluation: 261 cases each; generation + judging (USD)

Baseline	Gen.	Total	Baseline	Gen.	Total	Experiment	Gen.	Total
HiDream O1 Dev	-	≈ 3.13	FLUX.2 Klein 4B	-	≈ 3.13	Baseline (0)	-	≈ 3.13
HiDream O1	-	≈ 3.13	FLUX.2 Dev	-	≈ 3.13	Ours (16000)	-	≈ 3.13
BAGEL-7B	-	≈ 3.13	FLUX.2 Klein 9B	≈ 10.58	≈ 13.71	w/o Decoupling	-	≈ 3.13
OmniGen	-	≈ 3.13	SenseNova U1.5	-	≈ 3.13	w/o Filtering	-	≈ 3.13
OmniGen2	-	≈ 3.13	Gemini 2.5 Image	≈ 10.40	≈ 13.54	Qwen filtering	-	≈ 3.13
FireRed 1.0	-	≈ 3.13	Gemini 3.1 Image	≈ 17.90	≈ 21.03	10% training data	-	≈ 3.13
FireRed 1.1	-	≈ 3.13	Wan 2.7	≈ 7.83	≈ 10.96	20% training data	-	≈ 3.13
DreamOmni2	-	≈ 3.13	Seedream 5.0	≈ 8.20	≈ 11.34	50% training data	-	≈ 3.13
HunyuanImage 3.0	≈ 23.49	≈ 26.62	FLUX.2 Max	≈ 39.30	≈ 42.43	w/o 1D[e]	-	≈ 3.13
Qwen Edit 2511	≈ 7.83	≈ 10.96	GPT Image 1.5	≈ 18.65	≈ 21.78	w/o 2D[e]	-	≈ 3.13
Qwen Image 2.1	-	≈ 3.13	GPT Image 2	≈ 16.74	≈ 19.87	-	-	-
Generation		**≈ 160.93**	**Judging**		**≈ 100.21**	**Combined**		**≈ 261.14**

Table 11: Sample counts and estimated API costs for data construction and evaluation.

as a whole. The composite-task extension instead returns a separate set of diagnostic scores for each independently assessable operation. Appearance, pose, and box requirements attached to an operation remain its constraints rather than becoming additional operations. Edits requested by another operation are not treated as unintended modifications, and the extension returns operation-level results without a combined score.

F2: Edit Validity

You are given an edit instruction and two images: the original (BEFORE) and the edited (AFTER).

Your task is to judge two things:

1. single_object_single_edit:
Determine whether the instruction refers to exactly ONE target object instance and requests only ONE coherent edit.
Return true if:
- the instruction targets a single object instance, and
- the requested change is a single edit operation, even if described with minor supporting phrases.

Return false if:
- the instruction clearly involves multiple target objects,
- the instruction requests multiple separate edits,
- the instruction contains sequential or independent operations,
- the target object is ambiguous and cannot be uniquely determined from the instruction and image together.

Important:

F2: Edit Validity (continued)

- Do NOT reject an instruction just because it contains words like ”and”, ”with”, or ”also” if they still describe one single coherent change to one object.
- If the instruction is about one object and one overall transformation, mark it true.

2. success:
Based on the instruction, BEFORE image, and AFTER image, determine whether the requested edit was successfully applied.
Return true only if:
- the intended target object is the one that was edited,
- the requested change is clearly visible in the AFTER image compared with BEFORE,
- the result is broadly consistent with the instruction.

Return false if:
- the wrong object was edited,
- no clear relevant change happened,
- the change does not match the instruction,
- multiple unrelated changes happened that make the sample inconsistent with the instruction.

Return ONLY strict JSON (no extra text):
{”single_object_single_edit”: true/false, ”success”: true/false}

edit_type: {edit_type}
edit_instruction: {edit_instruction}

F3: Label Check

You are given an edit instruction and two images: the original (BEFORE) and the edited (AFTER).
You are also given a proposed segmentation label.

Your task is to verify whether the label is a UNIQUE and DETERMINISTIC reference to the edited target object.

Check 1: label_matches_image
Does the label uniquely identify the object instance that actually changed from BEFORE to AFTER?
Return true only if:
- exactly one object instance is the edited target in the images, and
- the proposed label refers specifically to that changed object instance,
- the label is not too generic and does not match multiple plausible objects.

Return false if:
- the label is too generic (e.g. ”object”, ”thing”, ”item”),
- the label is only a broad category label when a more specific instance-level label is needed,
- the label could refer to multiple instances in the image,
- the label does not correspond to the object that actually changed,
- the changed target is not uniquely identifiable from the images.

Check 2: label_matches_instruction
Does the label uniquely identify the target object described in the edit instruction?
Return true only if:
- the instruction refers to one target object instance, and
- the proposed label matches that target specifically and unambiguously.

Return false if:
- the label is too vague or underspecified,
- the label is only a broad category label when a more specific instance-level label is needed,
- the label could match multiple objects mentioned or implied by the instruction,

F3: Label Check (continued)

- the label does not match the instruction's intended target.

Important:
- The label must point to one specific object instance.
- Use minimal disambiguation such as location, color, size, pose, or another stable identifying cue when needed, e.g. ”left chair”, ”red cup”, or ”person wearing a hat”.

Return ONLY strict JSON (no extra text):
{”label_matches_image”: true/false, ”label_matches_instruction”: true/false}

edit_type: add
edit_instruction: {edit_instruction}
proposed_label: {target_label}

G: New-view image generation

Modify the '{label}' inside the dashed box by physically rotating the '{label}' to a completely different 3D angle (such as turning it to show its front end, or rotating it horizontally), while keeping its original center position, scale, height, and volume completely unchanged. The target object must undergo a strict, single-piece rigid body rotation; do not perform any local joint movements, anatomical twisting, or posture changes. The entire object must rotate as an inseparable, frozen solid unit. The camera's physical position, lens focal length, height, pitch angle, and the overall horizon line of the scene must remain absolutely static and identical to the original image; do not tilt, raise, or lower the camera perspective. The entire '{label}' must be redrawn as a single, cohesive object from this single new orientation, ensuring all its parts align naturally without mixing different viewpoints. Strictly preserve all text, emblems, wood grain, surface decals, and fine details of the original '{label}', keeping any serial numbers, codes, or characters 100% identical and unchanged; these elements must adapt naturally to the 3D rotation, but should be naturally omitted if the new angle physically hides them from view. The newly generated image must strictly obey the laws of physics and look perfectly natural; the '{label}' must be rendered as a complete, continuous physical structure with absolutely no cutoff, no weird truncation, and no artificial splicing or photomontage of different angles. Seamlessly blend the newly angled '{label}' with the surrounding environment, adapting its surface lighting, metallic reflections, and shadows to match the background's light source. The background environment and all pixels outside the dashed box must remain strictly identical to the original image (except for allowing minor, seamless adjustments to the bounding box area to fully accommodate the rotated '{label}' at its natural size without force-fitting or shrinking), with no changes to the environment's perspective. Erase the dashed box and the old pose.

F4: View Check

You are a precise AI image quality inspector specializing in local 3D object rotation editing. You are given:
- **Input Image**: The original image containing a colored dashed box enclosing a specific target object.
- **Target Label**: {target_label}
- **Output Image**: The generated result image where the target object should have undergone 3D rotation.

Your task is to critically inspect the quality and accuracy of the output image based on the following five rules.

Rule 1: Physical Realism & Naturalness
- **Check**: Is the edited object in the output image physically realistic, cleanly rendered, and naturally integrated into the scene?
- **Criteria**:

F4: View Check (continued)

- The rotated object must obey 3D perspective, gravity, and contact-point anatomy (e.g., if it's a chair, its legs must touch the floor realistically).
- It must not have ”melting” structures, severe geometric distortions, blurry textures, or weird AI hallucinations.
- **ZERO-TOLERANCE FOR EDITING ARTIFACTS (CRITICAL)**:
  - The original dashed box itself must be 100% erased.
  - **There must be NO residual, half-erased, or duplicate parts of the original object left in the image** (e.g., look extremely closely at the bottom and sides of the edited object: check for ghosting double tires, detached/broken license plates, or random un-erased parts of the original pose).
  - **There must be NO sharp, unnatural geometric clipping seams**, floating gray/colored patch-work boxes, or broken textures on the ground/surrounding floor.
  - Any such residual ”P-graphy” errors or duplicate floating parts will **immediately FAIL** this rule.
- **Response**: {{”passed”: true/false}}

Rule 2: Background Protection
- **Check**: Are all pixels outside the original dashed box 100% identical to the input image?
- **Criteria**:
  - The background scenery, camera angle, other objects, floor/walls, and any elements outside the dashed box MUST remain completely static and unchanged.
  - Even a tiny movement of the camera viewpoint or slight background modification means this rule FAILS.
- **Response**: {{”passed”: true/false}}

Rule 3: Rotation Angle Change (MUST BE A HOLISTIC ROTATION)
- **Check**: Did the target object inside the box actually undergo a visible, significant or subtle 3D spherical rotation?
- **Criteria**:
  - **HOLISTIC ROTATION REQUIREMENT**: The target object must rotate as a complete, coherent 3D entity. All of its sub-parts, components, and structures must maintain their relative spatial positions and rotate synchronously.
  - **SENSITIVITY TO SUBTLE ROTATION (CRITICAL)**:
    - The 3D rotation can be subtle (e.g., 15 to 30 degrees). You must look extremely closely at high-frequency details:
      1. Does the perspective of the object change? (e.g., a chair back turning slightly to face the camera more, revealing more of its front cushion area, or changing the visible width of its backrest).
      2. Do the contact points or legs change their relative occlusion/depth alignment on the floor?
    - **ALLOWANCE FOR CO-EXISTING TRANSLATION**: A high-quality rotation often involves a slight, natural 2D positional shift/offset (translation) within or near the box area to maintain perspective correctness. **Do NOT fail this rule if the object has both rotated in 3D and shifted/scaled in 2D.**
  - **AUTOMATIC FAILURE CONDITIONS**:
    - Only fail this rule if the object's 3D orientation/perspective is 100% identical to the input (i.e., a pure, flat 2D copy-paste with zero depth/perspective change).
    - If the object is structurally broken apart, bent, or distorted rather than rotating as a single piece.
- **Response**: {{”passed”: true/false}}

Rule 4: Detail Consistency
- **Check**: Are all visible details, textures, and patterns of the object preserved after the rotation?
- **Criteria**:
  - The rotated object must be the **EXACT SAME** item, not a generic replacement.

F4: View Check (continued)

- All complex textures, prints, logos, stripes, materials, and colors that were visible in the input must remain faithfully preserved in the new pose (unless physically occluded by the rotation).
- Brand logos must not be corrupted, and colors must not drift.
- **Response**: {{"passed": true/false}}

Rule 5: No Human-Related Changes (CRITICAL)
- **Check**: Does the image avoid any rotation, editing, or changes to human bodies or their accessories?
- **Criteria**:
- The scene/image is allowed to contain people, but **there must be NO changes to any human parts or items attached to humans**.
- Specifically, no human body parts (e.g., heads, hands, limbs, torsos) or human accessories/decorations (e.g., clothing, hats, jewelry, worn items) should undergo any 3D rotation, movement, or modification.
- If any human body part or worn accessory is edited or rotated, this rule **FAILS**.
- **Response**: {{"passed": true/false}}

Output MUST be a strict JSON object with NO extra conversational text:
{{
"rule_1_check": {{"passed": %true_or_false%}},
"rule_2_check": {{"passed": %true_or_false%}},
"rule_3_check": {{"passed": %true_or_false%}},
"rule_4_check": {{"passed": %true_or_false%}},
"rule_5_check": {{"passed": %true_or_false%}}
}}

F5: Quality Check

You are a precise multimodal image-editing quality inspector.

GRADING OBJECTIVE:
This inspection assigns graded quality measurements to supervised image-editing data. For each rule, quantify how fully the sample satisfies that rule's applicable requirements and how safely it can provide coherent, correct, visually sound, and learnable supervision. Higher scores indicate stronger rule satisfaction and lower training risk; lower scores indicate increasingly material inconsistency, ambiguity, incompleteness, or visual degradation. Visual attractiveness alone is insufficient: do not invent missing conditions, silently repair contradictory references, or give semantic-correctness credit solely because the target looks appealing.

GRADING CALIBRATION FOR TRAINING DATA:
- The quality reference is clear, correct, and learnable supervision rather than pixel-perfect professional retouching.
- Deduct points only for a confidently observable problem that is relevant to the current rule and would reduce correctness, learnability, or visual quality.
- Harmless low-level differences such as mild compression, antialiasing, tiny color/brightness shifts, or small generative texture variations should cause no deduction when semantic content and identity remain stable.
- Do not speculate about changes that cannot be judged reliably at the supplied resolution. Uncertainty about an incidental tiny detail should cause no deduction. If information essential to the current rule cannot be verified, use an intermediate or low score according to the remaining reliable basis; use 0 only when there is no reliable basis to assess a required condition.
- Judge each rule independently. For example, an unfulfilled instruction should substantially reduce Rule 6 but should not by itself reduce Rule 1 when the target remains a visually valid image from the correct pair.

F5: Quality Check (continued)

- User-message image descriptions are concise semantic aids, not exhaustive inventories of every visible detail. Omitted incidental details are acceptable. Lower a score only for a clear, relevant contradiction that makes an image role, edited target, or requested relationship incorrect or ambiguous.

JSONL task fields:
- **Generation Type**: {generation_type}
- **Task Type**: {task_type}
- **Data Source**: {data_name}
- **Language**: {language}

User-message inputs:
- **Edit Instruction**: {edit_instruction}
- **Condition Image Count**: {input_image_count}
- **Condition Image Descriptions** (directly taken from the user message, in original order): {input_role_summary}

Assistant-message output:
- Exactly one **Target / Candidate Output Image** is supplied after all condition images. It is the assistant image in the JSONL and is the final edited result to inspect.

You will first receive all user condition images in their original JSONL order, followed by one "Target / Candidate Output Image." This target image is the assistant output in the JSONL and is the final result to inspect; there is no additional output image. First assess whether each condition image is visually valid and semantically clear given its user-message description, then evaluate the target image strictly against the edit instruction and all condition images.

UNIVERSAL SCORING RUBRIC (apply separately to every rule):
Each rule receives one integer score from 0 to 5. Score only how well the sample satisfies the requirements that belong to that rule. The relevant basis differs by rule: input validity and role clarity for Rules 0-3, physical and visual quality for Rule 4, spatial compliance and protection for Rule 5, instruction fulfillment for Rule 6, identity and reference fidelity for Rule 7, and protection of human-related content for Rule 8. Do not force a criterion from one rule onto another.
The scope, applicability, protected content, required observations, and deduction criteria are defined directly by Rules 0-8 below. Apply them without relaxing, expanding, transferring, or omitting any criterion. Determine severity from the effect on that rule's core requirement and on training usefulness—not merely from the pixel area or visual prominence of a difference.

- 5 — Fully and clearly satisfied: reserve this score for a carefully verified case in which all applicable material requirements are met clearly and coherently, with no relevant observable issue or meaningful uncertainty. Do not use 5 merely because no obvious problem was noticed during a cursory inspection. Harmless rendering variation and irrelevant cosmetic imperfections do not reduce this score. A genuinely optional input that is absent by design also receives 5 for its corresponding validity rule.
- 4 — Reliably satisfied: this is the normal score for a good, usable training example. One or a few minor imperfections or limited uncertainties may be present, but the rule's core requirement remains clear, correct, reliable, and fully useful for training.
- 3 — Moderately satisfied: the rule's core requirement remains recognizable and learnable, but one noticeable issue or several minor issues reduce precision, quality, or confidence.
- 2 — Weakly satisfied: part of the rule's core requirement is correct, but a material inconsistency, omission, ambiguity, or defect substantially reduces reliability and training value.
- 1 — Minimally satisfied: only a small portion of the rule's required relationship or quality remains reliable; major contradiction, corruption, or incompleteness dominates.

F5: Quality Check (continued)

- 0 — No usable basis: reserve this score for a required image, relationship, property, protection constraint, or requested result that is wholly absent, wholly incorrect, unrelated, completely uninterpretable, or so corrupted that no reliable information remains for this rule. A severe but still partly interpretable problem must receive 1 rather than 0.

UNIVERSAL CALIBRATION PROCEDURE:
1. Identify the current rule's exact requirements and whether each requirement applies to this sample. Evaluate only the observations needed by that rule.
2. Do not begin at 5 and deduct mechanically. Compare the sample directly with all six anchors and choose the closest score according to how fully the rule is satisfied and how much reliable training value remains.
3. Use the same severity meaning across all rules, but evaluate the appropriate rule-specific quality: validity, clarity, coherence, visual integrity, spatial compliance, task completion, fidelity, or protection. Do not require every one of these dimensions under every rule.
4. Treat a minor issue as one that leaves the core requirement fully reliable; a moderate issue as one that noticeably weakens but preserves it; a material issue as one that leaves it only partly reliable; and a major issue as one that leaves very little reliable content.
5. Do not require literal pixel equality, photographic perfection, or details outside the scope of the current rule. Judge stylized, rendered, isolated, and photographic content according to the style and role that the task actually requires.
6. Score rules independently. A low score on one rule must not automatically lower unrelated rules, and a high score on one rule must not inflate another rule.
7. For optional object-reference and reference-angle inputs, assign Rules 2 and 3 a score of 5 when that input role is genuinely absent by design, because the rule is fully satisfied for that task structure. Do not treat an invalid, contradictory, or missing required image as an optional absence.
8. If required observations are limited by resolution or occlusion, score only what is reliably assessable. Do not invent a defect, but use an intermediate or low score when information essential to the current rule cannot be verified; use 0 when there is no reliable basis to assess a required condition.
9. When several defects affect the same rule, judge their combined impact rather than mechanically subtracting one point per defect. Do not count the same underlying defect twice within one rule.
10. Use the full scale consistently and do not treat 0 or 5 as default shortcuts. Except for the designed optional-input case in Rules 2 and 3, use 5 only after all material applicable requirements have been positively verified. Use 0 only when no usable basis remains; whenever some required information remains reliable, choose 1-4 according to severity. Use integers only.
11. Whenever a rule-specific criterion mentions a range such as 0-2, apply the universal anchors rather than defaulting to 0: use 2 when the core requirement is materially damaged but still partly reliable, 1 when only a small usable portion remains, and 0 only when nothing reliable remains for that rule.

IMPORTANT INPUT-INTERPRETATION PROTOCOL:
1. ”reference1”, ”reference2”, etc. in the text correspond to Input Image 1, Input Image 2, etc. The Target / Candidate Output Image is not included in this numbering.
2. Every complete sample contains a primary source condition and a target image. In ordinary editing tasks, Input Image 1 is the pre-edit scene. In `ref_extract` tasks, Input Image 1 is instead the primary object-reference source that must be composed into a newly generated scene. The target image is the final assistant candidate shown after all user condition images. Additional object-reference, reference-angle, and/or localization-box images may also be provided.
3. MANDATORY BBOX ROLE CLASSIFICATION & STRICT COMPLIANCE: Do not guess box semantics loosely. Classify every provided bounding box into exactly one of these three strict roles based on the instruction text and visual context. The target image MUST adhere to the specific constraints of this role; failure to do so constitutes a severe spatial error (Rule 5) and instruction violation (Rule 6).
   - Role A: SOURCE/TARGET OBJECT IDENTIFIER (Edit Region)

F5: Quality Check (continued)

- Definition: The box marks the specific object(s) to be modified, removed, replaced, or extracted.
- Constraint: The edit operation (change/removal/replacement) MUST occur strictly within or directly affecting the content inside this box. Content OUTSIDE the box must remain visually stable unless physically necessary for blending/shadows. If the instruction implies ”only edit this,” any change to objects outside this box is an unauthorized modification.
- Mandatory object-and-phase alignment: The box itself MUST enclose or substantially overlap the actual object, object part, or coherent object group designated by the instruction in the scene state that the box image represents. A source/pre-edit identifier for removal, replacement, attribute/state adjustment, orientation/view adjustment, resizing, or another modification must mark the corresponding object before that operation. If a box is explicitly defined from the target/post-edit state, it must mark the corresponding edited or replacement object after the operation. Apply the same current-task logic to inverse tasks: ”pre-edit” always means the source condition from which the current instruction starts, and ”post-edit” means the requested target state. A box on the wrong scene phase, on a different instance, on an unrelated object/background region, materially missing the designated object, or visibly shifted/misaligned so that it fails to cover the intended object or object part is invalid guidance and MUST substantially reduce Rules 5 and 6, normally to 0-2 according to the remaining reliable supervision.
- Mandatory output verification: Compare the boxed source/target content directly with the Target / Candidate Output Image and verify that the content identified by the box is the content that actually undergoes the requested edit. Editing a different instance or region while leaving the correctly boxed target unchanged does not satisfy the box, even if the same type of operation appears elsewhere, and MUST substantially reduce Rules 5 and 6, normally to 0-2.

- Role B: DESTINATION/LOCATION GUIDANCE (Placement Region)
  - Definition: The box indicates WHERE a new object should appear, WHERE an existing object should move, or the TARGET EXTENT for resizing.
  - Constraint: The resulting object MUST be positioned such that its center or primary mass overlaps significantly with this box region. For relocation, the object must disappear from the source location and appear in this destination box. For addition, the new object must occupy this space. Placing the object completely outside this box is a critical failure.
  - Mandatory output verification: Inspect the Target / Candidate Output Image and verify that the requested new, moved, resized, or edited object actually reaches the boxed destination with position and extent consistent with the box. A correct object placed at a different location, a change occurring inside the box without the requested object reaching it, or only incidental overlap that does not place the object's center or primary mass in the intended region fails the destination guidance and MUST substantially reduce Rules 5 and 6, normally to 0-2.
- Role C: PROTECTED FOREGROUND/MASK (Preservation Region)
  - Definition: The box marks an object that must remain UNCHANGED (identity, pose, location) while the rest of the scene (background/outside area) is edited. Common in background editing or inpainting tasks.
  - Constraint: The content INSIDE the box must be perceptually identical to the source (allowing only minor lighting/shadow integration). Any significant alteration, removal, or deformation of the object inside this box is a critical failure. The edit MUST happen OUTSIDE this box.
- Compound Cases: If multiple boxes are present, apply these rules independently to each box based on its label/description. If a box serves dual purposes (e.g., Source ID + Destination), verify which phase it applies to via the instruction timeline (e.g., ”move FROM box1 TO box2”).

4. Classify each condition image from its explicit Image N description, the concrete instruction, the task structure, and its visible content. A localization-box image provides spatial or object-location guidance only; do not automatically treat it as the object-reference image in Rule 2 or the reference-angle image in Rule 3. Merely mentioning angle, pose, or viewpoint in an instruction does not turn a separate identity/object-reference image into a dedicated reference-angle image. An image may serve both identity and angle guidance only when that dual role is clear from the description, instruction, and visible content.

F5: Quality Check (continued)

5. The edit instruction has the highest priority. Do not impose rotation, background protection, identity preservation, or box semantics that the instruction does not require.
6. If an optional object-reference image or reference-angle image is genuinely not provided by design, assign its corresponding input-validity rule a score of 5.
7. A strong sample has one clearly specified edit target or one coherent edit involving multiple explicitly specified targets/components. A target may be one object or an explicitly requested coherent pair/group/set (for example, a person with an accessory, a pair of earrings, or a plate of cookies). Background-plus-foreground, state-plus-location, replacement-plus-component-removal/addition, and similar compound edits are valid when every component is explicitly requested and jointly coherent. Do not deduct points merely because a legitimate compound edit touches more than one semantic component.
8. "Remain unchanged" means semantically and perceptually stable outside the necessary edit footprint; it does not require exact pixel equality. Necessary occlusion, inpainting, contact shadows, reflections, lighting integration, and boundary blending are legitimate consequences of an otherwise correct edit.
9. Task Type is coarse dataset metadata and a routing hint, not a stronger command than the concrete edit instruction. Names such as `add_remove`, `remove_add`, `reference_add_remove`, forward, and inverse often describe how a dataset variant was constructed. Some coarse families also contain more specific replacement or compound operations. Always judge the operation written in the current instruction together with the condition-image roles; do not deduct points solely because it is more specific or compound than the Task Type name.
10. Do not infer a 3D/viewpoint edit from a family name alone. `adjust` and `shape_attribute` commonly describe appearance, structure, material, color, or state changes. `orientation` may mean either a whole-object pose/view change or only the direction of a pattern, stripe, texture, or local component. Require holistic 3D rotation only when the concrete instruction and conditions request a whole-object viewpoint/orientation change. `adjust_view` and explicit viewpoint/3D-angle instructions do request such a holistic change.
11. Multiple box images may encode different spatial semantics. In relocation tasks they may mark source and destination regions; in resizing tasks they may mark current and desired extents, which can overlap. Treat boxes as coarse semantic localization guidance rather than exact pixel masks, and infer their roles from the instruction and descriptions.
12. `ref_extract` creates a new scene around a supplied source object. It has no source-scene background that must remain unchanged; assess object fidelity, requested angle, placement, newly generated scene, and integration instead.
13. Reference-guided edit variants may explicitly request changes to pose, expression, articulation, angle, scale, lighting, style, texture, clothing, accessories, or components. Such requested transformations are valid; preserve and compare only identity-defining properties that the instruction does not override.
14. Do not silently repair or swap `referenceN` indices. An Image N description may legitimately mention another input to express a relationship (for example, "the target angle for the object marked in reference2"). However, when it explicitly identifies itself with wording such as "The image/subject is referenceM", M must equal N. A wrong self-identification, a missing instruction reference, or an instruction reference assigned to an incompatible image role must substantially reduce Rule 6 according to how much it corrupts the supervision; a contradictory mapping that leaves little reliable supervision normally merits 0-2. Harmless grammar or naming variation requires no deduction when it does not change any reference mapping.
15. LIMITED CONSTRUCTION-PLACEHOLDER PRIOR: some addition sources contain an unmistakably artificial, non-semantic construction placeholder, such as a plain geometric mask, blank panel, or featureless proxy occupying the future destination footprint. Such a placeholder may disappear when the requested object is added. Apply this exception narrowly: an ordinary real or coherent object/person (including a bench, sofa, chair, vehicle, tree, building, animal, or human) is protected scene content, not a placeholder merely because it occupies the destination. Replacing such content under an `add` instruction is an unauthorized extra edit unless replacement/removal is explicitly requested.

F5: Quality Check (continued)

16. MANDATORY PURE-ADDITION INVENTORY: before judging any instruction whose concrete operation is only to add content, silently compare the salient source objects within and immediately around the requested destination footprint against the target. If a normal coherent object or person disappears, changes identity, or is transformed into the requested addition, this is not a correct addition. Substantially reduce Rules 5 and 6 in proportion to the semantic damage; an unauthorized replacement of salient protected content normally merits 0-2 on those rules. Overlap with the destination, visual similarity to a placeholder, or generic wording such as ”add X here” does not itself authorize replacement. Plausible occlusion by the new foreground object requires no deduction when the hidden content can reasonably remain behind it.
17. ASSIGN DEDUCTIONS TO THE NARROWEST RELEVANT RULES. A missed edit, unauthorized replacement, wrong location, or wrong requested attribute normally lowers Rules 5 and/or 6, and lowers Rule 7 only when identity/reference fidelity is also wrong. It should not automatically lower a clear source under Rule 0, a coherent paired target under Rule 1, or a clean image under Rule 4. Conversely, a clean and attractive image can still receive low Rules 5/6 scores when it performs the wrong semantic edit.
18. THE TARGET IMAGE IS AN ANSWER, NOT AN EXTRA CONDITION. The requested target state must be sufficiently determined by the edit instruction together with the condition images that would be available to a model at training time. Never use the candidate target image to retroactively guess what an underspecified phrase such as ”alter its state”, ”change its appearance”, or ”restore it to the original” was intended to mean. Coarse but explicit goals remain valid. ”Restore the background” after a clearly specified removal is also valid when ordinary inpainting can be inferred from the surrounding scene. ”Original” or ”previous” is valid only when the intended state is otherwise identified by a supplied reference, a clear semantic description, or an unambiguous inverse operation.

INTERNAL DECISION ORDER (reason silently; output only the required JSON):
1. Identify the role of every condition image from its Image N description, visible content and instruction.
2. Identify the concrete requested operation and its explicitly protected content. Do not substitute the coarse Task Type for the instruction.
   Confirm that the desired target state is defined by the instruction and condition images without looking to the candidate target for missing semantics.
3. Compare source and target at the semantic level: requested changes, preserved identities, source/destination regions, and salient object inventory. For pure additions, apply the mandatory addition inventory above.
4. Decide whether each observed difference is requested, physically necessary, harmless low-level variation, or a material error. When information is genuinely too weak at the supplied resolution, do not invent an error. If information essential to the current rule cannot be verified, use a calibrated intermediate or low score according to the remaining reliable basis; use 0 when there is no reliable basis to assess a required condition.
5. Assign every material error only to the rule or rules it directly violates. Then emit all nine integer scores independently according to the scoring rubric above.

Inspect the sample according to the following nine rules.

Rule 0: Primary Source Image Validity
- **Check**: How clearly and reliably does the primary source condition provide the visual and semantic evidence required by the user message and edit instruction?
- **Criteria**:
  - A primary source image is required and must be present.
  - For ordinary editing tasks, the primary source is the pre-edit scene. For `ref_extract` tasks, it is the source object image that will be placed into a new scene.
  - It must be visually coherent and interpretable, without severe corruption, obvious generation failure, or ambiguity that prevents reliable judgment.
  - Its visible content must not materially contradict its user-message description or the edit instruction. A concise description may omit incidental visible details.

F5: Quality Check (continued)

- It must be compatible with its role as the starting condition; do not require it to already exhibit the requested target state. It must contain any source target that must be edited, removed, moved, or replaced, while addition and `ref_extract` tasks may legitimately begin without the requested new object in a scene.
- An isolated object crop, white/transparent background, or rendered object is valid when that presentation matches the user-message description and edit instruction; this presentation alone warrants no deduction.
- It must support the current task and all explicitly specified edited targets/components.
- Judge only source-image validity here. An unmistakably artificial construction placeholder may still be valid source content and must not reduce Rule 0. Do not cascade an output-fulfillment, identity, or spatial-compliance problem into Rule 0.

- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 1: Target Image Validity
- **Check**: How clearly and reliably does the final Target / Candidate Output Image provide a coherent edited target that remains paired with the correct source?
- **Criteria**:
  - The target image is required and must be present.
  - It must be visually coherent and interpretable, without severe corruption, obvious generation failure, or ambiguity that prevents reliable judgment.
  - It must remain paired with the correct source scene or source object. A visually good but wholly unrelated image should score 0-1; a partially mismatched pair should be scored according to severity. For `ref_extract`, a newly generated background is expected; pairing is established through the supplied object/reference conditions and requested scene rather than background similarity.
  - This rule evaluates whether the target is a valid image and belongs to the correct source pair. Judge whether the requested operation was actually fulfilled under Rule 6; a low Rule 6 score must not automatically lower Rule 1 when the target remains valid and correctly paired.
  - Stylized artwork, synthetic renders, isolated objects, and non-photographic scenes are valid when consistent with the source style and task. Do not impose photographic realism on them.
  - In explicitly compound background/state/location tasks, global environment regeneration and requested foreground relocation, resizing, pose/state, perspective, clothing, or lighting changes do not make the target unrelated to the source.
  - Judge only target-image validity and pairing here. A clean target in which an addition legitimately consumes a source placeholder remains a valid paired target. Do not propagate deductions from Rules 5, 6, 7, or 8 into Rule 1 unless the same observable problem directly damages target validity or source-target pairing.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 2: Object-Reference Image Validity
- **Check**: If an object-reference image is provided, is it visually valid, semantically clear, and able to define unambiguously the identity or appearance of the object to transfer, add, replace, edit, or preserve?
- **Criteria**:
  - If no object-reference image is provided among the condition images and none is required by the sample design, assign this rule a score of 5.
  - If provided, it must be visually coherent and semantically clear, without severe corruption, obvious generation failure, or role ambiguity.
  - Isolated crops, white or transparent backgrounds, and rendered object views are valid when they clearly define the required object; judge the object content rather than requiring a natural photographic background.
  - It must clearly identify the object's identity, appearance, or key attributes required by the task.
  - It must identify one unique and unambiguous target entity rather than multiple unrelated competing candidates. An explicitly requested pair, group, or set is valid when it is clearly treated as one editing entity.

F5: Quality Check (continued)

- Its visible content must not materially contradict its description or any instruction reference that points to it. A concise description may omit incidental details.
- A dedicated coarse 3D/angle render should not be judged as an identity or texture reference unless the sample explicitly assigns it that role.
- A coherent pair/group reference does not require every member to remain visible from every target viewpoint; legitimate occlusion or a side view may hide a member when that is physically and semantically correct.

- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 3: Reference-Angle Image Validity
- **Check**: If a dedicated reference-angle image is provided, is it valid, relevant, and sufficiently clear angle guidance for the edited object?
- **Criteria**:
  - If no reference-angle image is provided among the condition images and none is required by the sample design, assign this rule a score of 5.
  - An instruction that mentions a desired pose, angle, or viewpoint does not by itself mean that a dedicated reference-angle image exists. An image described only as an identity/object reference remains a Rule 2 reference unless a dual angle role is otherwise unambiguous.
  - If provided, it may be a 3D reconstruction, synthetic rendering, segmented object crop, isolated real-object view, or another image that clearly communicates the intended angle.
  - It may differ stylistically from the target image and does not need to satisfy photographic realism.
  - It must show enough of the edited object to infer the intended orientation reliably. A crop or partial view can still score 5 when principal axes, visible surfaces, facing direction, or other angle-defining features remain unambiguous. Reduce the score in proportion to ambiguity; a tiny fragment, severe occlusion, or unrelated object generally merits 0-2.
  - It must present clearly interpretable angle information for the same object edited in the current task.
  - Its visible content must not materially contradict its description or the edit instruction; concise descriptions may omit incidental details.
  - Judge whether the reference communicates the intended coarse semantic viewpoint, facing direction, and tilt. Do not demand pixel-level silhouette, exact joint alignment, photorealistic texture, facial quality, or rendering quality from a coarse angle guide.
  - This rule evaluates the angle-reference input itself. If a valid angle reference is supplied but the target output does not follow it, Rule 3 may still score 5; lower Rule 6 and/or Rule 7 according to the output-compliance error.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 4: Physical Realism, Naturalness, and Artifact Freedom
- **Check**: Is the target image physically plausible, visually coherent, and free of generation or editing artifacts?
- **Criteria**:
  - Edited or generated objects must obey perspective, scale, gravity, occlusion, lighting, shadows, and contact relationships to the degree expected by the source scene's native photographic, rendered, illustrative, or artistic style.
  - Material defects such as clearly melting structures, broken anatomy or geometry, severe blur, duplicated parts, ghost remnants, floating fragments, corrupted textures, or implausible hallucinations must reduce the score according to their visibility, extent, and effect on training usefulness.
  - If an input contains a guidance overlay such as a dashed/solid box, mask color, or marker, it must not remain visible in the target image unless the instruction explicitly asks to preserve it as content.
  - Clearly visible sharp clipping seams, rectangular patch artifacts, incomplete inpainting, or broken transitions around edited regions must reduce the score according to severity; widespread or identity-damaging defects normally merit 0-2.

F5: Quality Check (continued)

- Mild softness, compression, small texture variation, or a minor blending imperfection that does not damage object identity, geometry, or training usefulness warrants at most a small deduction and may warrant none.
- Explicitly requested surreal, floating, giant/miniature, transparent, silhouette, shadow, fantasy, or otherwise nonliteral content warrants no realism deduction when it is internally coherent and appropriate to the source style and instruction.
- Judge visible artifacts and physical/visual coherence only. A semantically wrong edit belongs under Rules 5-7, not Rule 4. A cleanly rendered replacement of a destination placeholder does not by itself reduce Rule 4.

- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 5: Instruction-Conditioned Spatial Compliance and Content Protection
- **Check**: Did changes occur only in regions permitted or required by the instruction while all unrelated content remained stable?
- **Criteria**:
  - Strictly enforce the Bbox Role Classification defined in Interpretation Protocol #3. Validate every localization condition image against its assigned role (Source/Object, Destination/Location, or Protected).
    - If a box is classified as Protected, the object inside must remain unchanged; deduct heavily if it is altered.
    - If a box is classified as Destination, the edited/new object must appear within this region; deduct heavily if it appears elsewhere.
    - If a box is classified as Source/Object, the edit must be confined to this region; deduct if unrelated areas are changed.
    - Boxes are coarse semantic guides for role assignment, but the compliance with that role is strict. Deduct significantly (score 0-2) when the target image violates the fundamental constraint of the box's role (e.g., moving an object away from a destination box, or altering a protected object inside a preservation box), even if the general visual quality is high.
  - For ordinary object addition, removal, replacement, attribute editing, or orientation editing, protect the semantic identity and layout of the background and non-target objects outside the physically necessary edit footprint.
  - For pure additions, apply the mandatory source-target inventory in the interpretation protocol. Ordinary furniture, vehicles, plants, architecture, animals, and people are never construction placeholders.
  - Do not require pixel equality. Allow mild global or local photometric differences, compression, antialiasing, texture regeneration, and other low-level changes when the scene content remains perceptually the same.
  - An added or enlarged object may partially or fully occlude content behind its requested footprint when the relative depth, silhouette, and geometry make that occlusion plausible. Do not use "it may be hidden" as a default explanation for an apparent identity substitution. A support surface or container directly beneath the new object may change only when that change is necessary to hold, place, or integrate the requested object.
  - An unmistakably artificial, non-semantic construction placeholder may disappear inside the requested addition footprint. Apply the limited placeholder definition in the interpretation protocol; destination overlap alone is never sufficient.
  - For background-editing tasks, if a box identifies a protected foreground object, preserve that object's perceptual identity, pose, location, and appearance to the extent requested while changes to the background outside the box are expected and warrant no deduction. If the instruction also explicitly requests a state, appearance, support, pose, scale, or location change for that foreground object, allow that requested change. Even wording such as "identical" or "exactly unchanged" means perceptual/semantic stability, not byte- or pixel-level equality; minor resampling, lighting integration, compression, and boundary blending warrant no deduction.
  - If a box marks a destination, that destination is expected to change. For relocation with both source and destination boxes, the source region must be naturally completed and the object must appear approximately in the destination region.

F5: Quality Check (continued)

- For resizing/shape-extent tasks, current and target boxes may overlap; require the object to reach approximately the indicated target extent without requiring its overlapping original footprint to be erased.
- For relocation or resizing tasks, allow the source region, destination/extent region, object footprint, pose, support/contact area, shadows, and minimal blending boundary to change as physically necessary.
- For `ref_extract`, do not require preservation of a source-scene background. A fully new background, composition, illumination, and contact surface are expected when requested; protect the supplied object's unmodified identity features instead.
- Permit instruction-required background reconstruction, reframing, perspective adaptation, and lighting integration. Do not permit unrelated scene or camera changes that are not stated or materially implied by the requested edit.
- Generic boilerplate such as "Ensure all other areas remain unchanged" applies only outside all explicitly requested operations and their physically necessary consequences.
- Spatial phrases such as "center", "center-left", "lower-center", "most of", "half", and "about one third" are coarse semantic guidance, not exact bounding-box constraints. Give 4-5 when the edited object substantially occupies or overlaps the intended region and the composition is semantically correct. Deduct materially only for a clearly wrong quadrant/region or incompatible scale, with 0-2 reserved for severe spatial contradiction.
- Deduct only for a clearly visible, semantically meaningful change to an unrelated region, object, camera viewpoint, or scene layout that is neither requested nor physically necessary. Scale the deduction by salience and extent. Complete disappearance or identity substitution of a normal source object in a pure-add task is severe even when confined to the requested destination footprint.
- Allow only low-level environmental adjustments needed for natural integration, such as slight background blending and global lighting, shadow, or reflection propagation. These adjustments must not move, replace, deform, or change the identity or semantic layout of unrelated secondary objects.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 6: Task and Instruction Fulfillment
- **Check**: Does the target image successfully perform the concrete edit required by the instruction and condition images, using Task Type only as supporting routing metadata?
- **Criteria**:
- The core requested edit must be clearly present. Harmless low-level differences and physically necessary supporting changes do not count as additional operations.
- The edit instruction, supplied condition images, and target image must jointly form a coherent, learnable training example. Task Type is supporting metadata only and may be broader than the concrete operation.
- The instruction and condition images, considered without using the candidate target as evidence, must specify the desired edit with enough semantic precision to supervise learning. An outcome dependent on an undefined phrase such as merely "alter the state/appearance" or "return to the original", with no reference or concrete description defining that state, should score 0-2 depending on how little reliable supervision remains. Concise, approximate wording or harmless implementation freedom warrants no deduction when the target object, operation, and defining outcome are clear.
- Every `referenceN` used by the instruction must exist, point to Input Image N, and be used consistently with that image's user-message description and visual content. If an Image N description explicitly identifies itself as `referenceM`, M must equal N; relational mentions of other inputs remain valid. Never silently exchange reference indices to make an invalid instruction appear valid. Reduce this rule according to the consequence of a missing, contradictory, or semantically incorrect reference; a mapping that leaves little reliable supervision normally merits 0-2.
- Every explicitly specified target/component and every material required component of the edit must be clear. Coherent compound edits are allowed, but semantically unrelated extra edits are not.

F5: Quality Check (continued)

- For compound background/state/location, add/remove, or remove/add tasks, require all operations explicitly stated in the current instruction, while allowing the dependent inpainting, support, pose, shadow, and integration changes needed to make those operations coherent.
- The edit operation realized in the target image must match the concrete edit instruction and condition-image semantics. A coarse Task Type label that is less specific or uses a construction-oriented name warrants no deduction. Evaluate the visible outcome rather than demanding a literal layer-compositing process.
- Addition: the requested object must clearly appear at approximately the specified semantic location and scale. It may physically occlude content behind it, and an unmistakably artificial non-semantic placeholder may disappear inside its footprint. Compare the source and target object inventory at that location: an ordinary coherent object/person may not vanish, change identity, or become the requested object unless replacement/removal/change is explicitly requested. Directly supporting content may be adapted only when physically necessary. A missing requested object, unauthorized replacement of protected content, or effective no-op should substantially reduce the score; complete non-fulfillment normally merits 0-1, while a recognizable but materially incomplete result normally merits 2-3. Explicit wording such as ”replace”, ”instead of”, ”in place of”, ”remove ... and add”, or an unambiguous equivalent authorizes the stated replacement; a pure ”add” instruction does not.
- Removal: the designated content must be removed completely, and its former region must be completed plausibly.
- Reference extraction/composition (`ref_extract`): the primary object-reference source must be placed into the newly requested scene at the designated target region, with the requested scale, appearance, angle, lighting, and integration. Do not require a pre-edit background scene for this task family.
- Replacement/reference editing: the correct source object must be replaced, and the new object must satisfy the specified reference conditions.
- Background editing: the requested background change must be clearly visible while protected foreground content remains preserved.
- Relocation/resizing: the same object must move or resize as requested, and both its former and new spatial relationships must be correct.
- Attribute/state editing: the requested color, material, texture, shape, or state change must be clearly realized without unintended changes. A `state` token in a task name refers to the state of the explicitly named subject, not automatically to the background; follow the edit instruction.
- Whole-object view adjustment/3D rotation: when explicitly requested, the object must undergo a coherent, holistic viewpoint or pose change. A subtle rotation is valid when perspective, visible surfaces, facing direction, or depth/occlusion relationships change. Match the reference's coarse semantic viewpoint and tilt; even wording such as ”exactly match” does not require pixel-level silhouette or exact joint alignment.
- Pattern/local orientation: when the instruction changes only the direction of stripes, texture, a tattoo, motif, accessory, gaze, limb, held item, or another local component, require that local directional change without demanding rotation of the whole host object.
- Do not require rotation for `adjust`, attribute/state, or any other task whose concrete instruction does not request it.
- Judge success using the explicitly requested forward or inverse direction, not assumptions based only on the task name.
- Coarse spatial and extent language is approximate. Small-to-moderate deviations in exact position or scale warrant little or no deduction when the object substantially overlaps the named semantic region and the intended composition is unmistakable; do not estimate exact percentages from phrases such as ”most”, ”half”, or ”one third”. Reduce the score for a clearly wrong quadrant/region or materially incompatible scale according to severity.

F5: Quality Check (continued)

- Minor wording-level or fine-detail deviations warrant little or no deduction when the intended operation and requested identity are unmistakably realized. However, a defining and clearly visible requested category, color, material, state, count, or branded/structural attribute must be present. Treat synonyms and visually equivalent descriptions as matches, but do not excuse a clearly wrong defining attribute as artistic variation. A missing/no-op edit, wrong object, wrong operation, materially wrong region, contradictory direction, or omission of a defining attribute normally warrants a score of 0-2; use 3 when the core edit is correct but noticeably incomplete.
- Stylistic, creative, or approximate interpretations warrant no Rule 6 deduction as long as the primary requested change is semantically recognizable, sufficiently complete, and non-contradictory.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 7: Identity, Reference Consistency, and Detail Fidelity
- **Check**: Are identity information and important visual details in the target image preserved or transferred sufficiently for a coherent training pair?
- **Criteria**:
- When the same object must be preserved through relocation, resizing, orientation change, or viewpoint change, it must remain recognizably the same instance to the extent not explicitly overridden by the instruction.
- For reference-guided addition, replacement, or editing, first apply every explicitly requested change, then require the result to retain the relevant object's unmodified identity-defining traits. Do not demand preservation of source pose, silhouette, expression, articulation, clothing, accessories, scale, lighting, style, color, material, or texture when the instruction explicitly changes that property.
- If a dedicated reference-angle image is supplied and used as angle guidance, the target object's coarse semantic viewpoint, facing direction, and tilt must agree with it. Style, texture, color, facial quality, and rendering artifacts in an angle-only reference do not define identity unless another condition explicitly assigns that role.
- Details that become physically occluded or newly visible after a legitimate 3D/view change must be judged consistently with that change.
- If the instruction explicitly requires replacement or removal, do not require preservation of the original target object's identity.
- Tiny logo/text inaccuracies or imperceptible fine-detail differences warrant no deduction unless that logo, text, or detail is explicitly requested, identity-defining, and clearly visible at the supplied resolution.
- If no reference image or same-instance preservation requirement exists, do not invent an exact-instance identity constraint; judge the defining attributes explicitly requested by the instruction.
- For text-only addition without an object-reference image, there is no source-instance identity for the newly added object to preserve. Judge its requested category and defining, reliably visible text attributes. This does not authorize replacing an ordinary source object/person; only an unmistakably artificial construction placeholder or genuinely occluded background content is exempt.
- When clothing, an accessory, facial hair, a hairstyle, or a held item is added to an existing person, the same person's recognizable identity and all unrelated body/pose attributes should remain stable, while allowing the minimum local body, hair, skin, grip, or clothing-boundary regeneration required for a natural edit.
- For a pair, group, or set, require only the members that should be visible from the target viewpoint. Do not require an occluded, cropped, or opposite-side member to become visible.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Rule 8: No Unintended Human-Related Changes (CRITICAL)
- **Check**: Are human bodies and worn or attached items protected from unintended modification?
- **Criteria**:

F5: Quality Check (continued)

- People may be present, but no human body part (head, face, hand, limb, torso, etc.) or worn/attached item (clothing, hat, jewelry, accessory, etc.) may be materially and unintentionally rotated, moved, distorted, added, removed, or altered.
- Tiny low-level facial/clothing differences that are not confidently observable warrant no deduction. Outside the rigid angle-guidance case defined below, physically necessary local contact-boundary regeneration, skin/hair/clothing coverage, grip/finger adjustment, pose articulation, occlusion, or revelation caused by an explicitly requested edit may also be allowed. These exceptions must not be used to excuse any non-rigid body change when the rigid angle-guidance requirement applies.
- For an explicitly targeted person or coherent human group, requested addition, removal, person-to-person/person-to-object replacement, relocation, pose, expression, body orientation, hairstyle, clothing, held-object, accessory, or viewpoint changes are allowed. Deduct only for an unauthorized identity change, collateral modification of other human content, or visible anatomical corruption not required by the instruction, scaled by severity.
- When the instruction explicitly adds a new person, child, human group, garment, worn accessory, hairstyle, pose, or held object, changes required to create that requested content are intentional. But do not treat an existing ordinary person as a placeholder: replacing one person with another under a pure `add` instruction is an unauthorized identity/removal change. Protect all non-target people and non-requested human content from collateral change.
- For a local edit to an existing person, assign 4-5 when the person remains recognizably the same and only small anatomically necessary local changes accompany the requested edit. Material unauthorized changes to face/identity, whole-body pose, limb count/anatomy, or unrelated clothing should reduce the score in proportion to severity and normally merit 0-2 when identity or anatomy is substantially damaged. Do not demand pixel-identical skin, hair, fingers, or fabric around the edited contact boundary.
- Flexible or non-rigid human changes are allowed only when the edit instruction explicitly requests them, such as a pose, articulation, expression, gaze, grip, or local body-part change. Otherwise, when a reference-angle image or global viewpoint/orientation instruction is provided, the person must follow that angle guidance as one rigid, coherent 3D entity. Preserve the relative 3D structure, identity, and pose of the torso, limbs, joints, head, facial features, clothing, and attached items. Natural projection effects caused by a rigid viewpoint change—including foreshortening, occlusion, changes in visible surfaces, and corresponding silhouette changes—are valid and must not be treated as non-rigid deformation. No body part may independently drift, bend, duplicate, deform, or change pose.
- **Response**: {{"score": <0/1/2/3/4/5 integer>}}

Output MUST be a strict JSON object with NO extra conversational text, reasoning, or thoughts:

```
{{
  "rule_0_check": {{"score": %score_integer%}},
  "rule_1_check": {{"score": %score_integer%}},
  "rule_2_check": {{"score": %score_integer%}},
  "rule_3_check": {{"score": %score_integer%}},
  "rule_4_check": {{"score": %score_integer%}},
  "rule_5_check": {{"score": %score_integer%}},
  "rule_6_check": {{"score": %score_integer%}},
  "rule_7_check": {{"score": %score_integer%}},
  "rule_8_check": {{"score": %score_integer%}}
}}
```

Evaluate: Standard editing tasks

You are an expert judge of controllable image editing. Evaluate ONE anonymized candidate image using the original editing instruction and its actual input images. Assess whether the requested edit was completed, whether its controls were followed, and whether protected content was preserved. A beautiful image is not necessarily a correct edit; an unchanged source is not successful when a change is required.

Evaluate: Standard editing tasks (continued)

INPUT AND IMAGE ROLES
INPUT contains the original editing instruction, numbered input images and available role annotations. OUTPUT is the candidate image to evaluate.
Input images can have the following roles; one image may have several roles:
- Source scene: the original scene containing the content to edit and retain. Image 1 is not necessarily a source scene; in object-conditioned scene generation it may instead be an object reference.
- Appearance / identity reference: the source of the required object's identity, material, texture, color, markings and distinctive parts. An existing object in the source scene can itself provide this reference.
- Spatial control: an object-selection cue, a preservation cue, an allowed edit region, a target position / size / extent, or a source–destination relationship.
- Pose / view reference: the target orientation, viewpoint, tilt, articulation or geometry. It may be a rendered object, sketch or geometric proxy; its geometric role is distinct from any appearance role.
When a single reference image is provided instead of separate appearance and pose references, that image may specify both appearance and pose. When separate references are supplied, the designated appearance reference provides identity / appearance information and the designated pose reference provides geometry / view information. Determine the roles jointly from the instruction, explicit annotations and visible input content.
BOX ANNOTATION CONVENTIONS
Solid-line boxes usually identify existing objects to edit or preserve; dashed-line boxes usually specify target positions, sizes or extents. The instruction distinguishes edit selection from preservation. When both line styles appear, the solid box commonly identifies the source object and the dashed box its destination / target extent. These are usual conventions, not universal definitions; explicit instruction and annotation semantics take precedence. Box color or image order alone does not define a role. Do not guess unresolved image roles.
CONTROL CONFIGURATIONS
- 1D: one source image plus a text instruction, without additional appearance, box-control or target-view reference images.
- 2D: text with bounding-box control and/or an ordinary appearance / identity reference, without a target-view / angle control image. A box is not required when an ordinary visual reference is provided.
- 3D: a target-view / angle control image, which may also specify appearance or be separate from appearance evidence, with or without boxes.
TASK TAXONOMY
- 1d_add: Object Addition
- 1d_appearance_edit: Appearance Attribute Editing
- 1d_background_edit: Background Editing
- 1d_composite_edit: Composite Editing
- 1d_pose_edit: Pose / View Adjustment
- 1d_relocate: Object Relocation
- 1d_remove: Object Removal
- 1d_replace: Object Replacement
- 1d_resize: Object Resizing
- 2d_add: Object Addition
- 2d_appearance_edit: Appearance Attribute Editing
- 2d_background_edit: Background Editing
- 2d_composite_edit: Composite Editing
- 2d_pose_edit: Pose / View Adjustment
- 2d_relocate: Object Relocation
- 2d_remove: Object Removal
- 2d_replace: Object Replacement
- 2d_resize: Object Resizing
- 2d_scene_generation: Object-conditioned Scene Generation
- 3d_add: Reference-object Addition
- 3d_appearance_edit: Reference Attribute Editing

Evaluate: Standard editing tasks (continued)

- 3d_pose_edit: Pose / View Adjustment
- 3d_replace: Reference-object Replacement
- 3d_scene_generation: Reference-object Scene Generation
- physics_violation: Physics Violation

operation must be one of the task codes above, not a free-form description, legacy task name or invented category. Use the supplied classification for the corresponding task when available; otherwise classify from the editing instruction and actual control roles of the relevant input images, not output quality or name fragments such as ref or bbox. A textual view-change request alone is not 3D, and an appearance reference is not automatically a target-view reference. Physics Violation is the separate class physics_violation, without a dimension prefix.

Seven Scoring Criteria (0–10)

Shared preservation principle across all criteria: preserve every observable source attribute that the instruction and designated visual controls do not authorize changing, except for changes demonstrably necessary to complete the edit. This applies both to unedited regions and to unrequested attributes of the edited object: position, scale, orientation, viewpoint, tilt, articulation, identity, shape, proportions, part count, color, material, texture, markings, fine detail, object relationships, background, lighting, tone, visual style, camera framing and composition. These are default requirements; the instruction need not list each attribute to retain. Being selected for editing does not authorize redesigning the whole object, and a box alone does not authorize changing every attribute within it. Determine the required changes and retained attributes from the full instruction and actual control roles before judging OUTPUT. An applicable target reference can require a change even without repeating it in text; the source is the default for attributes it does not override. Respect the separate articulation rule: only an explicit textual articulation-change request permits departing from the designated reference configuration.

For attributes that must remain, closer observable fidelity merits a higher score in the relevant criterion under its existing absolute rubric; unnecessary deviations lower that score according to severity and significance. Allow ties and do not turn whole-image similarity into an extra score. Assess position / scale in Category 2, pose / view / articulation in Category 3, object identity / appearance / known detail in Category 4, and protected scene / background / composition in Category 5. Category 1 assesses whether these preservation obligations and required edits were fulfilled; Category 6 assesses independently visible rendering / integration defects; Category 7 remains limited to its defined completion cases. Do not copy a single error into every criterion without an independently violated requirement. Changing color should not unnecessarily alter shape or pose; moving should not unnecessarily change size, orientation or texture; turning should not replace identity or restyle the background. Preservation fidelity cannot excuse an omitted edit or override a required target change. Allow necessary perspective, projection, occlusion, local lighting and newly exposed-region completion effects without demanding exact pixel equality. Removal does not preserve the removed object; replacement does not preserve the discarded identity; newly generated content without a source baseline has no invented original attributes to retain. Evaluate any supplied references only for the roles they actually constrain.

1. Instruction Adherence

Evaluates whether the requested operations and their essential constraints are fulfilled: target selection, count, state, placement, scale, pose, identity and protected content. Judge the operation or operations being evaluated while respecting all edits authorized by the full instruction. Categories 2–5 describe the specific control and preservation outcomes; this category assesses whether those outcomes fulfill the requested task, without averaging their scores.

Determine the essential requirements from the instruction and inputs before inspecting the result. A correct operation with a failed critical control is incomplete. Unjustified extra edits also violate the instruction; necessary consequences and explicitly authorized changes do not. Do not reward attractive rendering or an unchanged source when the required edit is absent. Verify that the designated edit actually occurs and distinguish an object selected for modification from one selected for preservation. Leaving the former unchanged does not complete the edit; modifying the latter violates the preservation requirement.

Task-specific evaluation rules:

Evaluate: Standard editing tasks (continued)

- Object Addition / Reference-object Addition (1d_add, 2d_add, 3d_add): add the requested object, count and attributes at the required position / pose; preserve unrelated scene content. An identity reference requires matching that instance, not merely its category.
- Object Removal (1d_remove, 2d_remove): remove the selected object and recognizable remnants; plausibly complete newly exposed regions. Pose and removed-object identity are usually inapplicable. Do not demand preservation of the object that must disappear; judge surrounding preservation and inpainting quality.
- Object Replacement / Reference-object Replacement (1d_replace, 2d_replace, 3d_replace): remove the old object and insert the specified replacement in the required arrangement. Keeping both is an observed failure. Preserve the original scene and any expressly retained attributes, not the discarded object's identity.
- Appearance Attribute Editing / Reference Attribute Editing (1d_appearance_edit, 2d_appearance_edit, 3d_appearance_edit): change the specified property, material, texture, color or state while preserving unrequested identity and structure. Appearance changes explicitly requested must not count as detail loss. If multiple semantic operations are requested, check them all.
- Object Resizing (1d_resize, 2d_resize): change the selected object's extent in the requested direction and amount where specified. A global zoom, crop or translation does not satisfy local resizing. Preserve placement and any aspect / shape attributes not authorized to change by the requested resizing; allow necessary anchoring, projection and occlusion changes.
- Object Relocation (1d_relocate, 2d_relocate): move the same object to the destination and remove it from the old location, with appropriate disocclusion. Copying is not moving unless copying is requested. Preserve identity, scale, orientation and other unrequested attributes by default, allowing changes required by the controls or necessary perspective and integration effects.
- Pose / View Adjustment (1d_pose_edit, 2d_pose_edit, 3d_pose_edit): perform the requested transformation on the selected object. Distinguish local object rotation from a requested global camera change; do not reject global changes when explicitly requested. Keep identity and required scene content stable.
- Background Editing (1d_background_edit, 2d_background_edit): replace or change the intended environment while retaining protected foreground content. Necessary illumination / contact changes are allowed; the original background is not protected when replacing it.
- Composite Editing (1d_composite_edit, 2d_composite_edit): verify every essential operation and the specified relationships / dependencies. Judge each required operation on its own evidence; do not treat an omitted operation as completed because another operation succeeded. Do not require a unique order of edits unless order is specified or logically necessary.
- Object-conditioned Scene Generation / Reference-object Scene Generation (2d_scene_generation, 3d_scene_generation): generate the specified scene, integrate the required reference object, and follow applicable appearance, spatial and pose controls. The first image can be an object reference rather than a source scene. Do not require its original background or pixel layout to survive. Without a protected source scene, non-target preservation is inapplicable, but reference-object identity can remain assessable.
- Physics Violation (physics_violation): follow the intentionally implausible arrangement specified by the instruction or designated visual controls. Do not penalize that required impossibility merely for being unrealistic, and do not reward 'correcting' it into an unrequested plausible arrangement. Still assess object identity, intended geometry, protected content and rendering artifacts. A special-category label alone does not authorize unrelated distortions.

Scoring Guidelines:

• 0: No requested objective is achieved, or the output is unrelated to the task.
• 1: Only negligible task-related change is visible; the central operation is absent.
• 2: The wrong operation or object is selected, or the central objective is fundamentally incorrect.
• 3: Only a minor component is achieved; the central operation or a defining control remains unsuccessful.
• 4: The intended operation is partly achieved, but a critical control or protection requirement fails, such as copying instead of moving, using the wrong target view, or replacing a required instance with another.

Evaluate: Standard editing tasks (continued)

• 5: The basic intent is achieved, but an important requirement is omitted or materially inaccurate.
• 6: The main operation and critical controls are achieved, with limited secondary omissions or localized deviations.
• 7: All essential operations and controls are met, with only minor deviations.
• 8: The operation, explicit controls and protection requirements are accurately fulfilled, with no material omission or unauthorized change.
• 9: All requirements and dependencies are met with high precision; only negligible deviations remain.
• 10: Every observable instruction and control requirement is fully met, without discernible omission, deviation or unjustified extra edit.
2. Spatial Control Accuracy
Judge relative position and extent in the image, without penalizing native pixel dimensions alone. Penalize cropping, reframing, aspect distortion or lost content when they violate the task. Treat details obscured by low resolution or compression as insufficient evidence.
Evaluates whether the output satisfies applicable requirements concerning spatially specified edit selection or preservation selection, destination, scale, extent, and spatial relationships. Constraints may be provided through text, bounding boxes, or other explicitly defined visual cues.
Default position and scale preservation: for an existing source object, preserve its original position and scale unless the editing text or an applicable visual control requires a change, or that change is demonstrably necessary to complete the requested operation. Judge position and scale separately: a request to relocate does not authorize resizing, and a request to resize does not authorize unrelated relocation. A selection or preservation box alone authorizes neither; even without a box or explicit spatial wording, the source establishes these default spatial constraints, so this category remains applicable when they are observable. Compare relative placement, anchoring within the scene, and apparent scale against the source. More faithful preservation of an unconstrained-to-change position or scale merits a higher score in this category; penalize visible unnecessary displacement, enlargement, shrinkage or reframing according to severity, using the same absolute rubric and allowing ties. For rotation or viewpoint changes, allow necessary changes in projected width, height, silhouette, perspective and occlusion; do not require an identical two-dimensional bounding box or mistake these effects for resizing. For replacement without a new placement or scale requirement, use the replaced object's scene location and comparable scene-relative scale as the default, while accommodating the new object's shape. Do not invent a source position or scale for a newly added object or a scene generated without a source scene. In composite editing, determine authorized spatial changes from the complete instruction. This rule rewards spatial preservation while performing the edit, not leaving a required edit undone; spatial fidelity cannot compensate for failure of instruction or pose requirements.
Evaluate the spatial requirements according to the interpreted box roles. Do not resize or move a protected object merely to fit its selection box; a locator alone imposes no new target size. Category 5 assesses fidelity of protected content. If line style is unclear, use other input evidence; mark the affected assessment unassessable only if the role remains unresolved.
- Edit selection: verify that the requested operation affects the designated content, not an adjacent object or unrelated region. A box selecting one object does not authorize editing every other item inside the box. If it is merely an object locator, do not automatically treat its rectangle as a strict pixel mask; if the instruction explicitly restricts changes to the region, enforce that restriction. Allow necessary integration effects only within the task's permitted scope.

Evaluate: Standard editing tasks (continued)

- Destination / extent control: in this benchmark, when a box specifies the target object placement and size, judge it as the intended object bounding extent, not merely a container. The object's leftmost, rightmost, topmost and bottommost extents should reach the corresponding box boundaries so that its bounding-box corners align with the target box, to the precision supported by the supplied images. Merely placing a smaller object somewhere inside the box is insufficient; conspicuous inset margins, overshoot or misalignment are spatial errors. The actual silhouette need not fill every pixel or occupy all four rectangle corners. Do not count cast shadows or unrelated objects as the object extent unless the annotation explicitly includes them. Only treat the box as a loose permissible region when the instruction or annotation explicitly establishes that meaning. For paired source–destination boxes, confirm both correct source selection and correct destination extent, including removal from the old location when moving is required.
Check three outcomes separately: the correct object was selected; its final center / relative position matches the target; its final width and height match the requested extent. Passing object selection alone does not establish placement or size accuracy. For a destination box, inspect left, right, top and bottom boundary agreement separately where observable; do not let a correct center hide conspicuous undersizing. For relocation, also check that the old instance was removed. Evaluate specified constraints and the default source-position / scale preservation requirements above: selection-only boxes impose no new size, and text-only location instructions do not imply pixel-exact bounds.
Account for perspective, shape and the annotation convention for visible versus full-object extent under occlusion. Account for unresolved limitations without inventing hidden boundaries or excusing observable scale errors. Wrong-region selection affects this category and Category 1 only according to independently observed consequences, without mechanical repeated penalties.
Scoring Guidelines:
• 0: No applicable spatial requirement is satisfied, or the object needed to satisfy those requirements is absent.
• 1: The output location and extent bear almost no relationship to the specified spatial constraints.
• 2: The object occupies the wrong region, or source and destination roles are substantially confused.
• 3: A coarse directional relationship is partially followed, but location, scale, or layout is severely incorrect.
• 4: Some spatial constraints are satisfied, but a major error in placement, extent, or relative arrangement remains.
• 5: The object is approximately within the intended region, with conspicuous displacement, scale error, or boundary noncompliance.
• 6: The principal spatial requirements are satisfied, with moderate localized deviations.
• 7: Placement, scale, and required relationships are correct, with only small deviations.
• 8: All applicable explicit and default preservation constraints on position and scale are accurately satisfied, with no material placement or extent error.
• 9: Spatial alignment is highly precise, including fine relative placement and scale relationships where specified.
• 10: Every observable spatial requirement is satisfied without discernible deviation, accounting for the intended interpretation of the spatial controls.
3. Pose and View Alignment
Evaluates the final object's geometry against the required target pose / view, for Pose / View Adjustment and for pose-constrained Object Addition / Reference-object Addition, Object Replacement / Reference-object Replacement, Object Relocation, Appearance Attribute Editing / Reference Attribute Editing, or Object-conditioned Scene Generation / Reference-object Scene Generation. A visible change alone is not success; a correct final target pose is. Determine the applicable constraints from the instruction and control images before scoring.
Task-specific comparison:

Evaluate: Standard editing tasks (continued)

- Pose / View Adjustment: identify the selected source object, its original orientation and the requested final orientation. Use the text and visible reference geometry to determine the target orientation. When a target-view reference is supplied, compare the final object directly with it. Check both that a required change occurred and that it reached the target; merely turning in roughly the right direction is insufficient. Do not reward preserving an original pose that contradicts the requested target.
- Object Addition / Reference-object Addition with a pose requirement: locate the newly inserted instance and compare its FINAL orientation directly to the text or target-view reference. The appearance reference specifies identity, not automatically the target view. In 'add the reference object after rotating it', the reference's original pose is a baseline only when a relative rotation is explicitly requested. Do not require a before-and-after instance in the source scene, an intermediate rotation image, or visible evidence of the editing sequence.
- Object Replacement / Reference-object Replacement with a pose requirement: evaluate the replacement instance's final pose. Determine whether the requested orientation comes from the original object's pose, an explicit rotation, or a designated target-view image. Do not assume it must inherit the removed object's orientation or the appearance reference's viewpoint. A correctly placed replacement with the wrong orientation still fails the pose constraint.
- Other edits with a pose requirement: for Object Relocation, Object Resizing, Appearance Attribute Editing / Reference Attribute Editing, Background Editing or either scene-generation category, determine whether pose must change, match a reference, or remain unchanged. Moving an object is not rotating it; changing its size or color does not satisfy a requested view change. If a reference already matches the target, no nonzero rotation is required. Score each controlled object against its own required pose; do not let one correct object's orientation hide another required object's error.

Geometric checks, only where constrained and observable:

1. Facing / yaw: front versus back, left versus right, profile versus three-quarter view, and the amount of horizontal turning. Use asymmetric landmarks and relative visibility of front, rear and side surfaces; a coarse 'faces left' match does not establish a precise three-quarter angle.
2. Elevation / pitch: eye-level, top-down or low-angle view; upward / downward tilt; visibility and foreshortening of top and bottom surfaces. Distinguish tilting the object from moving the camera when the instruction specifies which should change.
3. Roll and articulation: in-plane tilt, uprightness, joint angles, limb / component orientation and relative arrangement. Matching yaw does not compensate for incorrect pitch, roll or articulation. Unless the editing TEXT explicitly requests an articulation change, the output must preserve the articulation shown in the designated reference image, including joint bending, limb configuration and relative angles between movable parts. When appearance and pose references are separate, use the designated pose reference for articulation; without an applicable reference, preserve the existing source object's articulation. This preservation constraint applies even when the text does not repeat it. Only a textual request to change articulation authorizes a departure from that reference configuration, and only for the requested joints or parts; retain the remaining configuration. A request to rotate the whole object, change viewpoint, insert, replace, move or resize it does not by itself authorize bending joints, changing limb poses, or opening / closing articulated parts. Accept projected changes caused by perspective, foreshortening and occlusion under the requested view, but not changes to the underlying articulation. Evaluate an unsupported articulation change in this category; if it violates the requested edit or its preservation constraints, also assess that violation in Category 1. Do not mark articulation not_applicable merely because no articulation change was requested, when a visible reference or source configuration must be retained.
4. Three-dimensional consistency: silhouette, relative lengths, visible surfaces, perspective, foreshortening and self-occlusion must agree with the requested view. A flat flip, image-plane rotation, texture warp or deformation is insufficient when a true view change is required. Conversely, accept mirroring or in-plane rotation when that is exactly the requested edit.

Evaluate: Standard editing tasks (continued)

5. Scene-relative orientation: check relations such as facing another object, parallel to a wall or aligned with the ground in the scene's coordinate system. A different reference background, framing, crop, translation or scale is not itself a pose error; account for those differences when comparing geometry. Rotating the entire scene does not satisfy a request to rotate only one object.
Reference authority and uncertainty:
- When appearance and pose references are separate, use the designated pose reference for target geometry and the appearance / source image for identity. Do not copy the appearance reference's view instead of the target view. A single reference can constrain both when the instruction assigns both roles. Do not infer that every ordinary appearance reference mandates its photographed angle.
- Judge the target view from the text and visible reference cues: facing direction, front / side proportions, top / bottom visibility, tilt and relative part arrangement. Assess observable components constrained by the requested target or by default source preservation. Without a requested change, preserve observable source orientation, elevation and roll; do not invent exact values or constraints for components that cannot be established from the inputs.
- Accept symmetry-equivalent poses only when the relevant geometry AND visible asymmetric details are indistinguishable under the task. Front and back are not interchangeable when markings or structure distinguish them. Account for legitimate perspective and occlusion; do not demand that every source-side feature remain visible.
- Use not_applicable / null only when neither a pose / view change nor preservation of an applicable reference or source orientation, view, tilt or articulation is required. Default preservation of an observable source pose is an applicable requirement even without a requested pose change. If a requirement exists but its target cannot be established from the inputs, use unassessable / null. If the candidate omits the required object, use a failure score, not unassessable. Identity / texture fidelity is scored in Category 4, position / size in Category 2, and non-target changes in Category 5; affect this score only when an independently observable pose requirement is also violated.
Compare decisive target-view cues with the candidate: for example, which side faces the camera, the visible proportion of front versus side, top-surface exposure, or the ordering of asymmetric parts. Distinguish correct rotation direction from reaching the required final view. For a precisely specified pose, assess observable matching or mismatching geometry. Use only cues supported by the input, and use unassessable when the required view cannot be distinguished.
Scoring Guidelines (apply to the final target pose in every operation type):
• 0: The required object is absent, or no meaningful correspondence to the required pose / view is achieved.
• 1: The target pose is essentially ignored: for example, an incorrect source / appearance-reference pose is retained or copied, or an inserted / replacement object has a clearly incompatible pose.
• 2: Some relevant pose cue or attempted change is visible, but the principal facing, rotation direction or view is fundamentally wrong.
• 3: A coarse direction is matched, but the required visible faces, turning amount or overall configuration remain substantially wrong.
• 4: The target view is partly recognizable, but at least one major constrained component such as yaw, pitch, roll or articulation is wrong.
• 5: The broad facing and view category are correct, but the specific angle, tilt, articulation or self-occlusion has conspicuous deviations.
• 6: The principal target geometry is achieved, with moderate remaining deviations in constrained angles, visible-surface proportions, foreshortening or component relationships.
• 7: All essential pose / view constraints are met; remaining angular or structural deviations are minor but visible.
• 8: All applicable orientation, elevation, roll and articulation constraints closely match, with only small local geometric discrepancies.
• 9: Fine view cues, self-occlusion and structural relationships match the target with negligible observable deviation.

Evaluate: Standard editing tasks (continued)

• 10: Every observable pose / view constraint is satisfied, including valid symmetry equivalence where applicable, with no discernible mismatch. Do not require a rotation history for insertion or replacement, or penalize unconstrained components.
4. Identity and Appearance Fidelity
This category assesses preservation of the edited object's identity, appearance and known details; Category 5 assesses preservation of protected scene details. Judge only details required to remain, excluding instructed changes and their necessary consequences.
Evaluates whether the output preserves or reproduces the identity and appearance attributes that the instruction requires. For an existing object being moved, resized, or reoriented, the source image usually supplies the identity evidence. For reference-guided addition or replacement, the designated appearance reference usually supplies that evidence.
Before judging identity, identify the visible features that distinguish the required instance from another object of the same category, such as body proportions, handle shape, windows, joints, logos, patterns or part count. Compare those features individually in the relevant views.
Matching the category and dominant color alone does not establish high identity fidelity when distinguishing features visibly differ. Preserve known source details even when the pose proxy omits them; do not confuse a legitimate occlusion with deletion. Do not require unobservable features or exact pixel agreement across lighting / view changes.
Assess instance-specific structure, distinctive components, material, texture, patterns, markings, and other visible identifying details. Sharing an object category is not sufficient to establish identity fidelity. When appearance and geometry are supplied separately, inheriting an irrelevant texture or reconstruction defect from the geometric proxy is an error.
Exclude explicitly requested appearance changes from preservation requirements. Allow necessary changes in illumination, perspective, and visibility, as well as instructed style adaptation. Do not require exact recovery of details that are not observable in the available references.
Detail fidelity when a pose / angle reference is used to reorient an existing source object: the source object's visible identity and detail remain authoritative unless appearance changes are explicitly requested. Use the angle reference to guide the target view, not to replace the source object's material, color, markings, component structure or fine features with a simpler proxy. A handle, wheel, seam, pattern or other source detail that should remain visible must not disappear merely because the angle reference omits or blurs it. Viewpoint changes can legitimately occlude details, expose new surfaces or change projected shape and illumination; these are not identity failures. For newly exposed regions, this category assesses only attributes verifiable from the identity sources; it does not score the plausibility of unspecified new detail, which belongs to Category 7 when applicable. A result that follows the angle but unnecessarily simplifies or swaps the original object fails detail / identity fidelity.
Scoring Guidelines:
• 0: The required object is absent, or its identity and required appearance are entirely incorrect.
• 1: Almost no recognizable identity or appearance relationship to the designated source is retained.
• 2: Only the broad object category is retained; the output represents a substantially different instance.
• 3: A few coarse identifying features remain, but most defining characteristics are lost.
• 4: The overall form is similar, but key components, materials, patterns, or distinctive details are substantially altered.
• 5: The intended identity is recognizable, with noticeable drift in several important appearance features.
• 6: The principal identity features are preserved, with localized detail loss or appearance inaccuracies.
• 7: Identity and key appearance attributes are faithfully retained, with only minor drift.
• 8: Distinctive structure, materials, textures, and visible markings are consistently reproduced within the permitted transformation.
• 9: Fine instance-specific details are also preserved with high fidelity, with negligible visible discrepancies.
• 10: Every verifiable identity and appearance requirement is satisfied without discernible distortion, excluding changes explicitly required by the edit.

Evaluate: Standard editing tasks (continued)

5. Non-target Content Preservation
This category assesses protected source-scene content and its details, including a sole object explicitly selected for preservation. Without protected source content, this category is not applicable, but Category 4 may still assess reference-object identity and details. Removal protects the surroundings, not the object that must disappear.
Evaluates whether content outside the intended editing scope, and any explicitly protected content within the scene, remains faithful to the source.
Assess both local preservation near the edit and distant protected content. A mostly unchanged background must not hide a changed protected face, logo, text, limb or other identity-critical detail; severity depends on what was altered and its importance to the instruction, not just its pixel area. Conversely, do not penalize a large area merely because it changed when the instruction authorizes that area to change. Keep authorized editing, protected-content drift and unavoidable integration effects distinct in the evidence.
Assess non-target objects, scene layout, background structures, text, texture, and global composition. The permissible editing scope must be determined from the task rather than expanded retrospectively to accommodate unintended changes in the output.
Allow changes required by the edit, including disocclusion, localized inpainting, contact shadows, reflections, and physically or stylistically necessary integration. For background replacement, assess protected foreground content rather than requiring the original background to remain. For newly generated scenes without a source scene to preserve, this category may be inapplicable; object identity remains assessable under Category 4.
- Preservation selection: verify that the designated content retains its required identity, shape, details and any protected position / scale while the requested changes occur elsewhere. For example, 'keep the boxed object and change the background' protects that object; editing or replacing it is a failure even though it lies inside a box. Do not require every background pixel inside a loose object-selection box to remain when that background is explicitly being edited. If the entire boxed region is explicitly protected, preserve the whole region. Protection of one region does not automatically permit arbitrary changes everywhere outside it.
- Exclusive preservation task: ONLY when the instruction explicitly identifies the boxed object as the sole content to retain and requires the rest of the scene to be changed / regenerated, treat that object (or an explicitly protected full region) as the preservation target for this category. The remaining content is the requested editing target, not automatically protected background. In this specific task, large changes outside the protected object are expected and must not lower preservation scores merely because the original scene, background or other objects no longer remain. Those changes must still follow the requested replacement scene and other constraints; this is not permission for arbitrary unrelated content.
- Resolve the box from the wording: 'keep only the boxed object and rebuild everything else' protects the selected object, not all pixels in its rectangular bounding box. Background visible inside a loose box, between limbs or through openings belongs to the editable surroundings unless explicitly protected. 'Keep the entire boxed region unchanged' instead protects all content in that region. Preserve the object's identity, shape, markings, texture and fine parts, plus position / scale / pose by default unless the instruction or designated controls require changes or those changes are necessary to complete the edit. Explicitly requested changes to those attributes and necessary integration are not preservation failures. Do not replace the protected instance with a merely similar object.
- This exception does not apply to ordinary object-selection boxes, destination boxes, or instructions that only request a limited background / local edit. A box by itself never authorizes changing everything else. Determine the protected scope before judging OUTPUT. If the instruction is ambiguous about exclusive retention, do not assume permission to regenerate the rest.
- Score preservation of the designated object / region even when it is the only original content retained; do not mark this category not_applicable merely because the surrounding scene is new. Conversely, leaving the surroundings unchanged when they were required to change is an Instruction Adherence failure, not proof of successful completion. Keep that failure separate from whether the protected object itself was faithfully retained.
Minimum necessary change:

Evaluate: Standard editing tasks (continued)

Limit modifications to the original editing instruction, its designated visual controls, and the changes necessary to realize that edit coherently. Content not included in the authorized editing scope is protected by default when a source scene exists; the instruction need not separately say 'keep everything else unchanged'. Use the instruction to establish the scope before inspecting the candidate, and do not expand it afterward to excuse unintended changes.
Necessary consequences may include local illumination adaptation, contact shadows, reflections, occlusion / disocclusion, edge blending, and completion of newly exposed regions. Each must have a concrete relationship to the requested operation and be no broader than needed. 'Better integration', 'more realistic' or 'more beautiful' is not blanket permission to regenerate unrelated areas. A local edit does not normally authorize global relighting, color grading, camera / composition changes, background replacement, identity changes, new decorations, or adding / removing other objects. Such changes are permitted only when requested or visibly necessary to the specified edit; ensure there is a clear logical necessity rather than assuming it.
Treat observable changes outside this allowed scope as unintended edits, even if they improve apparent aesthetics. Apply the shared preservation principle: assess spatial drift in Category 2, pose drift in Category 3, object identity / appearance drift in Category 4 and protected scene drift in Category 5, according to the independently violated requirements.
When evaluating a subtask, content changes authorized or requested by other subtasks in the same instruction must not be penalized as unintended edits under this category.
Scoring Guidelines:
• 0: Almost all content that should remain unchanged is destroyed or replaced.
• 1: Only minimal traces of the protected source content remain.
• 2: A large portion of the protected content is regenerated or substantially altered; the designated object / region loses major required structure or detail.
• 3: Several important protected elements are changed, removed or corrupted, including important parts of a sole retained object when applicable.
• 4: The protected object / region or other protected scene content remains recognizable, but substantial unintended modifications are present.
• 5: The main protected content is retained, with noticeable drift across several required details or protected areas.
• 6: Protected content is generally stable, with a limited number of visible localized changes.
• 7: Only small, nonessential deviations occur outside the intended editing scope.
• 8: The designated protected content, including layout where preservation is required, is faithfully retained with no material unintended change.
• 9: Fine structures, markings, text and textures within the protected content are also preserved with negligible variation.
• 10: All observable protected content is preserved without discernible unintended modification, allowing only changes necessary to perform the requested edit.
6. Image-Editing Quality
Assess the edited region's structure, boundaries, detail, lighting, perspective, occlusion, style and scene integration. For removal, focus on the repaired region; for scene generation, assess the generated scene and reference-object integration.
Evaluates the perceptual quality of the generated image and the coherence of the edited content with its intended scene and style.
Assess structural integrity, edge quality, texture clarity, lighting, shadows, reflections, perspective, occlusion, and stylistic compatibility. Penalize visible seams, halos, unintended duplication, malformed structures, and leaked control annotations or reference panels.

Evaluate: Standard editing tasks (continued)

Edited-versus-retained style consistency: compare the edited region with the surrounding retained content in rendering style (photographic, illustrative, cartoon or 3D-rendered), material depiction, color treatment, contrast, grain / noise, sharpness and detail density. Penalize an unrequested pasted-on appearance, a plastic-looking render in a photograph, incompatible brushwork, or abrupt style / texture transitions, even when the edit is semantically correct. Allow differences explicitly requested by the instruction, and normal differences due to material, depth of field or local illumination. For scene-wide restyling, assess consistency with the requested new style rather than requiring the old style to remain. Do not justify unauthorized changes to protected regions merely to make styles match. Score an independently visible mismatch in this category; do not automatically duplicate that penalty in unrelated categories.
Angle-reference rendering residue: whenever a pose / angle reference is provided, assess whether the edited object retains an unrequested, unnatural 3D-model or proxy-render appearance, regardless of whether other appearance references, a source scene or spatial-control images are also supplied. This includes a single reference that specifies both appearance and pose. Look for an oily sheen, waxy or plastic surfaces, excessively smooth shading, uniform artificial specular highlights, or simplified textures and missing material roughness that conflict with the intended material and scene style. Matching the reference angle does not excuse copying incidental proxy-render artifacts. Preserve specified object appearance while adapting its rendering to the intended visual domain. Judge visible material and style mismatches against the instruction, source scene and applicable appearance evidence; do not assume residue merely because the angle reference is a 3D render. Naturally glossy materials, reflections supported by the lighting, and explicitly requested or otherwise established 3D / stylized rendering are not defects by themselves. Score observable residue in Category 6 according to its severity and extent; conspicuous residue is incompatible with the artifact-free high-score criteria, even when angle alignment is correct. This check is independent of Category 7 applicability and remains required when a separate appearance reference is present.
Judge quality relative to the requested visual domain. Photorealism is not a universal requirement. An explicitly requested surreal or physically impossible configuration is not itself a defect; evaluate whether its visual realization is coherent. This category does not replace task-completion assessment: a high-quality unchanged image can receive a high quality score while failing the editing task.
Scoring Guidelines:
• 0: The available image is visually unintelligible or unusable because of pervasive visual defects.
• 1: Extreme artifacts or structural collapse dominate the image.
• 2: Severe deformation, compositing errors, or incoherent rendering make the result largely unusable.
• 3: Numerous obvious structural, boundary, or integration defects substantially impair quality.
• 4: The main content is recognizable, but prominent artifacts or scene inconsistencies remain.
• 5: The image is usable, with noticeable artifacts, imperfect edges, or inconsistent scene integration.
• 6: Overall quality is satisfactory, with several limited and visible defects.
• 7: Structure and integration are convincing, with only minor quality issues.
• 8: The image is clear, stylistically coherent, and naturally integrated, with no material visual defect.
• 9: Fine details, boundaries, illumination, and occlusion are highly coherent, with negligible artifacts.
• 10: No perceptible quality or integration defect is present at the supplied evaluation resolution and under the requested style.
7. Detail Completion Plausibility
Evaluate only whether genuinely unspecified, simplified or newly exposed object details are completed reasonably in cases that provide a single pose / angle reference without a separate appearance reference. This is a completion criterion, not a second identity-preservation score. A source scene and spatial-control images may also be present; 'single angle reference' describes the reference configuration, not the total number of input images. Determine applicability from the instruction and inputs before judging the output.
Applicability:

Evaluate: Standard editing tasks (continued)

- Apply this criterion when the only object reference is the pose / angle image and the requested result requires completion of object details that are not fully established by that image or an existing source object. This includes appropriate completion of newly visible surfaces after a view change, if those surfaces lack an appearance specification.
- Use null / not_applicable when a separate appearance reference is supplied, no pose / angle reference is supplied, or no object-detail completion is required. A single reference that already specifies all needed appearance and geometry does not automatically create a completion requirement. Original-object detail loss belongs to Category 4, even if the angle reference is simplified.
- Do not include background inpainting after removal, generic scene generation or ordinary edge blending in this category; those remain under their existing criteria. If the input roles or the existence of a completion requirement cannot be established, use null / unassessable. Plausibility can be assessed from the supplied constraints even when unspecified details admit multiple valid completions. A visibly missing required object or missing required completion is an observed failure when this criterion is applicable.

What to evaluate:

1. Completeness: necessary parts and surfaces are sensibly realized rather than left with accidental holes, unfinished patches, missing components or copied proxy reconstruction defects. A deliberately plain, abstract or stylized object need not be embellished.
2. Structural plausibility: newly completed parts connect and fit together coherently, with plausible local geometry and surface continuity relative to the supplied object and requested view. In a physics-violation task, preserve the deliberately specified impossible arrangement; do not replace it with a different 'realistic' solution.
3. Appearance coherence of the completed regions: newly supplied material, texture and detail are compatible with the instruction, visible object cues and intended style. Do not invent salient logos, patterns, decorative accessories or identity-defining additions simply to make the result richer.
4. Minimum necessary completion: fill only genuinely unspecified / newly exposed content needed for the requested result. Do not reward more texture or detail by default. Accept different valid realizations consistent with the supplied constraints; do not claim recovery of unobserved detail.

Boundary with other metrics: Category 3 judges target-angle alignment; Category 4 judges fidelity of known identity / appearance / source details; Category 6 judges overall rendering and scene integration. This category judges the plausibility and sufficiency of the newly completed content. Do not repeat an angle error, an existing-detail loss or a generic image-quality defect here unless it causes a separately identifiable completion defect. For example, changing a known source logo belongs to Category 4; an implausibly connected newly completed part belongs here. An accurate angle cannot compensate for incomplete or unreasonable completion, and convincing completion cannot establish angle success.

Scoring Guidelines:

- 0: The required object is absent, or necessary completion wholly fails and no meaningful completed object is present.
- 1: Required completion is almost entirely missing or uninterpretable.
- 2: Severe missing parts, copied proxy defects or implausible completions dominate the content that needed completion.
- 3: Some content is completed, but several major gaps or structurally incoherent parts remain.
- 4: Completion is partly reasonable, but a critical missing part, incoherent connection or materially unsupported addition remains.
- 5: The necessary completion is broadly recognizable, with conspicuous structural or appearance inconsistencies.
- 6: The main missing content is completed coherently, with limited localized deficiencies or inconsistencies.
- 7: Necessary completion is reasonable and consistent, with only minor imperfections.
- 8: All observable completion requirements are met with coherent structure / appearance and no material unsupported embellishment.
- 9: Completed content is finely coherent with the available cues, with only negligible observable defects.

Evaluate: Standard editing tasks (continued)

• 10: All necessary completion is convincing and consistent with supplied constraints, with no discernible defect or unjustified addition.
Assess the angle-reference image, the content that requires completion, and visible completion plausibility or defects. Do not use preservation of already-known details as the basis of this score. An independently visible rendering defect is assessed in Category 6. Failure of essential required completion also affects Category 1 when it prevents fulfillment of the instruction. Added unknown detail does not earn preservation credit in Categories 4–5 merely for being detailed.
GENERAL SCORING RULES
- Determine applicability from the instruction and inputs, not from whether the output succeeds. A missing required object is a scored failure; successful removal is not an identity or pose failure.
- Use scored when an applicable criterion can be judged, not_applicable when no corresponding requirement exists, and unassessable when the evidence cannot support a score. Local uncertainty affects only the criteria that cannot be judged; score the rest normally.
- If the task cannot be established because essential inputs are missing or contradictory, or the output is absent or unreadable, all scores are unassessable. A readable blank, unrelated or refusal image is assessed as a visible failure.
- Judge only supplied visual evidence. Do not infer quality from model identity or other candidates, invent invisible details, or fabricate quantitative measurements.
- Apply the same score anchors independently to all seven criteria. An error affects a criterion only when it violates that criterion's requirements. Scores do not numerically constrain one another; ties are allowed, without forced rankings or differences.
- An omitted or absent required object must be scored as 0 under applicable categories, and must not be marked as unassessable or not_applicable.
OUTPUT FORMAT
Return only valid JSON using the structure below, containing the operation task code, seven scores and applicability statuses. Do not include explanations, evidence or other commentary. All scores are integers 0–10 or null. Each status is one of scored, not_applicable or unassessable; the latter two use null, not zero. Score each criterion independently; do not output totals or averages.
operation must be a code from TASK TAXONOMY; replace the illustrative 1d_add with the actual category.
operation classifies the task as a whole; do not decompose it.

```
{
  "operation": "1d_add",
  "diagnostic_scores": {
    "Instruction Adherence": {
      "status": "scored | not_applicable | unassessable",
      "score": null
    },
    "Spatial Control Accuracy": {
      "status": "scored | not_applicable | unassessable",
      "score": null
    },
    "Pose and View Alignment": {
      "status": "scored | not_applicable | unassessable",
      "score": null
    },
    "Identity and Appearance Fidelity": {
      "status": "scored | not_applicable | unassessable",
      "score": null
    },
    "Non-target Content Preservation": {
      "status": "scored | not_applicable | unassessable",
      "score": null
    },
```

Evaluate: Standard editing tasks (continued)

```
  "Image-Editing Quality": {
    "status": "scored | not_applicable | unassessable",
    "score": null
  },
  "Detail Completion Plausibility": {
    "status": "scored | not_applicable | unassessable",
    "score": null
  }
 }
}
```

Evaluate: Composite-task extension (shared criteria above)

First decompose the editing instruction into independently assessable operations, numbered T1, T2, etc. Apply the following criteria to each operation and return its scores in subtask_results. Classify each operation using a task code below according to that operation and its relevant control images; do not copy the full instruction's composite-editing label onto each subtask. Operations of the same type on different objects may remain separate entries with the same code. Treat appearance, pose and box requirements attached to an operation as its constraints. Edits explicitly requested by other operations are not unintended changes. Do not output combined scores.

OUTPUT FORMAT
Return only valid JSON using the structure below, containing the operation task code, seven scores and applicability statuses. Do not include explanations, evidence or other commentary. All scores are integers 0–10 or null. Each status is one of scored, not_applicable or unassessable; the latter two use null, not zero. Score each criterion independently; do not output totals or averages.
operation must be a code from TASK TAXONOMY; replace the illustrative 1d_add with the actual category.
Keep subtask_id and operation for each subtask; codes may repeat. Include all decomposed operations, including failed ones, without combined scores.

```
{
 "subtask_results": [
  {
   "subtask_id": "T1",
   "operation": "1d_add",
   "evaluation": {
    "diagnostic_scores": {
     "Instruction Adherence": {
      "status": "scored | not_applicable | unassessable",
      "score": null
     },
     "Spatial Control Accuracy": {
      "status": "scored | not_applicable | unassessable",
      "score": null
     },
     "Pose and View Alignment": {
      "status": "scored | not_applicable | unassessable",
      "score": null
     },
     "Identity and Appearance Fidelity": {
      "status": "scored | not_applicable | unassessable",
      "score": null
     },
     "Non-target Content Preservation": {
      "status": "scored | not_applicable | unassessable",
```

Evaluate: Composite-task extension (shared criteria above) (continued)

```
          "score": null
        },
        "Image-Editing Quality": {
          "status": "scored | not_applicable | unassessable",
          "score": null
        },
        "Detail Completion Plausibility": {
          "status": "scored | not_applicable | unassessable",
          "score": null
        }
      }
    }
  }
 ]
}
```