UNIEDITBENCH: A UNIFIED AND COST-EFFECTIVE BENCHMARK FOR IMAGE AND VIDEO EDITING VIA DISTILLED MLLMS

Anonymous authors
Paper under double-blind review

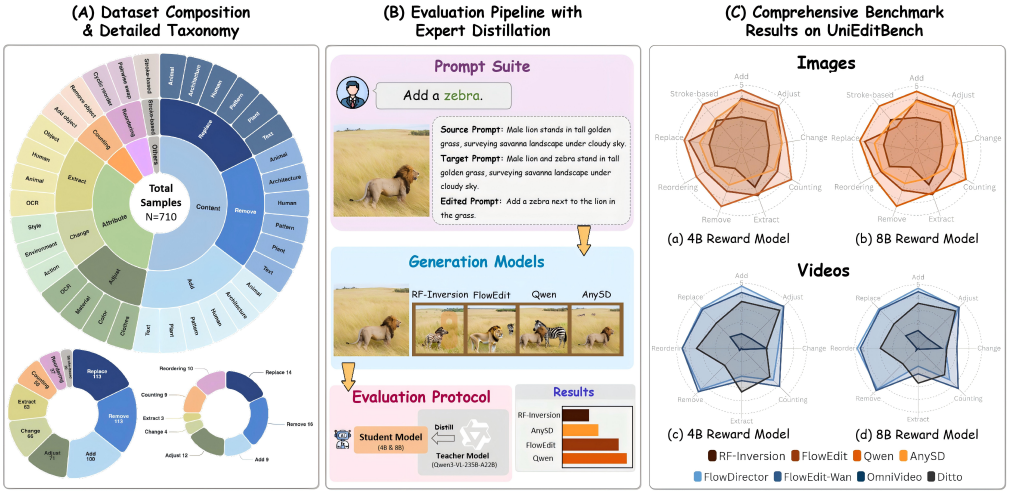

Figure 1: UniEditBench overview: task taxonomy and 710 source instances (633 images, 77 videos), the prompt-triplet scoring interface, and selected editor profiles under the 4B/8B evaluators. The judges produce absolute rubric scores.

ABSTRACT

The evaluation of visual editing models remains fragmented across methods and modalities. Existing benchmarks are often tailored to specific paradigms, making fair cross-paradigm comparisons difficult, while video editing lacks reliable evaluation benchmarks. Furthermore, common automatic metrics often misalign with human preference, yet directly deploying large multimodal models (MLLMs) as evaluators incurs prohibitive computational and financial costs. We present **UniEditBench**, a unified benchmark for image and video editing that supports reconstruction-based and instruction-driven methods under a shared protocol. UniEditBench includes a structured taxonomy of nine image operations (Add, Remove, Replace, Change, Stroke-based, Extract, Adjust, Count, Reorder) and eight video operations, with coverage of challenging compositional tasks such as counting and spatial reordering. To enable scalable evaluation, we distill a high-capacity MLLM judge (Qwen3-VL-235B-A22B Instruct) into lightweight 4B/8B evaluators that provide multi-dimensional scoring over structural fidelity, text alignment, background consistency, naturalness, and temporal-spatial consistency (for videos). Experiments show that the distilled evaluators maintain strong agreement with human judgments and substantially reduce deployment cost relative to the teacher model. UniEditBench provides a practical and reproducible protocol for benchmarking modern visual editing methods.

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

1 INTRODUCTION

Recent advances in visual generative models—particularly diffusion-based frameworks Ho et al. (2020); Rombach et al. (2022) and MLLMs Liu et al. (2023); Bai et al. (2025b)—have substantially improved visual content creation. Research has shifted from synthesis toward user-controllable, fine-grained editing, including object replacement, attribute manipulation, and motion-aware video editing Hertz et al. (2022); Brooks et al. (2023); Zhang et al. (2023), making visual editing increasingly important for interactive multimedia systems.

Despite rapid progress, evaluation remains fragmented across method families. We identify three practical gaps. **(1) Paradigm fragmentation.** Existing benchmarks typically target either reconstruction-based methods, including inversion-based Couairon et al. (2022); Rout et al. (2024) and inversion-free approaches Kulikov et al. (2025); Jiang et al. (2026), or instruction-driven methods Brooks et al. (2023); Wu et al. (2025a), limiting fair cross-paradigm comparison. **(2) Limited video coverage.** Standardized video editing benchmarks remain scarce despite the importance of temporal coherence and inter-frame consistency Liu et al. (2024); Chen et al. (2024). **(3) Evaluator cost–fidelity tension.** Efficient metrics such as PSNR and CLIP-based scores Radford et al. (2021) may misalign with human preference, whereas very large MLLM evaluators increase both computational requirements and the cost of repeated evaluation.

To address these gaps, we present UniEditBench, a unified benchmark for image and video editing under a shared protocol. It supports reconstruction-based and instruction-driven methods through nine image operations (Add, Remove, Replace, Change, Stroke-based, Extract, Adjust, Count, Reorder) and eight video operations, including reasoning-intensive tasks such as counting and spatial reordering. UniEditBench further standardizes heterogeneous model inputs using visually grounded prompt triplets comprising a source prompt, target prompt, and editing instruction. As shown in Figure 2, the pipeline integrates multi-source data aggregation, prompt unification, quality filtering, and evaluator distillation into a consistent workflow.

To enable scalable and interpretable scoring, we distill the visual assessment capability of Qwen3-VL-235B-A22B into 4B/8B evaluators initialized from Qwen3-VL-4B and Qwen3-VL-8B. The students assess structural fidelity, text alignment, background consistency, naturalness, and temporal-spatial consistency for videos. Two-stage image/video training reduces disagreement with the teacher, and a blind user study supplies complementary human score profiles. The students require substantially less weight storage, supporting repeated evaluation at a smaller deployment scale. Teacher agreement and human validity are assessed separately.

Our contributions are summarized as follows:

- **Unified benchmark across paradigms and modalities.** We introduce UniEditBench, which evaluates both image and video editing under a shared fine-grained taxonomy and protocol, connecting previously isolated method families while retaining modality-specific criteria.
- **Benchmark construction with standardized prompt interfaces.** We build a multi-source, quality-controlled dataset and unify heterogeneous model inputs with visually grounded prompt triplets, improving evaluation consistency across different editing paradigms.
- **Cost-aware distilled evaluators with interpretable scoring.** We propose a teacher-student evaluation framework that distills a high-capacity MLLM judge into lightweight 4B/8B models and performs dimension-specific assessment, with teacher-agreement, human-profile, and deployment evidence reported separately.

2 RELATED WORK

2.1 VISUAL GENERATION AND EDITING

Recent advances in visual content generation have been driven by diffusion models (Ho et al., 2020; Rombach et al., 2022) and multimodal LLMs (MLLMs) (Liu et al., 2023; Bai et al., 2025b). Research has shifted toward controllable visual editing, such as object manipulation, attribute editing, and video transformation (Hertz et al., 2022; Brooks et al., 2023; Zhang et al., 2023). Existing methods fall into two paradigms: reconstruction-based methods (inversion-based (Couairon et al., 2022;

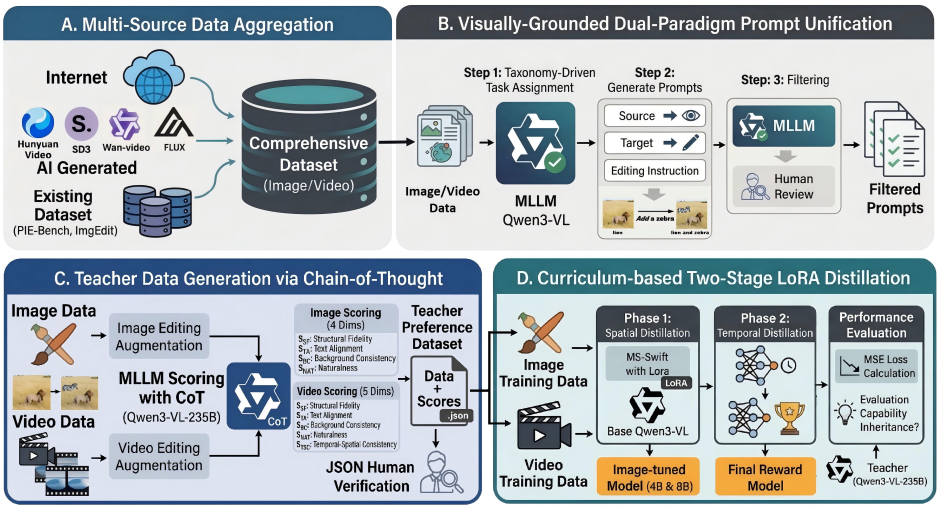

Figure 2: UniEditBench construction and evaluator training. Source assets are paired with prompt triplets, editing outputs receive teacher scores and rationales, and LoRA training proceeds from images to videos. The scoring dataset contains absolute ratings; the diagram's preference-data label does not denote pairwise comparisons.

Rout et al., 2024) and inversion-free (Kulikov et al., 2025; Jiang et al., 2026)), which emphasize structural preservation, and instruction-driven methods (Brooks et al., 2023; Liu et al., 2025; Deng et al., 2025; Wu et al., 2025a), which leverage natural language for flexible semantic edits. These paradigms complicate direct comparison, motivating unified evaluation.

2.2 Evaluation Benchmarks for Visual Editing

EditVal and EditBench evaluate text-guided edits (Basu et al., 2023; Lin et al., 2024). I2EBench studies multidimensional instruction-based image evaluation, and HATIE addresses scalable human-aligned text-guided evaluation (Ma et al., 2024; Ryu et al., 2025). Video protocols include TGVE, FiVE-Bench, and IVEBench (Wu et al., 2023; Li et al., 2025b; Chen et al., 2026). In particular, IVEBench evaluates temporal content through a dedicated video suite. UniEditBench complements this work with one instance interface across modalities and distilled local judges. Its 77-video subset is a compact cross-modality diagnostic, not a replacement for broader dedicated video benchmarks. This focus separates two benchmark design choices: which transformations are represented and how their outputs are assessed. Sharing operation names supports comparison of related tasks, while a video-specific temporal criterion retains information that an image-only rubric cannot express. The prompt-triplet interface similarly connects input conventions without requiring all editing architectures to use the same conditioning mechanism.

2.3 MLLM-based Evaluation and Distillation

Recently, large language models serve as automated evaluators (LLM-as-a-Judge) (Gu et al., 2024), demonstrating strong alignment with human judgments. In the visual domain, MLLMs enhance this by capturing fine-grained semantic and perceptual attributes beyond traditional metrics. However, directly employing large-scale MLLMs introduces practical limitations. Proprietary APIs incur substantial financial costs and privacy concerns, while deploying large open-source models like Qwen3-VL-235B-A22B (Bai et al., 2025b) requires significant computational resources. Knowledge distillation (KD) (Gu et al., 2023; Hsieh et al., 2023) offers a principled approach to transferring capabilities from large teacher models to smaller student models. Building on this idea, we develop a distilled MLLM-based evaluation framework that transfers capabilities from a high-capacity model to lightweight variants (4B/8B), enabling scalable, efficient evaluation for visual editing. Here, distillation transfers an evaluation interface as well as numerical responses: the students retain named dimensions and explanatory text. We therefore report teacher-score error separately from editor

quality and human profiles, so deployment benefits can be assessed while preserving the different roles of agreement, quality assessment, and practical resource requirements.

3 UNIEDITBENCH

We next describe benchmark construction and prompt unification (Section 3.1), taxonomy and dataset composition (Section 3.2), multi-dimensional metrics (Section 3.3), and the two-stage distilled evaluator (Section 3.4).

3.1 BENCHMARK CONSTRUCTION METHODOLOGY

UniEditBench aggregates diverse, high-fidelity data from three sources (Figure 2A). We first sample benchmarks including PIE-Bench and ImgEdit (Ye et al., 2025) for backward comparability, and crawl high-resolution web images and videos to capture real-world motion, lighting, and background complexity. We construct taxonomy-driven synthetic data: GPT-5.4 generates and filters 20 prompts per category, while FLUX (Labs, 2024) and SD3 (Esser et al., 2024) synthesize images, and Wan (Wan et al., 2025) and HunyuanVideo (Kong et al., 2024) synthesize videos. This strategy improves diversity while maintaining category coverage.

Visually-Grounded Dual-Paradigm Prompt Unification: To bridge reconstruction-based and instruction-driven evaluation, we use an MLLM to inspect each source, assign a taxonomy label, and generate a prompt triplet comprising a source prompt, target prompt, and editing instruction (Figure 2B). The triplet provides a common interface across paradigms while making the intended transformation explicit. Semantic equivalence remains a property to inspect, rather than a consequence of converting text formats. We reduce labeling noise through secondary MLLM verification, rejection of ambiguous samples, and final expert review.

Rigorous Quality Control: The candidate pool undergoes a strict two-stage purification process (Figure 2B). First, coarse-grained MLLM filtering removes samples with severe artifacts, watermarks, or poor text-media alignment. Second, expert annotators manually cross-validate the remaining assets to confirm edit feasibility, resolve prompt ambiguity, and rigorously verify temporal consistency for videos. We retain only samples that satisfy both semantic correctness and stringent visual quality criteria. This process supplies the benchmark with feasible editing requests and usable source assets, so that subsequent scores concern the requested transformation and its visible effects.

The editing instance as the unit of evaluation. Write an instance as $z = (x, p_{\mathrm{src}}, p_{\mathrm{tgt}}, u, c)$, where x is the source media, u is the requested edit, and c is its operation label. The source and target descriptions express the starting and desired states of the same instance. An editor produces y; the evaluator receives $(x, p_{\mathrm{src}}, p_{\mathrm{tgt}}, y)$, as specified by the templates. This notation separates the task record from any particular editor's input adapter. For example, relocating a chair is one task even when one editor consumes two scene descriptions and another consumes an imperative sentence. Additional spatial controls remain part of the editor configuration; their accounting is detailed in Appendix G.

This separation also preserves the meaning of the source count. The 710 source instances identify benchmark tasks, while multiple editors can generate different results for each source. Scores index those results and their dimensions. Thus neither the number of editor outputs nor the number of rubric entries should be confused with the number of independently sourced scenes.

3.2 TAXONOMY AND DATASET COMPOSITION

To evaluate visual editing models systematically, UniEditBench constructs a granular taxonomy. Image editing tasks are categorized into 9 operations: Add, Remove, Replace, Change, Stroke-based, Extract, Adjust, Count, and Reorder. To accommodate temporal dynamics, video editing is adapted into 8 operations, excluding Stroke-based editing.

Beyond Appearance: The Challenge of Spatial Reasoning. Traditional benchmarks emphasize appearance changes or simple entity swaps, leaving compositional reasoning underexplored. UniEditBench organizes its taxonomy into content modification (Add, Remove, Replace), attribute transformation (Change, Adjust, Extract), and compositional reasoning (Count, Reorder). Count requires

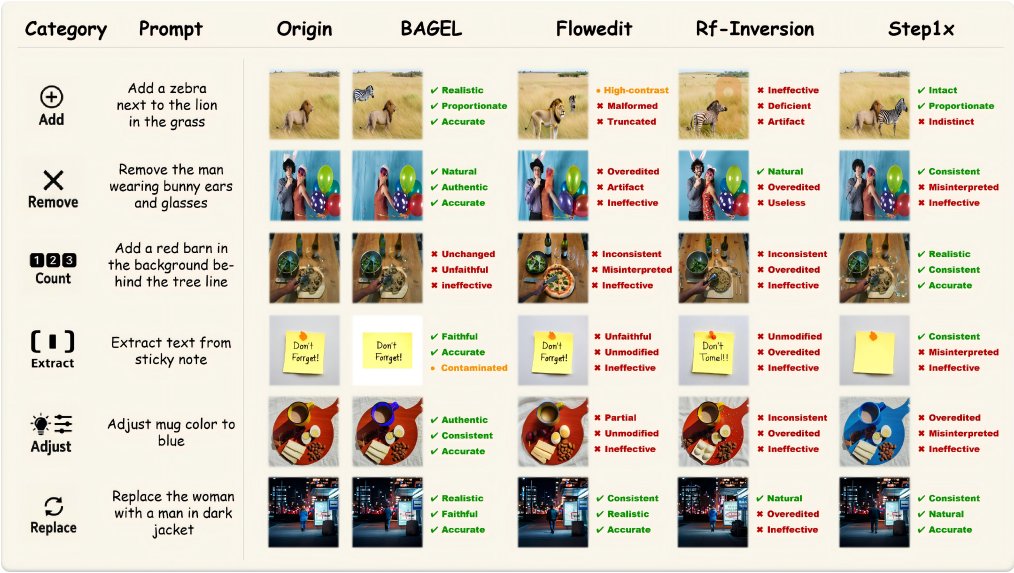

Figure 3: Visual examples from the editing benchmark. The instruction "Add a red barn in the background" illustrates Add; the panel's Count heading does not contain an explicit counting constraint. Additional examples appear in Appendix C.

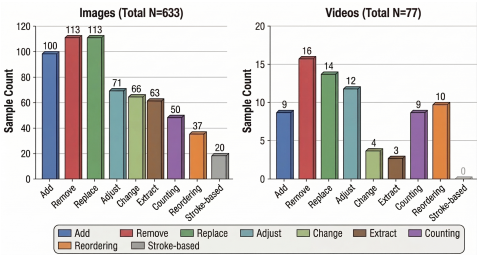

Figure 4: UniEditBench source counts by editing category and modality.

The distribution makes the benchmark's coverage explicit: an operation label identifies the requested transformation, while its count indicates how many source instances support that part of the evaluation. Image and video counts are reported separately because the modalities have different sample sizes and applicable operations. In particular, Stroke-based editing belongs to the image taxonomy. The figure complements the dimension-level result tables by showing the task composition behind their summaries. Category coverage and visual diversity describe different aspects of the resource: one operation can involve diverse objects and scenes.

changing object quantities while preserving unedited regions, whereas Reorder requires relocating objects while maintaining object identity and scene consistency. These tasks expose explicit spatial and multiplicity constraints, although task-level teacher MSE measures agreement with the teacher and does not directly assess an editor's reasoning process.

Dataset Statistics and Domain Diversity. The UniEditBench dataset comprises 633 source images and 77 source videos (710 prompt triplets in total). Figure 4 reports the category counts for both modalities. For coverage, the visual assets are curated for broad domain diversity, encompassing a wide spectrum of artistic styles (realistic photography, 2D anime, 3D rendering, oil painting) and rich semantic scenes (portraits, animals, landscapes, urban environments).

3.3 Dimension-Specific Scoring

Each image receives four integer scores in $\{0, 1, \ldots, 5\}$; video adds a fifth score. These are conceptually distinct rubric dimensions, not statistically independent variables.

- **Structural fidelity** (S_{SF})**:** preservation of geometry and identity outside the requested change. Intentional relocation is part of the requested transformation, rather than an error in preserving the original layout.

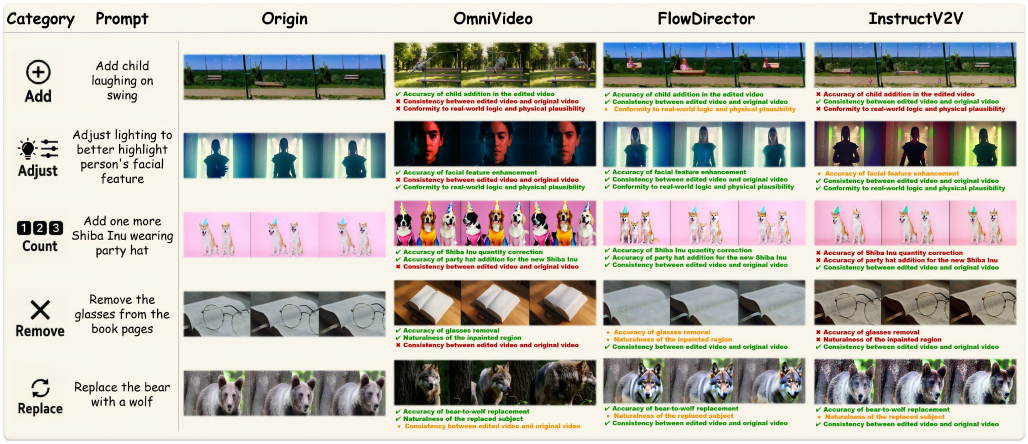

Figure 5: Visual results of selected video editing methods on UniEditBench. Appendix C supplies additional outputs spanning the benchmark's editing operations.

- **Text alignment (S_{TA}):** whether the requested transformation is applied to the intended content, including quantities and spatial relations.
- **Background consistency (S_{BC}):** preservation of context that the instruction leaves unchanged. It does not require unchanged pixels inside an intended global edit.
- **Naturalness (S_{NAT}):** visual plausibility, coherent lighting, and absence of generative artifacts, interpreted relative to the source's visual style.
- **Temporal-spatial consistency (S_{TSC}):** for video, coherent identity and trajectories across frames, without flicker or discontinuous changes.

The Overall column is the unweighted mean of the applicable dimensions; image and video means therefore have different denominators and are not directly comparable measures of task difficulty. Appendix F formalizes aggregation and distinguishes score-level MSE from editor performance.

A score vector before an aggregate. For an image result, the evaluation response is summarized by

$$\mathbf{s}_i = (s_{i,SF}, s_{i,TA}, s_{i,BC}, s_{i,NAT}), \qquad o_i = \tfrac{1}{4}\mathbf{1}^\top \mathbf{s}_i. \tag{1}$$

The video vector appends $s_{i,TSC}$ and averages five components. The vector is the diagnostic output; o_i is its compact summary. Equal averages can arise from different allocations across adherence, preservation, and plausibility. Consequently, a method comparison retains the dimension columns even when an Overall column is available. This representation makes the benchmark suitable for inspecting trade-offs rather than requiring every use case to value an identical failure pattern.

Extending the rubric through time. For video, the requested transformation concerns a clip rather than a collection of unrelated images. An edited object's identity and relation to the scene should remain coherent as viewpoint, pose, or visibility changes. Temporal-spatial consistency therefore asks whether the intended content persists without abrupt or contradictory changes. It complements text alignment: one can satisfy an instruction in a selected frame yet fail to maintain it across the clip. Conversely, temporal smoothness does not establish whether the edited clip depicts the transformation specified by the instruction.

The temporal dimension also differs from preserving every frame unchanged. A motion or relocation request may intentionally alter trajectories; the rubric concerns whether that alteration is coherent with the instruction. The spatial dimensions retain their original roles within this interpretation. Naturalness addresses plausible appearance, structural fidelity concerns unaffected identities and geometry, and background consistency concerns context outside the requested transformation. Keeping these responsibilities explicit allows the image and video interfaces to share a vocabulary while respecting the additional temporal evidence required for video.

3.4 COST-EFFECTIVE EVALUATOR VIA DISTILLATION

Large multimodal evaluators require substantial weight storage and runtime resources. We use Qwen3-VL-235B-A22B as a teacher and initialize students from Qwen3-VL-4B and Qwen3-VL-8B. Distillation makes the shared scoring interface available in smaller models for local deployment, with model storage and measured teacher agreement reported separately to support evaluator selection.

Teacher scoring data. Image outputs are generated with Bagel, Step1X-Edit, Qwen-Image-Edit, RF-Inversion, and FlowEdit (Deng et al., 2025; Liu et al., 2025; Wu et al., 2025a; Rout et al., 2024; Kulikov et al., 2025). Video outputs use FlowEdit-Wan, FlowDirector, InstructV2V, and Ditto (Li et al., 2025b;a; Qin et al., 2024; Bai et al., 2025a). The teacher produces a dimension-specific rationale and absolute score for each output, using the templates in Appendix B. We refer to these as scoring data, since there are no pairwise preference labels in this construction.

Two-stage LoRA training. Using ms-swift (Zhao et al., 2025), students are fine-tuned on image evaluations for three epochs and then on video evaluations for three epochs, initialized from the image-stage weights. The response format contains both rationale text and scores. We evaluate the resulting teacher imitation by

$$\mathrm{MSE} = \frac{1}{N|\mathcal{D}|} \sum_{i=1}^{N} \sum_{d \in \mathcal{D}} (s_{i,d}^{\mathrm{student}} - s_{i,d}^{\mathrm{teacher}})^2, \tag{2}$$

where $\mathcal{D}$ is the applicable dimension set for one modality. MSE measures agreement between the numerical scores extracted from teacher and student responses. Training uses the structured response containing both rationale and scores; the stage comparison then measures how score agreement changes after image supervision and the subsequent addition of video supervision.

The scoring response has two levels of organization. Within an output, named fields retain the rubric dimensions so that adherence is not confused with visual preservation. Across outputs, the same fields permit operation-level and modality-level summaries. The distilled interface preserves this organization: a user can inspect a rationale for one edited result and then relate its scores to the corresponding model profile. This is especially useful when a model's aggregate score is driven by a strong background score but weak adherence on a particular operation.

What is transferred. The student learns to produce a structured evaluation response, comprising the explanation and the dimension scores. The rationale connects a numerical judgment to a visible property of the edit, such as a requested object being absent or an unrelated region changing appearance. The numerical fields then support aggregation across outputs. Maintaining both parts preserves the teacher's evaluation interface rather than reducing each result to one undifferentiated reward. Equation 2 measures score transfer, while the accompanying rationale keeps the stated basis of each judgment available for inspection.

The image stage supplies the four shared spatial criteria. Continuing from those weights with video examples introduces evidence about persistence, motion, and identity across frames while retaining the common dimension vocabulary. This sequence connects spatial assessment to its temporal extension. The stage comparison measures realized teacher agreement, and the 4B and 8B models provide separate deployment choices. Their shared output fields allow users to compare modality-specific errors and select the evaluator whose measured profile best fits the intended editing workload.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

We evaluate 25 image and eight video editors with the 4B/8B judges. Image scores cover structural fidelity, text alignment, background consistency, and naturalness; video adds temporal-spatial consistency. Tables report seed-averaged 4B/8B scores on a 0–5 scale (Appendix E). Overall averages the applicable dimensions; teacher-score MSE (Equation 2) measures imitation. A blind study of 50 participants in five groups supplies complementary human profiles. Appendices G and I detail input conditions and human evaluation.

Table 1: Quantitative Benchmarking on **Image Editing**. Scores (0-5) are evaluated by our distilled SFT Image+Video evaluators. Each cell presents assessment results in a 4B/8B dual-entry format. Best results within each paradigm are **bolded**, and second-best results are underlined. The table highlights why unified, multi-dimensional evaluation is necessary: different paradigms excel on different dimensions, and no single method dominates all criteria simultaneously.

Paradigm	Model	Structural (S_{SF})	Alignment (S_{TA})	Background (S_{BC})	Naturalness (S_{NAT})	Overall
Inversion-based	SDEditMeng et al. (2021)	1.74/2.12	2.60/2.78	1.49/1.70	1.72/1.99	1.89/2.15
	RF_InversionRout et al. (2024)	2.69/2.64	2.11/2.09	2.36/2.32	2.33/2.27	2.37/2.33
	RF_SolverWang et al. (2024)	3.26/3.29	**3.33/3.32**	2.52/2.61	2.61/2.69	2.93/2.98
	PNP_InversionJu et al. (2023)	3.63/3.68	2.51/2.50	3.10/3.17	2.70/2.82	2.99/3.04
	MasactrlCao et al. (2023)	**3.82/3.86**	2.37/2.33	**3.67/3.72**	**2.95/3.04**	**3.20/3.24**
Inversion-free	FloweditKulikov et al. (2025)	3.84/3.79	**3.36/3.33**	3.96/3.92	3.38/3.36	3.63/3.60
	FIA-EditYang et al. (2025)	3.61/3.65	3.18/3.13	3.37/3.50	2.98/3.06	3.28/3.33
	DNAEditXie et al. (2025)	**4.40/4.40**	2.64/2.58	**4.37/4.41**	**3.76/3.79**	**3.79/3.80**
Structured editing	InstructPix2PixBrooks et al. (2023)	1.65/1.74	1.65/1.79	1.34/1.39	1.23/1.29	1.47/1.55
	Step1X-EditLiu et al. (2025)	**4.53/4.47**	**3.88/3.87**	**4.58/4.58**	**4.01/3.94**	**4.25/4.21**
	ICEditZhang et al. (2025)	4.29/4.25	3.08/3.01	4.26/4.21	3.60/3.58	3.81/3.76
	control_v11e_sd15_ip2pZhang et al. (2023)	1.64/1.62	1.66/1.80	1.18/1.21	1.07/1.09	1.39/1.43
MLLM	HunyuanCao et al. (2025)	**4.62/4.69**	**4.63/4.61**	4.10/4.16	**4.13/4.15**	**4.37/4.40**
	Qwen-Image-editWu et al. (2025a)	4.54/4.53	4.36/4.35	4.17/4.18	4.09/4.09	4.29/4.29
	Bagel_with_thinkingDeng et al. (2025)	4.20/4.14	3.64/3.64	4.29/4.34	3.52/3.52	3.91/3.91
	Bagel_without_thinkingDeng et al. (2025)	4.17/4.17	3.71/3.70	**4.37/4.38**	3.60/3.58	3.96/3.96
	UniWorldLi et al. (2025d)	4.24/4.30	3.85/3.85	3.98/4.13	3.69/3.78	3.94/4.02
	Omni-Gen2Wu et al. (2025b)	4.39/4.38	3.65/3.60	4.07/4.17	3.71/3.79	3.96/3.99
	DreamOmni2Xia et al. (2025)	4.53/4.56	3.54/3.48	4.33/4.33	4.04/4.02	4.11/4.10
Masked Diff/AR	EditMGTChow et al. (2025)	2.59/2.61	2.67/2.75	2.83/2.91	1.89/1.94	2.49/2.55
	LaVida-OLi et al. (2025c)	**3.85/3.89**	**3.54/3.59**	3.49/3.55	2.88/2.96	3.44/3.50
	EditARMu et al. (2025)	3.29/3.37	2.56/2.55	3.03/3.07	2.50/2.60	2.85/2.90
	VAREditMao et al. (2025)	3.84/3.82	3.53/3.55	**4.21/4.29**	**3.50/3.51**	**3.77/3.79**
Others	AnyEditYu et al. (2025)	3.61/3.62	2.46/2.43	3.80/3.82	2.81/2.82	3.17/3.17
	FLUX.1-Kontext-devLabs et al. (2025)	**4.62/4.63**	**3.74/3.71**	**4.63/4.64**	**4.25/4.22**	**4.31/4.30**

Table 2: Quantitative Benchmarking on **Video Editing**. Scores (0-5) are evaluated by our distilled SFT Image+Video evaluators. Each cell presents assessment results in a 4B/8B dual-entry format. Best results within each paradigm are **bolded**, and second-best results are underlined. The results emphasize our motivation that video editing must be evaluated jointly on instruction following and temporal stability, rather than by single-aspect quality alone.

Paradigm	Model	Structural (S_{SF})	Alignment (S_{TA})	Background (S_{BC})	Naturalness (S_{NAT})	Temporal (S_{TSC})	Overall
Inversion-free	FlowDirectorLi et al. (2025a)	3.96/4.10	**3.87/3.87**	4.68/4.57	4.13/4.04	4.39/4.35	4.21/4.19
	FlowEdit-PyramidLi et al. (2025b)	3.77/3.56	3.55/3.42	4.04/3.86	3.45/3.19	3.69/3.39	3.70/3.48
	FlowEdit-WanLi et al. (2025b)	**4.25/4.40**	3.30/3.31	**4.74/4.74**	**4.47/4.48**	**4.66/4.69**	**4.28/4.32**
Structured editing	ICVELiao et al. (2025)	3.97/3.90	**4.00/3.94**	**4.45/4.43**	**3.70/3.70**	**4.08/4.00**	**4.04/3.99**
	InstructV2VQin et al. (2024)	**4.01/4.00**	3.38/3.31	3.90/3.84	3.55/3.61	3.86/3.90	3.74/3.73
MLLM	OmniVideoYang et al. (2026)	0.40/0.40	1.29/1.29	0.45/0.60	0.83/0.86	0.57/0.61	0.71/0.75
	DittoBai et al. (2025a)	**2.99/2.94**	**3.19/3.21**	**2.97/2.75**	**2.78/2.56**	**2.97/2.84**	**2.98/2.86**
Others	Lucy-Edit	3.42/3.44	3.17/3.14	4.05/4.09	3.30/3.27	3.51/3.52	3.49/3.49

Table 3: Mean squared error (MSE) against Qwen3-VL-235B-A22B, grouped by editing operation. The operation columns follow the reported weighted aggregation; the last two columns report modality-level averages. Lower is better. Final models improve over zero-shot; gains vary across stages and operations. Appendix A gives modality-specific results.

Model	Add	Adjust	Change	Count	Extract	Remove	Reorder	Replace	Stroke-based	Ovr. (Image)	Ovr. (Video)
Zero-shot 4B	2.16	2.27	3.71	3.87	5.98	3.62	3.73	2.47	1.47	2.95	5.40
Zero-shot 8B	1.50	1.58	2.73	2.83	5.24	3.25	3.84	2.03	2.11	2.51	4.00
SFT Image 4B	0.35	0.46	0.45	0.76	1.62	1.12	0.95	0.43	0.32	0.50	2.33
SFT Image 8B	0.27	0.31	0.56	0.60	1.15	0.93	0.87	0.40	0.31	0.47	1.56
SFT Image+Video 4B	**0.27**	**0.25**	**0.39**	**0.48**	**1.30**	**0.72**	**0.67**	**0.36**	**0.31**	**0.49**	**0.73**
SFT Image+Video 8B	**0.23**	**0.26**	**0.40**	**0.46**	**1.24**	**0.78**	**0.61**	**0.31**	**0.32**	**0.47**	**0.79**

4.2 Main Results and Analysis

Editor profiles. Tables 1 and 2 expose different trade-offs between adherence and preservation. On images, Hunyuan and Qwen-Image-edit have strong overall scores, while FlowEdit and DNAEdit retain competitive structural and background scores. On videos, FlowEdit-Wan scores highly for temporal consistency and naturalness, whereas ICVE has stronger text alignment. These dimension profiles explain why one aggregate rank is insufficient. Figures 3 and 5 show outputs that complement the quantitative profiles and make their visual differences directly inspectable.

The profiles also explain why preservation alone is insufficient to assess an edit. In the inversion-free image group, DNAEdit has higher structural and background scores than FlowEdit, but lower text alignment under both evaluators. This is a concrete adherence–preservation trade-off visible within

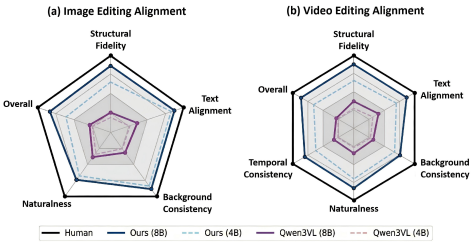

Figure 6: Aggregate human-rating profiles for zero-shot and distilled evaluators across the image and video rubric dimensions.

Human-evaluation profiles. Figure 6 suggests that fine-tuned evaluators better match the reported aggregate human profiles than their zero-shot initializations. The dimension-level view complements the MSE analysis by showing how the scoring interface reflects human assessments of adherence and preservation. This evidence concerns aggregate profiles; item-level agreement and calibration require separate validation. Appendix I details the human-evaluation scope and its relationship to teacher imitation.

Deployment scope. The deployment table separates weight storage, additional inference memory, and latency. These quantities describe different resource requirements and depend on the execution configuration. For the mixture-of-experts teacher, stored and activated parameters describe different resource demands. The smaller dense judges reduce weight storage, while runtime also depends on visual token counts and the image or video workload.

Table 4: Deployment & Inference Specifications for Evaluators (BF16 Precision)

Feature / Model Scale	**Our SFT 4B**	**Our SFT 8B**	**Qwen3-VL 235B-A22B**
Architecture Type	Dense	Dense	Mixture of Experts (MoE)
Total Parameters	4B	8B	235B
Activated Params (per Token)	4B	8B	22B
Weight VRAM (BF16)	≈ 8.8 GB	≈ 17.2 GB	≈ 470.8 GB
1K Image Inference (Extra VRAM)	≈ 1.2 GB	≈ 2.4 GB	$\approx 12-18$ GB
Suggested Minimum VRAM	12 GB	24 GB	640 GB ($8\times$80GB Cluster)
Avg. Inference Time (1K Image)	0.6 - 1.0 s	1.0 - 1.5 s	5.0 - 7.0 s (8 GPUs)

the same input paradigm. Keeping the four dimension columns therefore changes the interpretation of the Overall ordering: a high average need not identify the stronger method for a task where applying the requested change is the primary requirement. The two evaluator sizes expose the same direction of this comparison without requiring their absolute scores to coincide.

4.3 Teacher Imitation and Deployment

Two-stage training reduces image MSE from 2.95/2.51 to 0.49/0.47 and video MSE from 5.40/4.00 to 0.73/0.79 (4B/8B; Table 3). Image-only training already lowers video MSE to 2.33/1.56, with further improvement after video supervision. The final 8B has lower image error and 4B lower video error; some operations depart from the aggregate trend. Appendices J and A break down stages, operations, and dimensions to support modality-specific evaluator selection.

The students require 8.8/17.2 GB of BF16 weight storage versus 470.8 GB for the teacher (Table 4). Appendices E, D, and F detail training, deployment, and score aggregation, respectively.

5 Conclusion

UniEditBench connects image and video editing through a shared operation taxonomy, visually grounded prompt triplets, and dimension-specific evaluation. Its 633 image and 77 video sources support comparison of 25 image and eight video systems while retaining the distinction between following an instruction, preserving unaffected content, and producing plausible output. Distillation from a high-capacity teacher into 4B/8B judges makes the same scoring interface available with substantially smaller model-storage requirements; the two-stage training reduces teacher-score disagreement in both modalities. The benchmark's contribution is the combination of a consistent task representation, interpretable rubrics, complete editor profiles, and compact evaluators.

Limitations and future work. Video coverage remains compact, and human validation is at the aggregate-profile level. Future work will broaden video coverage and collect source-disjoint, item-level human judgments across more editing operations.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and consistency checking. The authors take responsibility for the final manuscript, including its mathematical derivations, implementation descriptions, experimental values, citations, and claims.

ETHICS STATEMENT

UniEditBench evaluates image and video editing with automated judges and a human-rating study. Edited media can misrepresent people or events, and automated scores can reproduce biases of the teacher and source data. These scores should support, rather than replace, human review in consequential applications. Participant privacy and the rights attached to source media require protection; benchmark use and any redistribution should respect the applicable licenses and consent requirements. Aggregate study results do not establish demographic fairness or universal human preference.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset, 2025a. URL https://arxiv.org/abs/2510.15742.

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*, 2025b.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.

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 preprint arXiv:2509.23951*, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7310–7320, 2024.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. In *International Conference on Learning Representations*, 2026.

Wei Chow, Linfeng Li, Lingdong Kong, Zefeng Li, Qi Xu, Hang Song, Tian Ye, Xian Wang, Jinbin Bai, Shilin Xu, et al. Editmgt: Unleashing potentials of masked generative transformers in image editing. *arXiv preprint arXiv:2512.11715*, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Lifan Jiang, Boxi Wu, Yuhang Pei, Tianrun Wu, Yongyuan Chen, Yan Zhao, Shiyu Yu, and Deng Cai. Snr-edit: Structure-aware noise rectification for inversion-free flow-based editing. *arXiv preprint arXiv:2601.19180*, 2026.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

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.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025a.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five-bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16672–16681, 2025b.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*, 2025c.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025d.

Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing. *arXiv preprint arXiv:2510.14648*, 2025.

Haonan Lin, Yan Chen, Jiahao Wang, Wenbin An, Mengmeng Wang, Feng Tian, Yong Liu, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *Advances in Neural Information Processing Systems*, 37: 115712–115756, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

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.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

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

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7899–7909, 2025.

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. Towards scalable human-aligned benchmark for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18292–18301, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023.

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.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. *arXiv preprint arXiv:2506.01430*, 2025.

Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-video 2: Scaling mllm-conditioned diffusion for unified video generation and editing. *arXiv preprint arXiv:2602.08820*, 2026.

Kaixiang Yang, Boyang Shen, Xin Li, Yuchen Dai, Yuxuan Luo, Yueran Ma, Wei Fang, Qiang Li, and Zhiwei Wang. Fia-edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing. *arXiv preprint arXiv:2511.12151*, 2025.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

A Detailed Alignment Error Analysis

The following tables report aggregate teacher-score MSE separately by modality, operation, and dimension. They are not item-level errors or estimates of human agreement. The main table uses the same modality-specific Overall columns.

Tables 5 and 7 detail the alignment errors categorized by fine-grained editing operations for the image and video datasets, respectively. To provide a deeper structural analysis, Tables 6 and 8 further decouple these errors across the scoring dimensions defined in the benchmark. The stage comparison has operation-specific exceptions. For image Extract, the image-only 8B error of 1.01 is lower than the two-stage errors of 1.26 and 1.23. Image-stage training can therefore remain preferable on an individual operation even when the final video average improves.

Table 5: Detailed MSE on **Image Editing** categorized by editing operations.

Model	Add	Adjust	Change	Count	Extract	Remove	Reorder	Replace	Stroke	Overall
Zero-shot 4B	1.94	1.91	3.43	2.85	5.92	3.38	4.00	2.11	1.47	2.95
Zero-shot 8B	1.34	1.46	2.46	2.34	5.24	3.00	3.95	1.87	2.11	2.51
SFT Image 4B	0.28	0.28	0.33	0.48	1.15	0.72	0.63	0.34	0.32	0.50
SFT Image 8B	0.24	0.23	0.34	0.47	1.01	0.71	0.57	0.33	0.31	0.47
SFT Image+Video 4B	0.25	0.24	0.36	0.46	1.26	0.65	0.61	0.35	0.31	0.49
SFT Image+Video 8B	0.22	0.24	0.32	0.44	1.23	0.69	0.59	0.28	0.32	0.47

Table 6: Detailed MSE on **Image Editing** categorized by scoring dimension.

Model	Structural (S_{SF})	Alignment (S_{TA})	Background (S_{BC})	Naturalness (S_{NAT})	Overall
Zero-shot 4B	3.62	2.26	2.68	3.25	2.95
Zero-shot 8B	2.73	2.40	2.35	2.55	2.51
SFT Image 4B	0.66	0.28	0.57	0.49	0.50
SFT Image 8B	0.63	0.28	0.51	0.45	0.47
SFT Image+Video 4B	0.62	0.29	0.57	0.48	0.49
SFT Image+Video 8B	0.57	0.28	0.57	0.47	0.47

Table 7: Detailed MSE on **Video Editing** categorized by editing operations.

Model	Add	Adjust	Change	Count	Extract	Remove	Reorder	Replace	Overall
Zero-shot 4B	4.43	4.02	8.14	8.50	7.28	5.04	3.01	6.16	5.40
Zero-shot 8B	3.16	2.19	6.88	5.07	5.29	4.80	3.54	3.70	4.00
SFT Image 4B	1.03	1.35	2.36	2.01	10.95	3.53	1.83	1.37	2.33
SFT Image 8B	0.60	0.68	3.93	1.19	3.85	2.25	1.69	1.08	1.56
SFT Image+Video 4B	0.52	0.28	0.80	0.57	2.06	1.13	0.84	0.51	0.73
SFT Image+Video 8B	0.36	0.34	1.73	0.56	1.50	1.34	0.65	0.64	0.79

Table 8: Detailed MSE on **Video Editing** categorized by scoring dimension.

Model	Structural (S_{SF})	Alignment (S_{TA})	Background (S_{BC})	Naturalness (S_{NAT})	Temporal (S_{TSC})	Overall
Zero-shot 4B	7.03	4.61	4.58	4.77	6.01	5.40
Zero-shot 8B	5.66	3.81	2.49	3.40	4.63	4.00
SFT Image 4B	3.42	1.67	2.08	2.07	2.43	2.33
SFT Image 8B	2.06	1.53	1.14	1.39	1.65	1.56
SFT Image+Video 4B	0.87	0.80	0.58	0.65	0.75	0.73
SFT Image+Video 8B	1.00	0.83	0.60	0.69	0.81	0.79

B PROMPT TEMPLATES FOR KNOWLEDGE DISTILLATION

The templates specify the rationale-and-score response format for teacher supervision and student inference. Generating an explanation before its scores preserves a common, inspectable interface. Score agreement is assessed numerically, while explanation correctness and causal faithfulness are separate validation questions.

Image Teacher Prompt for Image Editing Evaluation

```
You are an expert image editing evaluation model. You will evaluate the
quality of an edited image compared to the original image.
**Input Information:**
- Original Prompt: {original_prompt}
- Edited Prompt: {edited_prompt}
- Original Image: [First image]
- Edited Image: [Second image]
**Evaluation Task:**
Evaluate the edited image across the following four dimensions. For each
dimension, provide a score from 0 to 5: 0 (Completely fails), 1 (Poor
quality with major issues), 2 (Below average), 3 (Acceptable quality),
4 (Good quality with very minor flaws), 5 (Excellent quality, meets all
requirements).
**Evaluation Dimensions:**
1. **Structural Fidelity**: Evaluate whether the structure, pose,
orientation, and spatial relationships of unedited entities remain
perfectly consistent with the original image.
2. **Text-Image Alignment**: Evaluate how accurately the edited content
reflects the requested changes described in the edited prompt.
3. **Background Consistency**: Evaluate the consistency of all regions
EXCEPT the edited subject. Check for unwanted changes, color shifts, or
distortions in unedited areas.
4. **Naturalness**: Evaluate whether the overall scene appears natural.
Look for noticeable flaws such as inconsistent lighting, perspective
errors, structural distortions, or watermarks.
**Important Guidelines:**
- [CRITICAL] You must FIRST provide a detailed step-by-step explanation
for each dimension, and THEN output the final numerical scores.
- Be critical and use the full range of scores (0-5). A perfect score
(5) should be rare and reserved for truly excellent results. Scores
below 2 indicate significant failures.
**Output Format:**
Provide your evaluation strictly in the following JSON format:

{
  "explanation": {
    "structural_fidelity": "<reasoning>",
    "text_image_alignment": "<reasoning>",
    "background_consistency": "<reasoning>",
    "naturalness": "<reasoning>"
  },
  "scores": {
    "structural_fidelity": <score>,
    "text_image_alignment": <score>,
    "background_consistency": <score>,
    "naturalness": <score>
  }
}
```

Video Teacher Prompt for Video Editing Evaluation

```
You are an expert video editing evaluation model. You will evaluate the
quality of an edited video compared to the original video.
**Input Information:**
- Original Prompt: {original_prompt}
- Edited Prompt: {edited_prompt}
- Original Video: [First video]
- Edited Video: [Second video]
**Evaluation Task:**
Evaluate the edited video across the following five dimensions. For each
dimension, provide a score from 0 to 5: 0 (Completely fails), 1 (Poor
quality with major issues), 2 (Below average), 3 (Acceptable quality),
4 (Good quality with very minor flaws), 5 (Excellent quality, meets all
requirements).
**Evaluation Dimensions:**
1. **Structural Fidelity**: Evaluate whether the structure, pose,
orientation, and spatial relationships of unedited entities remain
consistent with the original video.
2. **Text-Video Alignment**: Evaluate how accurately the edited content
reflects the requested changes described in the edited prompt.
3. **Background Consistency**: Evaluate the consistency of all regions
EXCEPT the edited subject. Check for unwanted changes, color shifts, or
distortions in unedited areas.
4. **Naturalness**: Evaluate whether the overall video appears natural,
checking for flaws like inconsistent lighting, perspective errors, or
structural distortions.
5. **Temporal-Spatial Consistency**: Focus on the continuity and
logical rationality of the video across time and space. Check if the
object maintains identity consistency, if motion trajectories are smooth,
and if spatial logic (e.g., gravity, depth) is reasonable. Penalize any
flickering, teleportation, or physical violations.
**Important Guidelines:**
- [CRITICAL] You must FIRST provide a detailed step-by-step explanation
for each dimension, and THEN output the final numerical scores.
- Be critical and use the full range of scores (0-5). A perfect score
(5) should be rare. Consider the intrinsic difficulty of the editing
task while maintaining consistent scoring standards.
**Output Format:**
Provide your evaluation strictly in the following JSON format:

{
  "explanation": {
    "structural_fidelity": "<reasoning>",
    "text_video_alignment": "<reasoning>",
    "background_consistency": "<reasoning>",
    "naturalness": "<step-by-step reasoning>",
    "temporal_spatial_consistency": "<reasoning>"
  },
  "scores": {
    "structural_fidelity": <score>,
    "text_video_alignment": <score>,
    "background_consistency": <score>,
    "naturalness": <score>,
    "temporal_spatial_consistency": <score>
  }
}
```

C QUALITATIVE EXAMPLES ACROSS TASK TAXONOMY

To supplement the quantitative findings in the main manuscript, this appendix provides comprehensive qualitative visualizations across UniEditBench's full task taxonomy. These examples demonstrate the benchmark's coverage to assess diverse, fine-grained editing operations, ranging from simple appearance-level attribute modifications to complex compositional and spatial reasoning tasks.

By presenting results across multiple state-of-the-art models from distinct paradigms, we visualize the current performance bottlenecks in automated visual editing—especially concerning spatial relations and fine-grained text alignment—which often remain hidden when analyzing holistic scalar metrics.

Qualitative Image Editing Visualization: Figure 7 displays generated results from representative models across all nine image editing categories defined in our visually-grounded taxonomy: Add, Adjust, Change, Count, Extract, Remove, Reorder, Replace, and Stroke. To showcase paradigm diversity, this visualization covers examples from Inversion-based, Inversion-free, Structured Editing, and MLLM-based methods. These snapshots illustrate distinct failure modes, such as counting errors, object merging during reordering, and background distortions, specific to each task and paradigm.

Qualitative Video Editing Visualization: Figure 8 presents qualitative results across all eight video editing categories: Add, Adjust, Change, Count, Extract, Remove, Reorder, and Replace. The examples cover three major paradigms: inversion-free, structured editing, and MLLM-based methods. In particular, Count and Reorder require object permanence and logically consistent motion across frames. The selected key frames illustrate the difficulty of preserving temporal-spatial consistency while performing accurate semantic edits.

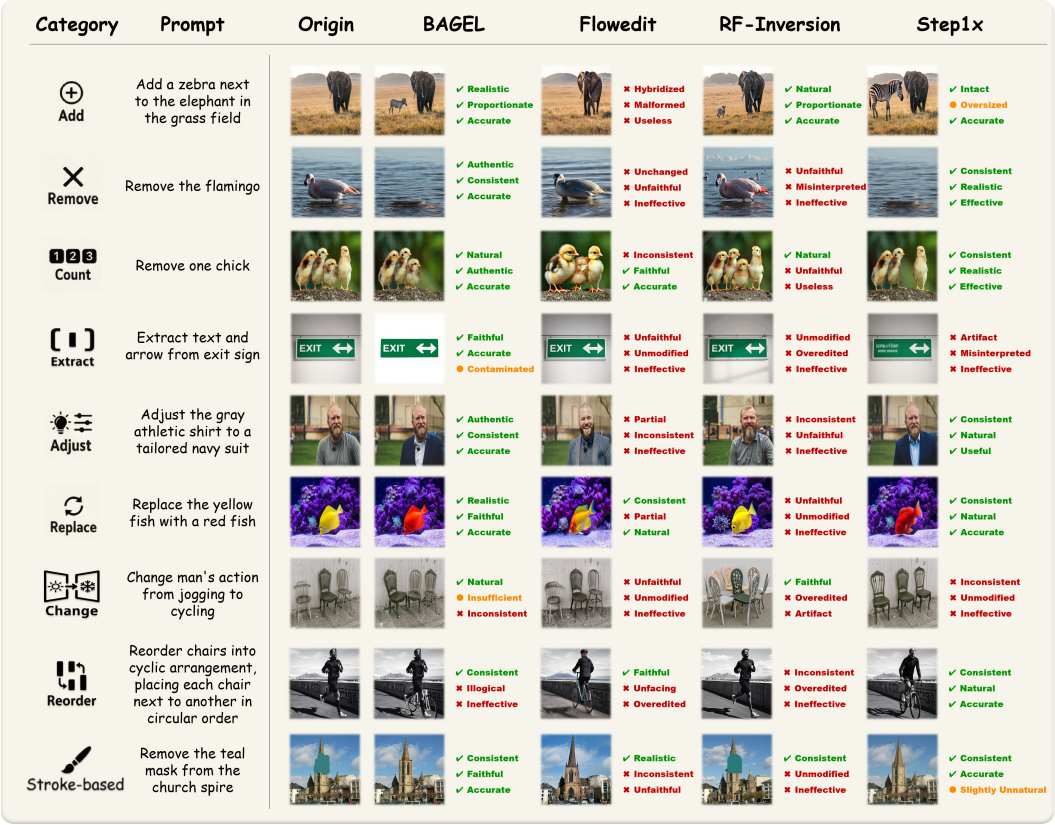

Figure 7: Comprehensive qualitative comparison on **Image Editing**. The examples cover the full 9-category taxonomy on the benchmark across four major paradigms: Inversion-based, Inversion-free, Structured Editing, and MLLM.

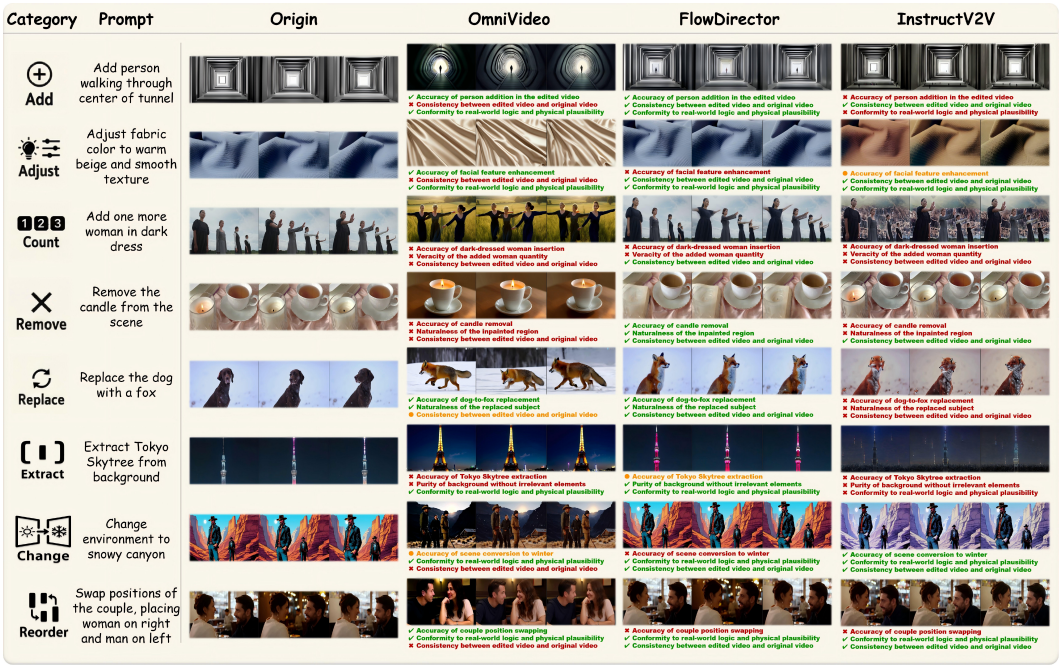

Figure 8: Comprehensive qualitative comparison on **Video Editing**. The examples cover the full 8-category taxonomy on the benchmark across three major paradigms: Inversion-free, Structured Editing, and MLLM.

D Deployment Cost and Ranking Resolution

Table 4 gives the reported deployment figures. The 4B and 8B models are dense; the teacher is a mixture of experts with 235B total and 22B active parameters per token. Its approximately 470.8 GB BF16 weight storage is much larger than the students' 8.8/17.2 GB. The teacher deployment uses eight GPUs, whereas the student configurations target one GPU. Activation memory and latency depend on visual tokens, batch size, decoding length, and hardware. The latency ranges should therefore be treated as configuration-dependent figures, not a controlled cross-hardware speedup or an API-cost comparison.

Editor scores also have finite interpretive resolution. For example, the leading image Overall values under the 4B evaluator are 4.37 and 4.31; the leading video values are 4.28 and 4.21. These gaps of 0.06 and 0.07 do not establish statistical separation without item-level uncertainty analysis. The reported tables are most useful for dimension profiles and larger differences, with narrow ordering treated cautiously. Appendix F explains why evaluator MSE cannot be converted directly into confidence intervals on these rankings.

Choosing a compact evaluator. Table 10 makes the deployment choice depend on modality. The final 8B evaluator has lower image imitation error than 4B (.47 versus .49), while 4B has lower video error (.73 versus .79). Parameter count alone therefore does not determine the preferred student. The modality profile and available memory jointly inform deployment, with the full dimension scores retaining more diagnostic value than a narrow Overall-rank difference.

E Training Configuration and Response Interface

Random seeds and result aggregation. Numerical experimental results in the main-text tables are averaged over runs with seeds 0, 42, and 100. Each run produces benchmark-level editor scores and evaluator errors under the stated image or video rubric; the reported value is their arithmetic mean across the three seeds. Seed 1000 provides an additional stability check and is not included in that mean. Results remain stable across the tested seeds. This consistency concerns repeated runs

of the evaluation, while agreement with human judgments is a separate property. Dataset counts, model sizes, and configuration entries retain their stated meanings.

The reported implementation uses ms-swift (Zhao et al., 2025) with LoRA fine-tuning, Qwen3-VL-4B/8B initialization, three image epochs, and three subsequent video epochs. The teacher is Qwen3-VL-235B-A22B. The templates define dimension names, integer score range, and rationale fields. Equation 2 evaluates numerical score disagreement and should not be interpreted as a claim that a separate regression head was trained.

The BF16 weight-memory figures concern model storage; they do not include all training states or imply a training-memory budget. The deployment table's additional-memory and latency ranges were reported for an image workload labeled "1K" in the deployment measurements. Since this label does not uniquely specify pixel dimensions or token count, those ranges are not normalized per-image-size performance claims. A workload definition, decoding configuration, and device model are necessary to transfer them to another deployment.

F Score Semantics and Aggregation

Let $s_{i,d}^{(j)}$ be evaluator j's score for instance i and dimension d. The per-instance average is $o_i^{(j)} = |\mathcal{D}|^{-1} \sum_d s_{i,d}^{(j)}$, and the benchmark average is $N^{-1} \sum_i o_i^{(j)}$. Images use four dimensions and videos use five. Means are rounded for presentation. The 4B/8B pair identifies two distinct evaluators. Each value is aggregated over the three run seeds described in Appendix E; the pair is not an uncertainty interval.

For an operation subset I_c, teacher MSE is

$$E_c^{(j)} = \frac{1}{|I_c||\mathcal{D}|} \sum_{i \in I_c, d \in \mathcal{D}} (s_{i,d}^{(j)} - s_{i,d}^{(T)})^2. \tag{3}$$

An overall MSE weights operation means by their instance counts, not equally across taxonomy names. The modality-separated tables report the respective errors; a mixed-modality operation mean would additionally require operation-specific sample weights. Keeping the modalities separate avoids obscuring those denominators.

The rubric is ordinal even though means and MSE treat adjacent score levels as equally spaced for reporting. A numerical gain of 0.1 should not be interpreted as a calibrated perceptual utility increment. Text alignment and preservation can legitimately move in opposite directions. For Count, the requested quantity belongs to alignment; unintended object deformation belongs to fidelity/naturalness. For Reorder, the requested positional change is exempt from preservation penalties. For Extract or a global style edit, the intended edit scope determines which background content should remain unchanged. These distinctions clarify interpretation without asserting that the dimensions are statistically independent.

Interpreting evaluator error. Let $s_{i,d}^{(k)}$ be the score assigned by student k to result i on dimension d, and let $s_{i,d}^{(T)}$ be the teacher score. The reported MSE measures disagreement between these scores over the specified result set. It is therefore an evaluator-imitation loss, whereas the editor tables report the scores themselves. A lower student MSE does not imply that a particular editor receives higher quality scores, nor does it establish better agreement with an independent human panel. For example, a student can closely reproduce a systematic teacher preference and retain the same error relative to people.

The Overall quality score averages the four applicable image dimensions or five video dimensions. Dimension profiles remain necessary: a successful edit may improve alignment while intentionally changing structure or appearance. Likewise, operation-level MSE depends on the mixture of results evaluated in that category. The modality-specific appendix tables make those comparisons explicit. We interpret score gaps as descriptive differences and use the human study as complementary evidence, rather than treating teacher imitation and human validity as interchangeable objectives.

Ranking uncertainty. Seed stability and uncertainty across benchmark instances address different sources of variation. The former is assessed in Appendix E; the latter requires instance-level

Table 9: Prompt-triplet interface and the conditions needed to interpret a comparison.

Interface	Natural text fields	Comparison condition
Description-based reconstruction	Source and target descriptions	Both descriptions must encode the same intended edit as the instruction.
Instruction-driven editing	Editing instruction and source media	The instruction must not add entity or spatial information absent from the matched description task.
Mask/stroke-conditioned editing	Applicable text plus spatial control	Spatial supervision must be declared; a supplied mask is additional information.
Video editing	Applicable text fields plus source clip	Duration, frame selection, and temporal request must be consistent.

comparisons. Global teacher MSE measures imitation error, not per-editor bias or the uncertainty of a score gap, so it cannot establish significance for margins of 0.06 or 0.07. Frames from the same video are also correlated and should not be counted as independent benchmark instances.

G Input Interfaces and Fair Comparison

The triplet $(p_{\mathrm{src}}, p_{\mathrm{tgt}}, u)$ records source description, desired description, and edit instruction. It supports the interface mapping in Table 9. This mapping describes admissible inputs; Appendix H discusses unsupported operations and parse failures. Neither establishes identical conditioning strength across implementations.

Why the shared schema matters. Consider an instruction that moves one object relative to another. A source description records their initial relationship, a target description specifies the desired relationship, and the instruction identifies the change. These fields expose the same semantic transformation to interfaces that otherwise accept different text formats. A spatial mask supplies a different kind of information: it can identify where to act even when the textual descriptions are unchanged. Keeping the fields explicit therefore makes the benchmark useful for comparing realized editing behavior while preserving the information needed to interpret that comparison. The operation taxonomy serves the same purpose at the task level: Count and Reorder test requested multiplicity and relations, rather than simply grouping results by visual appearance.

H Interface Failure Accounting

Unsupported operations, generation failures, and evaluator parse failures require distinct treatment. Comparisons restricted to supported tasks require a declared common subset and denominator; silently dropping failed or unsupported outputs changes the estimand. A parse failure by the evaluator is similarly different from a valid score of zero and should be accounted for separately. The aggregate tables cannot establish the per-method adaptation and failure policy by themselves, so cross-paradigm differences are descriptive outcomes of the evaluated configurations.

The label "unified" concerns the instance schema, common spatial rubric, and reporting interface. It does not require identical metric sets across modalities, identical generation architectures, or equal information in different control channels. Temporal consistency is an explicit video extension. HATIE's study of text-guided evaluation (Ryu et al., 2025) further motivates keeping conditioning and judging as separate protocol choices.

I Human Evidence, Teacher Bias, and Generalization

The human study and teacher-MSE tables answer different questions. A student can reduce imitation error while retaining the teacher's mistakes. For numerical human reference h, teacher t, and student s, the identity

$$\mathbb{E}(s-h)^2 = \mathbb{E}(s-t)^2 + \mathbb{E}(t-h)^2 + 2\mathbb{E}[(s-t)(t-h)] \tag{4}$$

Table 10: Teacher-score MSE by training stage. Image and video are summarized separately. Lower values indicate closer imitation, not necessarily stronger human agreement. Operation- and dimension-level values appear in Appendix A.

Training	Image 4B	Image 8B	Video 4B	Video 8B
Zero-shot	2.95	2.51	5.40	4.00
Image SFT	.50	.47	2.33	1.56
Image+video SFT	.49	.47	.73	.79

shows why the first term alone does not determine human error. The cross term can have either sign. Likewise, similar average radar profiles can conceal different item-level rankings.

The study reports 50 participants and five groups, but the aggregate figure does not establish a named correlation coefficient, inter-rater agreement, or a teacher–human baseline. Teacher outputs provide model-generated supervision rather than human ground truth. The selected human-study editors overlap with those used to construct teacher-scoring data. Editor overlap alone does not prove sample leakage; conversely, it cannot establish source-disjoint evaluation. Generalization to unseen source assets and unseen editor families requires those two notions of separation to be assessed independently.

Source provenance is another possible dependency. The benchmark mixes existing datasets, web assets, and synthetic outputs from FLUX/SD3 and Wan/HunyuanVideo. Editors from related families may be more compatible with some sources. Aggregate editor scores do not isolate this effect, so the reported results do not support source-invariant robustness. The 77 source videos supply limited coverage, especially for rare operations, long-duration consistency, and multi-event motion. A selected-frame illustration is not a complete temporal audit.

J What the Stage Ablation Establishes

Table 3 gives the stage errors used to decompose the total video-MSE reduction below.

Table 11: Derived decomposition of the video teacher-MSE reduction.

Quantity	4B	8B
Zero-shot to final reduction	4.67	3.21
Zero-shot to image-stage reduction	3.07	2.44
Image-stage share of total reduction	65.7%	76.0%
Additional video-stage reduction	1.60	.77

The final models improve overall video imitation while preserving broadly similar aggregate image error. This supports the value of adding video supervision. It does not separate ordering from training duration or establish superiority over joint, reverse-order, video-only, or equal-compute training. There are also individual image-operation regressions, so the aggregate should not be called evidence of complete prevention of forgetting.

Editor references. The evaluated systems are described in (Meng et al., 2021; Rout et al., 2024; Wang et al., 2024; Ju et al., 2023; Cao et al., 2023; Kulikov et al., 2025; Yang et al., 2025; Xie et al., 2025; Brooks et al., 2023; Liu et al., 2025; Zhang et al., 2025; 2023; Cao et al., 2025; Wu et al., 2025a; Deng et al., 2025; Li et al., 2025d; Wu et al., 2025b; Xia et al., 2025; Chow et al., 2025; Li et al., 2025c; Mu et al., 2025; Mao et al., 2025; Yu et al., 2025; Labs et al., 2025; Li et al., 2025a;b; Liao et al., 2025; Qin et al., 2024; Yang et al., 2026; Bai et al., 2025a).