VEFX-BENCH: A BENCHMARK FOR GENERIC VIDEO EDITING AND VISUAL EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI-assisted video creation becomes increasingly practical, instruction-guided video editing has become essential for refining generated or captured footage to meet professional requirements. Yet the field still lacks both a large-scale human-annotated dataset with complete editing examples and a standardized evaluator for comparing video editing models. Existing resources are limited by small scale, missing edited outputs, or the absence of human quality labels, while current evaluation often relies on expensive manual inspection or generic vision-language model judges that are not specialized for editing quality. We introduce `VEFX-Dataset`, a human-annotated dataset containing 5,049 video editing examples across 9 major editing categories and 32 subcategories, each annotated for three key aspects of editing quality: Instruction Following, Rendering Quality, and Edit Exclusivity. Building on `VEFX-Dataset`, we propose `VEFX-Reward`, a reward model designed specifically for video editing quality assessment. `VEFX-Reward` jointly processes the source video, the editing instruction, and the edited video, and predicts per-dimension quality scores via ordinal regression. We further release `VEFX-Bench`, a benchmark of 300 curated video-prompt pairs for standardized comparison of video editing models. Experiments show that `VEFX-Reward` aligns more strongly with human judgments than generic VLM judges and prior reward models on both standard IQA/VQA metrics and group-wise preference evaluation. Using `VEFX-Reward` as an evaluator, we benchmark representative commercial and open-source video editing models, revealing a gap between visual plausibility, instruction following, and edit locality in current models.

1 INTRODUCTION

The landscape of AI-assisted video creation is advancing rapidly. Recent video generation systems have shown impressive progress in producing photorealistic clips from natural-language prompts (Wan et al., 2025; Chen et al., 2025a; Kong et al., 2024; Polyak et al., 2024; OpenAI, 2024; Li et al., 2024; Wu et al., 2026; DeepMind, 2025; Kling AI, 2025; xAI, 2026). In professional production workflows, however, a prompt-generated video rarely satisfies the desired result in a single pass; it typically undergoes multiple rounds of targeted refinement, such as moving objects, adjusting camera motion, or adding visual effects, before it can be used. As a result, instruction-guided video editing (Jiang et al., 2025; Wei et al., 2025; Wan et al., 2025), where a user specifies a natural-language instruction to modify an existing video, has therefore become an essential component of AI-assisted filmmaking.

Despite rapid progress, the evaluation of video editing and visual effects (VFX) remains fundamentally unresolved. Unlike generic video generation, video editing must answer at least three distinct questions: Did the model execute the requested edit? Is the edited video visually coherent and temporally plausible? Did it preserve content that should have remained unchanged? These requirements expose two major bottlenecks. First, the field lacks large-scale human-annotated resources that contain complete editing triplets—the source video, the editing instruction, and the edited result—along with fine-grained quality labels. Second, evaluation still relies heavily on costly manual inspection or generic vision-language model (VLM) judges that are not designed for video-editing-specific assessment. The absence of a dedicated automatic evaluator makes both systematic benchmarking and preference-based optimization difficult.

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



Figure 1: **Overview of our framework.** We construct `VEFX-Dataset`, a human-annotated dataset with 5,049 video editing examples across 9 categories and 32 subcategories, scored along three decoupled dimensions: Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE). We train `VEFX-Reward`, a dedicated reward model for video editing quality assessment that takes the original video, editing instruction, and edited video as input and predicts per-dimension quality scores. We further release `VEFX-Bench`, a benchmark of 300 curated video-prompt pairs for standardized comparison of video editing models.

Existing resources address only parts of this problem. Benchmarks like EditBoard (Chen et al., 2025c), FiVE-Bench (Li et al., 2025), and IVE-Bench (Chen et al., 2025b) provide instructions without edited outputs; OpenVE (He et al., 2025) offers scale but relies heavily on automated generation and filtering rather than human annotation; VE-Bench (Sun et al., 2024) included edited videos and human scores but reduced quality to a single scalar and are built on older editing systems. On the reward-model side, prior work focuses on image editing or video generation quality rather than video editing itself (Wu et al., 2025a; Liu et al., 2025a). As a result, there is a pressing need for a benchmark and evaluator that jointly capture instruction faithfulness, rendering quality, and preservation of unedited content.

To address these gaps, we introduce `VEFX-Dataset`, `VEFX-Reward`, and `VEFX-Bench`. `VEFX-Dataset` contains 5,049 human-annotated video editing examples spanning 9 major categories and 32 fine-grained subcategories. Each example contains a source video, an editing instruction, and an edited result produced by a diverse mixture of commercial models, open-source models, and agentic editing pipelines. Trained annotators score each example along three dimensions that capture key aspects of editing quality: Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE). This design is central to the benchmark: an edit may be semantically wrong but visually clean, or visually strong while unnecessarily modifying non-target content. Building on `VEFX-Dataset`, we train `VEFX-Reward`, a reward model that takes the source video, the editing instruction, and the edited video as input and predicts per-dimension quality scores via ordinal regression. We further release `VEFX-Bench`, a standardized benchmark of 300 curated video-prompt pairs for systematic model comparison, and use `VEFX-Reward` to evaluate representative commercial and open-source video editing models under the same multi-dimensional protocol.

2 Related Work

Instruction-guided video editing modifies a source video according to natural-language instructions while preserving unrelated content (Jiang et al., 2025; Wei et al., 2025; Li et al., 2026; Motamed et al., 2026; Burgert et al., 2025; Gao et al., 2026a). Recent research models such as VACE, UniVideo, and the Wan family, together with commercial models such as Kling Omni, Grok Imagine, and Luma Ray2, have made the editing ecosystem highly heterogeneous (Wan et al., 2025; Cheng et al., 2025; Kling AI, 2025; xAI, 2026; Luma AI, 2025a). This heterogeneity makes standardized evaluation increasingly important. Existing editing benchmarks and datasets cover useful parts of the problem: EditBoard and FiVE provide task-oriented protocols, OpenVE-3M offers large-scale edited pairs with automatic multi-dimensional scoring, IVE-Bench defines a three-dimensional evaluation suite,

and VE-Bench includes edited videos with human scores (Chen et al., 2025c; Li et al., 2025; He et al., 2025; Chen et al., 2025b; Sun et al., 2024). However, these resources either lack human-scored edited-output triplets for reward-model training, rely primarily on automated supervision, or collapse editing quality into a single scalar. VEFX-Dataset is designed to fill this gap with human-annotated edited examples and decoupled labels for instruction following, rendering quality, and edit exclusivity.

Reward modeling for visual generation has progressed from image preference models such as ImageReward, HPS, and PickScore to image-editing and video-generation reward models (Xu et al., 2023; Wu et al., 2023; Kirstain et al., 2023; Wu et al., 2025a; He et al., 2024; Liu et al., 2025a; Wu et al., 2025b; Duan et al., 2025; Gao et al., 2026b). These models provide useful preference signals, but most are not specialized for video editing, where the evaluator must jointly reason over the source video, the instruction, and the edited result. Existing video-editing assessors such as VE-Bench also use a single scalar score, limiting diagnosis across different failure modes (Sun et al., 2024). In contrast, VEFX-Reward predicts multi-dimensional scores tailored to video editing quality assessment. We provide a more detailed discussion of related work in Appendix A.

3 VEFX-Dataset and VEFX-Bench

We present VEFX-Dataset, a human-annotated dataset for video editing quality evaluation, and VEFX-Bench, a standardized benchmark for systematic model comparison. VEFX-Dataset contains 5,049 editing examples—4,200 for training and 849 for testing—covering 9 major categories and 32 subcategories, each annotated along three decoupled quality dimensions: Instruction Following (IF), Rendering Quality (RQ), and Edit Exclusivity (EE). VEFX-Bench contains 300 curated (raw video, editing prompt) pairs for evaluating and comparing video editing models under a standardized protocol. This section describes the data collection process, annotation protocol, reliability check, and key dataset statistics.

Table 1 compares VEFX-Dataset with existing video editing datasets along properties that are particularly important for reward modeling: whether the dataset includes edited outputs, whether quality scores are direct human annotations, and whether the supervision distinguishes multiple failure modes. These properties matter because reward-model training requires actual edited results, reliable human supervision, and labels that separate instruction failure, rendering artifacts, and unintended edits. Several recent resources provide prompts without edited outputs (Chen et al., 2025c; Li et al., 2025; Chen et al., 2025b); others rely on automated filtering or judge models rather than trained annotators (Li et al., 2025; He et al., 2025); and some collapse editing quality into a single scalar. VEFX-Dataset is the only dataset in this comparison that provides edited outputs, human quality labels, and multi-dimensional supervision simultaneously.

Table 1: Comparison of VEFX-Dataset with existing video editing datasets. In the "#Cate" column, entries such as "8/35" denote 8 major categories and 35 subcategories. "Human Labels" indicates direct example-level human quality labels for edited outputs. "Multi-Dim." indicates whether quality evaluation is decomposed into multiple dimensions; † denotes automatic/MLLM-scored supervision, and ‡ denotes a protocol or metric suite without released human-scored edited-output triplets. "Editing Systems" summarizes the diversity of models used to generate edited videos.

Dataset	#Videos	#Pairs	#Cate	Edited Videos	Human Labels	Multi-Dim.	Editing Systems
VE-Bench (Sun et al., 2024)	169	1,170	6	✓	✓	✗	8 SD-based open-source (2024)
EditBoard (Chen et al., 2025c)	–	–	4	✗	–	–	–
FiVE (Li et al., 2025)	∼100	420	6	✗	–	–	–
OpenVE-3M (He et al., 2025)	1M	3M	8	✓	✗	✓†	Open-source + agentic (2025)
IVE-Bench (Chen et al., 2025b)	600	–	8/35	✗	✗	✓‡	–
VEFX-Dataset (Ours)	1,988	5,049	9/32	✓	✓	✓	4 Commercial + Open + Agentic (2026)

3.1 Data Collection

We curate source videos from open-source video datasets including Open-Sora (Zheng et al., 2024) and OpenVid-1M (Nan et al., 2024), supplemented with privately collected footage for additional diversity. We filter the initial pool for quality and usability, including sufficient resolution, duration, and temporal continuity, remove NSFW content, and then sample across scene categories and content types. The final set contains 1,419 source videos spanning 10 scene categories, summarized in Figure 3(c). We design instructions to cover 9 major editing categories and 32 subcategories, illustrated in Figure 2. We use Gemini 3 Flash (Google DeepMind, 2025c) to analyze video content,

assign suitable editing categories, and generate matched prompts. Low-confidence assignments are discarded. This process yields broad task coverage while keeping the editing instructions grounded in the source content. For each (source video, instruction) pair, we collect edited videos from a diverse mixture of commercial systems, open-source models, and agentic editing pipelines. This diversity is important because it exposes the benchmark to a broad range of quality levels and failure modes rather than the behavior of a single model family. Detailed model lists and pipeline descriptions are provided in Appendix E.

3.2 Annotation Protocol

Each editing example is evaluated on a 4-point scale along three decoupled dimensions. **Instruction Following (IF)** measures whether the edit satisfies the semantic requirements of the instruction. A score of 4 indicates that all requested edits are completed correctly, while a score of 1 indicates failure, contradiction, or an edit that is largely unrelated to the instruction. **Rendering Quality (RQ)** evaluates visual quality, including clarity, naturalness, temporal stability, and the absence of artifacts such as flickering, ghosting, blur, or distortion. This dimension is scored independently of whether the instruction is followed. **Edit Exclusivity (EE)** assesses whether the model changes only the intended target region without introducing unnecessary modifications elsewhere. In our annotation guide, score 4 means that no clear non-target change is introduced, score 3 corresponds to one localized non-target change, score 2 corresponds to two to three clear non-target changes or one large unintended background change, and score 1 indicates widespread or global over-editing.

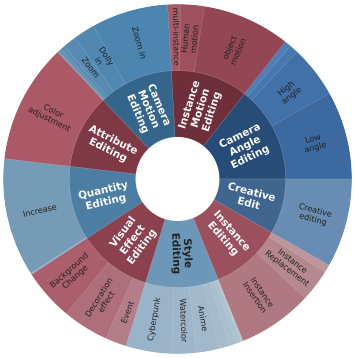

Figure 2: Task hierarchy of the 9 main editing categories and 32 subcategories in `VEFX-Dataset`.

IF, RQ, and EE are scored independently by 25 paid, trained annotators, with at least two independent ratings per example. An independent review pass flags score differences greater than two points on the 1–4 scale for re-annotation. Final dimension labels average all valid ratings, including those from re-annotation. Appendix F provides the complete rubric, bilingual annotation guide, and expert-guided training and calibration procedure.

3.3 Annotation Reliability

To assess annotation reliability, we conduct a targeted cross-check study. We randomly sample 550 examples from the dataset and assign them to a separate pool of six paid, trained annotators who did not participate in production labeling. Each example receives at least two independent cross-check ratings under the same rubric. We report two direct agreement measures: exact agreement and within-1-point agreement. We observe strong agreement across all three dimensions, with within-1-point agreement exceeding 91% (Table 2).

Table 2: Inter-annotator agreement on the 550-sample cross-check subset. Higher values indicate stronger agreement.

Metric	IF	RQ	EE
Exact Agreement (%)	75.2	87.2	72.2
Within-1 Agreement (%)	93.5	97.2	91.7

Table 2 shows strong agreement under this cross-check protocol. Within-1 agreement exceeds 91% on all three dimensions, reaching 93.5% for IF, 97.2% for RQ, and 91.7% for EE, while exact agreement remains high at 75.2%, 87.2%, and 72.2%, respectively. This pattern is intuitive: rendering quality is the easiest dimension to align on, while instruction following and edit exclusivity involve more borderline cases around partial success and acceptable non-target change. Although limited in scale, this study provides a useful sanity check that the three-dimensional labels are stable enough for training and evaluation. Additional details are provided in Appendix G.

3.4 Dataset Statistics and Analysis

We present analyses of `VEFX-Dataset` to characterize the dataset and motivate its three-dimensional design. Extended analysis is provided in Appendix H.

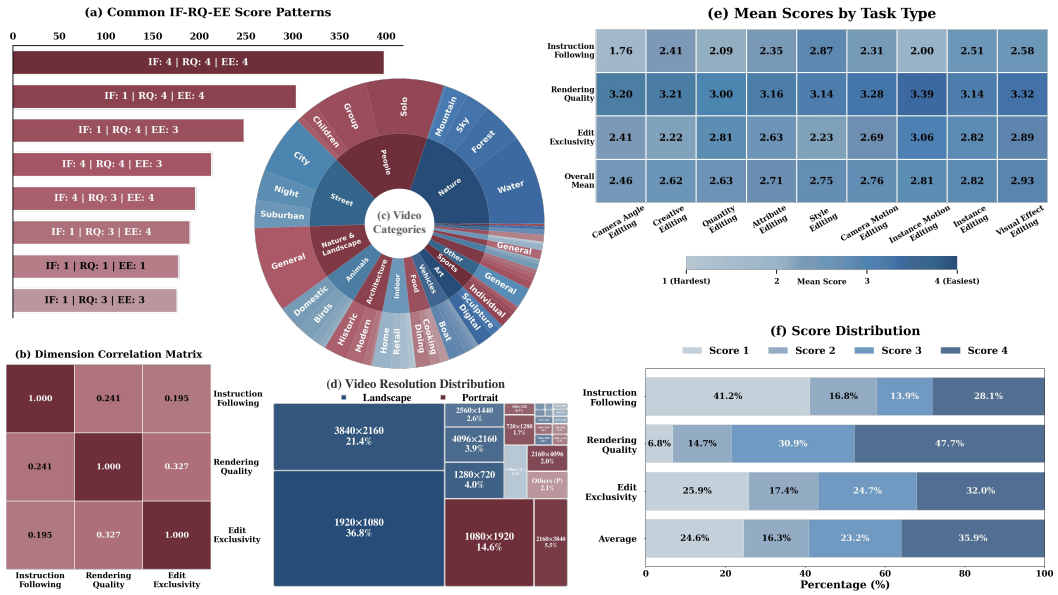

Figure 3: Overview of dataset statistics for VEFX-Dataset. Panel (a) shows common IF–RQ–EE score patterns; (b) reports pairwise dimension correlations; (c) summarizes video-category coverage; (d) shows the video-resolution distribution; (e) reports mean scores by task type; and (f) shows score distributions across annotation dimensions. Together they show that VEFX-Dataset spans diverse content and resolutions, exhibits heterogeneous task difficulty, and captures clear variation in difficulty across editing tasks.

Score distributions and score patterns. Figure 3 shows the distributions of three quality dimensions. IF is polarized, with many edits either failing outright or satisfying the instruction well; RQ is more concentrated at high scores, indicating that visual plausibility is often easier than semantic correctness; and EE remains more balanced. Common patterns such as $(1, 4, 4)$ and $(1, 4, 3)$ further show why a single scalar score would obscure cases where an output looks clean but does not follow the instruction.

Task difficulty across editing types. The task-type heatmap reveals variation in difficulty across editing categories. Camera Angle Editing is the hardest category overall, with IF = 1.76 and an overall mean of 2.46, while Quantity Editing is also challenging on IF at 2.09. By contrast, Style Editing reaches the highest IF at 2.87, and Visual Effect Editing attains the highest overall mean at 2.93. RQ remains relatively stable across task types, between 3.00 and 3.39, again suggesting that current systems are better at producing visually plausible outputs than at satisfying complex instructions. EE varies more strongly, with Instance Motion Editing reaching 3.06 while Creative Editing and Style Editing are lower at 2.22 and 2.23.

Coverage and dimension independence. Panels (c) and (d) show that VEFX-Dataset spans diverse scene types and video formats, including broad coverage over both landscape and portrait resolutions. Since editing difficulty depends on content and format, this diversity is important for robust evaluator training. The pairwise correlations between dimensions remain weak confirming that IF, RQ, and EE capture decompled failure modes.

4 VEFX-Reward: Human-Aligned Video Editing Reward Model

We present the design and training of VEFX-Reward, a reward model that predicts human-aligned quality scores for video editing results across three dimensions, instruction following, visual rendering quality, and edit exclusivity, by jointly considering the original video, the editing instruction, and the edited output.

Problem formulation. Given an original video $V_o = \{v_o^t\}_{t=1}^T$, an editing instruction P, and the corresponding edited video $V_e = \{v_e^t\}_{t=1}^T$, our goal is to predict quality scores that align with human judgment along the three annotation dimensions defined in Section 3.2: $[s_{\text{IF}}, s_{\text{RQ}}, s_{\text{EE}}] =$

$\mathcal{F}(V_o, P, V_e)$, where each score lies on the ordinal scale $\{1, 2, 3, 4\}$. The three dimensions require different reasoning. Instruction Following evaluates the semantic execution of the requested edit. Rendering Quality measures visual fidelity and temporal consistency. Edit Exclusivity compares the original and edited videos to detect unintended modifications outside the target region. A single aggregate score can obscure these distinct failure modes, which motivates the multi-dimensional formulation of `VEFX-Reward`.

Architecture. `VEFX-Reward` is instantiated on the Qwen3-VL-Instruct family (Bai et al., 2025b) at two scales, 4B and 32B, which correspond to `VEFX-Reward-4B` and `VEFX-Reward-32B` in the experiments. In both variants, the model jointly processes the original video, the edited video, and the editing instruction. This design allows the backbone to compare the edited result against both the requested change and the source content, which is essential for assessing semantic faithfulness, rendering quality, and unintended edits within one shared representation. We introduce three learnable special tokens, <|IF_reward|>, <|RQ_reward|>, and <|EE_reward|>, to query the three target dimensions. Their final hidden states are passed to a shared reward head, which produces the ordinal logits used for prediction. This token-based design gives each dimension its own query while preserving a single backbone for joint multimodal reasoning.

Ordinal regression objective. The bimodal distribution of IF scores (Section 3.4) and the ordinal nature of the 4-point scale motivate the use of ordinal regression rather than standard L2 loss. We adopt ordinal regression (Shaham et al., 2020), which models the score as a sequence of ordered threshold decisions instead of an unconstrained scalar regression target. For each dimension, the reward head predicts three ordered probabilities corresponding to whether the score is greater than 1, 2, and 3. Training applies binary cross-entropy to these ordered threshold predictions under the formulation:

$$\mathcal{L} = \sum_{d} \frac{1}{K-1} \sum_{k=1}^{K-1} \mathrm{BCE}\Big(\sigma(z_d^k),\ \mathbf{1}[y_d > k] \ \Big|\ y_d \geq k\Big), \tag{1}$$

where the conditional constraint $y_d \geq k$ ensures that each threshold is trained only on relevant samples, preserving the ordinal structure. At inference, we convert these ordered probabilities into a continuous score on $[1, 4]$ by taking their expected value: $\hat{s}_d = 1 + \sum_{k=1}^{K-1} P(Y > k)$. This soft prediction is used in all reported experiments.

Training details.

We train `VEFX-Reward` on 4,200 examples and evaluate on 849 examples from `VEFX-Dataset`, with source videos and prompt groups kept disjoint across splits. Original and edited videos are sampled at 4 FPS with aligned temporal indices and a maximum of 399,360 pixels per frame. We use a two-stage, 50-epoch AdamW training schedule in bf16 on 8 GPUs with an effective batch size of 8, selecting the checkpoint with the best validation loss. Full data processing, freezing schedules, and optimization settings are provided in Appendix B.

5 Experiments

We evaluate `VEFX-Reward` as a video editing quality assessor. We compare against generic VLM-as-judge baselines and prior reward models, analyze global agreement with standard IQA/VQA metrics, and further test whether the learned scores preserve local human preferences. Additional evaluations and ablations are provided in Appendix C.

5.1 Experimental Setup

Evaluation metrics. Our primary evaluation follows standard IQA/VQA protocol. We report Spearman Rank-Order Correlation Coefficient (SRCC), Kendall Rank-Order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE) in Section 5.2. SRCC and KRCC are computed on raw predictions, while PLCC and RMSE are computed after the standard four-parameter logistic calibration. We complement these global correlation metrics with a group-wise preference metric, Pairwise Accuracy, in Section 5.3. Detailed metric definitions and the calibration protocol are provided in Appendix J.1.

Baselines. We compare `VEFX-Reward` against three types of baselines:

- *VLM-as-a-Judge*: Qwen3.5-397B, Qwen3.5-122B (Qwen Team, 2026), Gemini-3.1-Pro, Gemini-3.1-Flash-Lite, Gemini-2.5-Flash (Google DeepMind, 2026b;a; 2025b), and Seed-2.0-Lite, Seed-1.6 (ByteDance Seed Team, 2025). Each model receives the source video, editing instruction, and edited video, and is prompted to score editing quality on the same rubric used in human annotation.
- *EditReward*: an image editing reward model with two output heads, one aligned with instruction following and one aligned with generic visual quality (Wu et al., 2025a).
- *VE-Bench*: a video editing reward model that predicts a single scalar score (Sun et al., 2024).

Evaluator supervision. Prompted VLM judges use the IF/RQ/EE definitions and annotation examples provided to human annotators. Further more, we compare our results with a textual SFT baseline in Appendix C.1 uses the same Qwen3-VL-4B backbone and 4,200 training examples as `VEFX-Reward-4B`, enabling a comparison of autoregressive score generation and ordinal reward prediction under shared supervision.

Implementation details. We instantiate `VEFX-Reward` at two scales, `VEFX-Reward-4B` and `VEFX-Reward-32B`, using Qwen3-VL backbones at 4B and 32B with the same architecture and training objective. Both models are trained on the 4,200-example training split and evaluated on the 849-example test split. We sample both the original and edited videos at 4 FPS with a maximum frame resolution of 399,360 pixels while preserving aspect ratio, and train for 50 epochs in bf16 on 8 GPUs with an effective batch size of 8, selecting the checkpoint with the best validation loss. For VLM-as-judge baselines, we use a shared rubric-aligned prompt over the same source-video/instruction/edited-video triplet. In our evaluation, the human overall score is defined as the arithmetic mean of IF, RQ, and EE. For `VEFX-Reward` and VLM-as-judge baselines, the overall prediction is the mean of the three predicted dimension scores; for EditReward, it is the mean of its two native heads; and for VE-Bench, it is the model's native scalar output. Additional implementation details are provided in Appendix C.

5.2 Results on Standard IQA/VQA Metrics

Following standard IQA/VQA practice, we evaluate all methods with SRCC, KRCC, PLCC, and RMSE. The Overall columns in Table 3 report agreement on the human overall score rather than a separate learned target.

Table 3: Results on standard IQA and VQA metrics. SRCC, KRCC, and PLCC are higher-is-better; RMSE is lower-is-better. PLCC and RMSE are computed after logistic calibration. Overall denotes correlation on the human overall score, defined as the mean of IF, RQ, and EE. For `VEFX-Reward` and VLM-as-judge baselines, the overall prediction is the mean of the three predicted dimension scores; EditReward uses the mean of its two heads; and VE-Bench uses its overall score.

Method	SRCC↑				KRCC↑				PLCC↑				RMSE↓			
	IF	RQ	EE	Overall	IF	RQ	EE	Overall	IF	RQ	EE	Overall	IF	RQ	EE	Overall
VLM-as-a-Judge																
Seed-1.6	0.686	0.618	0.504	0.630	0.605	0.573	0.447	0.508	0.684	0.608	0.591	0.701	0.918	0.798	0.917	0.565
Seed-2.0-Lite	0.545	0.544	0.697	0.720	0.483	0.497	0.527	0.607	0.616	0.594	0.729	0.768	0.984	0.815	0.791	0.510
Qwen3.5-122B	0.379	0.563	0.658	0.631	0.327	0.523	0.573	0.520	0.378	0.663	0.601	0.685	1.165	0.752	0.893	0.578
Qwen3.5-397B	0.572	0.422	0.654	0.601	0.506	0.384	0.587	0.497	0.615	0.624	0.692	0.657	0.992	0.785	0.820	0.598
Gemini-3.1-Pro	0.731	0.518	0.681	0.752	0.559	0.459	0.584	0.608	**0.754**	0.510	0.644	0.726	0.826	0.864	0.788	0.546
Gemini-3.1-Flash-Lite	0.309	0.302	0.661	0.574	0.283	0.277	0.555	0.436	0.316	0.425	0.673	0.505	1.194	0.910	0.840	0.685
Gemini-2.5-Flash	0.256	0.217	0.581	0.383	0.236	0.195	0.544	0.296	0.256	0.491	0.569	0.478	1.216	0.875	0.934	0.697
Previous Reward Models																
EditReward (Wu et al., 2025a)	0.453	-0.211	–	0.558	0.342	-0.164	–	0.411	0.455	0.317	–	0.580	1.113	0.844	–	0.631
VE-Bench (Sun et al., 2024)	–	–	–	0.214	–	–	–	0.150	–	–	–	0.238	–	–	–	0.752
Ours																
`VEFX-Reward-4B`	0.714	**0.690**	0.693	0.760	0.564	**0.574**	0.556	0.595	0.704	**0.793**	0.710	0.771	0.888	**0.642**	0.764	0.493
`VEFX-Reward-32B`	**0.754**	0.681	**0.717**	**0.780**	**0.612**	0.567	**0.597**	**0.616**	0.751	0.792	**0.732**	**0.790**	**0.825**	0.643	**0.740**	**0.475**

Results. Both `VEFX-Reward` variants outperform prior reward-model baselines in predicting human overall scores. `VEFX-Reward-32B` achieves the best overall performance across all evaluated dimensions, while `VEFX-Reward-4B` performs slightly worse than `VEFX-Reward-32B` but still surpasses all other reward models. The improvement over prior reward models is substantial: EditReward achieves an overall SRCC of 0.558 and an RMSE of 0.631, whereas VE-Bench achieves only 0.214 SRCC and 0.752 RMSE. Strong VLM-based judges remain competitive on some indi-

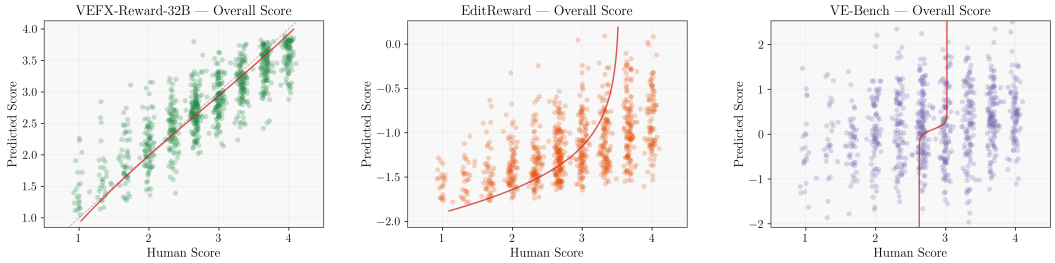

Figure 4: Predicted overall scores versus human overall scores for `VEFX-Reward-32B`, EditReward, and VE-Bench. Here the human overall score is defined as the mean of IF, RQ, and EE. `VEFX-Reward-32B` exhibits a tight monotonic trend that closely follows the human score axis, whereas EditReward shows a weaker and more nonlinear relationship, and VE-Bench displays substantially larger dispersion with limited sensitivity to score differences. Additional scatter plots are provided in Appendix J.1.

vidual dimensions, but they do not match the consistency of `VEFX-Reward` across dimensions and evaluation metrics.

Scatter-plot analysis. Figure 4 qualitatively confirms the table: `VEFX-Reward-32B` follows the human overall score with a tighter monotonic trend, whereas EditReward is more compressed and VE-Bench shows substantially larger dispersion.

5.3 Group-wise Preference Evaluation

Standard IQA and VQA metrics measure global correlation with human scores, but reward models are often used in the setting such that given several candidate edits for the same source video and instruction, the model should prefer the better one. We therefore add a ranking-based evaluation that measures whether a reward model preserves human ordering.

Pairwise Accuracy. Each ranking group g contains all candidate edits that share the same raw video and editing instruction; candidate edits may come from different video editing models, but comparisons are performed only within the group and never across groups. We enumerate all candidate pairs in each group and compare the predicted ordering with the ground-truth ordering. If the ground truth is tied, the pair is counted as correct regardless of the prediction; if the prediction is tied but the ground truth is not, it receives a score of 0.5. The dataset-level Pairwise Accuracy is

$$\mathrm{PairAcc} = \frac{\sum_{g=1}^{G} \sum_{(i,j)\in\mathcal{P}_g} \mathrm{Acc}_{ij}}{\sum_{g=1}^{G} |\mathcal{P}_g|}, \tag{2}$$

where $\mathrm{Acc}_{ij} = 1$ if the predicted order matches the human order. The Pairwise Accuracy depends only on relative ordering within each candidate group and is therefore insensitive to score-scale mismatch across models.

Preference results. Both `VEFX-Reward` variants substantially outperform previous reward models on group-wise preference consistency. `VEFX-Reward-32B` achieves the best overall Pairwise Accuracy at 0.872, while `VEFX-Reward-4B` remains close at 0.863, indicating that most relative preference signal is already captured at 4B scale. EditReward remains somewhat competitive on IF because one of its heads is naturally aligned with instruction following, but it performs poorly on RQ and EE because it is an image editing reward model without video-native temporal reasoning or a dedicated EE concept. VE-Bench shows moderate ordering ability, but its single-score design limits fine-grained candidate comparison. Together, these results confirm that `VEFX-Reward` is not only better aligned with human scores globally, but also more reliable for within-group candidate selection.

Table 4: Group-wise preference evaluation using Pairwise Accuracy. Overall denotes performance on the human overall score, defined as the mean of IF, RQ, and EE.

Model	IF	RQ	EE	Overall
EditReward	0.8283	0.5629	0.5317	0.7919
VE-Bench	0.7351	0.8127	0.7143	0.6651
`VEFX-Reward-4B`	0.9120	**0.9309**	0.9167	0.8628
`VEFX-Reward-32B`	**0.9366**	0.9111	**0.9196**	**0.8723**

Table 5: VEFX-Reward-32B-based evaluation of representative video editing models in IF, RQ, EE, Overall (Mean), and Overall (GeoAgg). Higher is better on all columns.

Model	Overall (GeoAgg)	Overall (Mean)	IF	RQ	EE
Commercial					
Kling o3 omni	**3.057**	**3.221**	3.033	**3.588**	3.043
Kling o1	2.985	3.183	**3.040**	3.534	2.976
Runway Gen-4.5	2.912	3.020	2.817	3.319	2.923
Seedance 2.0	2.766	3.107	2.811	3.421	3.088
Grok Imagine	2.723	3.109	2.606	3.346	**3.376**
Luma ray 3	2.717	2.936	2.702	3.403	2.705
Wan 2.6	2.146	2.592	2.012	3.317	2.446
Luma ray 2	1.804	1.977	2.038	2.532	1.363
Open-source					
UniVideo (Wei et al., 2025)	2.516	2.883	2.294	3.266	3.091
VACE (Jiang et al., 2025)	1.775	2.126	2.027	3.172	1.180

6 Benchmarking Existing Video Editing Models

Beyond baseline comparison, VEFX-Bench also enables a systematic evaluation of existing video editing models with our learned evaluator. We score 10 representative video editing models using VEFX-Reward-32B on the same 1–4 scale as VEFX-Dataset, including the commercial models Kling o3 omni (Kling AI, 2026), Kling o1 (Kling AI, 2025), Runway Gen-4.5 (Runway, 2025), Seedance 2.0 (ByteDance Seed Team, 2026), Grok Imagine (xAI, 2026), Luma ray 3 (Luma AI, 2025b), Wan 2.6 (Alibaba Cloud, 2025), and Luma ray 2 (Luma AI, 2025a), as well as the open-source models UniVideo (Wei et al., 2025) and VACE (Jiang et al., 2025). We report *Overall (Mean)* as the arithmetic mean of IF, RQ, and EE, and use *Overall (GeoAgg)* as the primary ranking metric. GeoAgg is a weighted geometric aggregate computed per sample before averaging, with IF receiving twice the weight of RQ and EE so that strong rendering quality cannot fully compensate for poor instruction following. For models with incomplete benchmark coverage, we report coverage-adjusted dimension scores; details are provided in Appendix J.2. Human validation of GeoAgg is provided in Appendix J.3.

6.1 Editing-Model Results

As shown in Table 5, Kling o3 omni ranks first under *Overall (GeoAgg)*, followed by Kling o1 and Runway Gen-4.5. The top models combine strong IF and RQ with competitive EE, while models with high visual quality but weaker instruction following or edit locality are penalized by the shortfall-sensitive aggregate. Grok Imagine achieves the strongest EE score, but its lower IF reduces its GeoAgg rank; UniVideo is the strongest open-source model and remains competitive with several commercial models, especially on EE. Overall, the results suggest that modern video editing models often produce visually plausible outputs, but reliable instruction following and locality preservation still separate the strongest models from the rest.

7 Conclusion

We introduced VEFX-Dataset, a human-annotated dataset of 5,049 video editing examples with decoupled labels for Instruction Following, Rendering Quality, and Edit Exclusivity, together with VEFX-Reward for automated evaluation and VEFX-Bench for standardized model comparison. Across both standard IQA/VQA metrics and group-wise preference evaluation, VEFX-Reward consistently outperforms generic VLM judges and prior reward-model baselines, showing the value of task-specific reward modeling for video editing. Using VEFX-Reward as a scalable evaluator, we further benchmark representative commercial and open-source video editing models and analyze their behavior across editing tasks. This analysis shows that current models often achieve plausible rendering quality without reliably satisfying instructions or preserving non-target content, highlighting the value of reporting instruction following, rendering quality, and edit exclusivity separately. We hope these resources provide a practical foundation for benchmarking, model selection, and reward-driven optimization in video editing.

AI USE STATEMENT

As described in Section 3.1, Gemini 3 Flash assisted with video-content analysis, editing-category assignment, and instruction generation, and video editing models produced edited outputs. Vision-language models were also evaluated as judges (Section 5.1); the dataset's quality labels were supplied by human annotators. Generative AI tools also assisted with manuscript editing, organization, and the interpretation and presentation of experimental results. The authors take responsibility for the final content of the paper, including its claims and reported results.

REFERENCES

Alibaba Cloud. Alibaba unveils wan2.6 series enabling everyone to star in videos. https://www.alibabacloud.com/blog/alibaba-unveils-wan2-6-series-enabling-everyone-to-star-in-videos_602742, December 2025. Accessed: 2026-04-16.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

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.

Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz. Motionv2v: Editing motion in a video. *arXiv preprint arXiv:2511.20640*, 2025.

ByteDance Seed Team. ByteDance Seed: Models and research. https://seed.bytedance.com/, 2025. Accessed: 2026-02-27.

ByteDance Seed Team. Seedance 2.0 official launch. https://seed.bytedance.com/en/blog/official-launch-of-seedance-2-0, February 2026. Accessed: 2026-04-16.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23516–23527, 2025a.

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. *arXiv preprint arXiv:2510.11647*, 2025b.

Yupeng Chen, Penglin Chen, Xiaoyu Zhang, Yixian Huang, and Qian Xie. Editboard: Towards a comprehensive evaluation benchmark for text-based video editing models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15975–15983, 2025c.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

DeepMind. Veo 3 technical report. Technical report, DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Accessed: 2026-02-18.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Xiangbo Gao, Renjie Li, Xinghao Chen, Yuheng Wu, Suofei Feng, Qing Yin, and Zhengzhong Tu. Pisco: Precise video instance insertion with sparse control. *arXiv preprint arXiv:2602.08277*, 2026a.

Xiangbo Gao, Mingyang Wu, Siyuan Yang, Jiongze Yu, Pardis Taghavi, Fangzhou Lin, and Zhengzhong Tu. The pulse of motion: Measuring physical frame rate from visual dynamics. *arXiv preprint arXiv:2603.14375*, 2026b.

Google DeepMind. Gemini 2.0 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.0-flash, 2025a. Accessed: 2026-09-25.

Google DeepMind. Gemini 2.5 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, 2025b. Accessed: 2026-09-25.

Google DeepMind. Gemini 3 Flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/, December 2025c. Accessed: 2026-09-25.

Google DeepMind. Gemini 3.1 Flash-Lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, March 2026a. Accessed: 2026-09-25.

Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026b. Accessed: 2026-09-25.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Ralph L. Keeney. Multiplicative utility functions. *Operations Research*, 22(1):22–34, 1974.

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, 1993.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Kling AI. Kling video o1 user guide. https://kling.ai/quickstart/klingai-video-o1-user-guide, December 2025. Accessed: 2026-04-16.

Kling AI. Kling video 3.0 omni model user guide. https://kling.ai/quickstart/klingai-video-3-omni-model-user-guide, February 2026. Accessed: 2026-09-25.

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.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. *arXiv preprint arXiv:2503.13684*, 2025.

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024.

Zirui Li, Xinghao Chen, Lingyu Jiang, Dengzhe Hou, Fangzhou Lin, Kazunori Yamada, Xiangbo Gao, and Zhengzhong Tu. Physics-aware video instance removal benchmark. *arXiv preprint arXiv:2604.05898*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025a.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, et al. Light-x: Generative 4d video rendering with camera and illumination control. *arXiv preprint arXiv:2512.05115*, 2025b.

Luma AI. Introducing ray2. https://lumalabs.ai/changelog/introducing-ray2, January 2025a. Accessed: 2026-04-16.

Luma AI. Luma ai launches ray3. https://lumalabs.ai/news/ray3, September 2025b. Accessed: 2026-04-16.

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Hantang Liu, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos. *arXiv preprint arXiv:2508.18633*, 2025.

Saman Motamed, William Harvey, Benjamin Klein, Luc Van Gool, Zhuoning Yuan, and Ta-Ying Cheng. Void: Video object and interaction deletion. *arXiv preprint arXiv:2604.02296*, 2026.

Kepan Nan et al. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

OpenAI. Sora: Creating video from text. https://openai.com/index/sora/, February 2024. Accessed: 2026-09-25.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Accessed: 2026-09-25.

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 *ICML*, 2021.

Nikhila Ravi et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Runway. Introducing runway gen-4.5: A new frontier for video generation. https://runwayml.com/research/introducing-runway-gen-4.5, December 2025. Accessed: 2026-04-16.

Shaun R. Seaman and Ian R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013.

Uri Shaham, Igal Zaidman, and Jonathan Svirsky. Deep ordinal regression using optimal transport loss and unimodal output probabilities. *arXiv preprint arXiv:2011.07607*, 2020.

Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment. *arXiv preprint arXiv:2408.11481*, 2024.

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.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025a.

Mingyang Wu, Ashirbad Mishra, Soumik Dey, Shuo Xing, Naveen Ravipati, Hansi Wu, Binbin Li, and Zhengzhong Tu. Consid-gen: View-consistent and identity-preserving image-to-video generation. *arXiv preprint arXiv:2602.10113*, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025b.

xAI. Grok imagine api. https://x.ai/news/grok-imagine-api, January 2026. Accessed: 2026-04-16.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zangwei Zheng et al. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Additional Related Work

A.1 Instruction-Guided Video Editing

Instruction-guided video editing aims to modify a video according to natural-language instructions while preserving unrelated content (Jiang et al., 2025; Wei et al., 2025; Li et al., 2026; Motamed et al., 2026; Burgert et al., 2025; Gao et al., 2026a). Early methods extended image editing pipelines to the temporal domain, typically by introducing temporal attention or consistency modules on top of text-to-image diffusion models (Guo et al., 2023; Yang et al., 2023). More recent approaches adopt video-native diffusion or flow-matching architectures. Representative research models include VACE (Jiang et al., 2025), UniVideo (Wei et al., 2025), and the broader Wan family (Wan et al., 2025; Cheng et al., 2025). Alongside them, commercial systems such as Kling Omni, Grok Imagine, Luma Ray2, and the commercial Wan 2.6 service variant have reached practical quality levels (Kling AI, 2025; xAI, 2026; Luma AI, 2025a; Alibaba Cloud, 2025). The resulting ecosystem is highly heterogeneous, with different systems excelling on different editing types, which makes standardized evaluation increasingly important.

A.2 Video Editing Quality Evaluation

Evaluating video editing quality is intrinsically multi-faceted. Conventional metrics such as CLIP score, SSIM, and LPIPS capture only narrow aspects of the problem and do not directly measure instruction fidelity, temporal consistency, or unintended edits (Radford et al., 2021; Zhang et al., 2018). VBench and VBench++ provide broad evaluation suites for video generation, but they are not designed for editing, where the source and edited videos must be considered jointly (Huang et al., 2024; 2026). Several editing-oriented resources have been introduced more recently. `EditBoard` (Chen et al., 2025c) and `FiVE` (Li et al., 2025) provide useful task-oriented protocols, but at limited scope or scale. `OpenVE-3M` (He et al., 2025) collects large-scale edited pairs and uses MLLM-based multi-dimensional scoring for filtering and benchmarking, but its quality supervision is primarily automatic rather than direct human annotation. `IVE-Bench` (Chen et al., 2025b) proposes a three-dimensional metric suite covering video quality, instruction compliance, and video fidelity, with human pairwise studies used to validate metric alignment; however, it provides source videos and instructions rather than human-scored edited-output triplets for reward-model training. `VE-Bench` (Sun et al., 2024) includes edited videos and human scores, but reduces quality to a single scalar MOS. In contrast, `VEFX-Dataset` provides large-scale human-annotated video editing examples with decoupled quality labels tailored specifically to the editing setting.

A.3 Reward Models for Visual Generation

The success of RLHF in language modeling has motivated analogous efforts in visual generation. For image generation, reward models such as ImageReward, HPS, and PickScore learn to approximate human preference signals from large-scale annotations (Xu et al., 2023; Wu et al., 2023; Kirstain et al., 2023). This line of work has extended to image editing: EditReward trains a multi-dimensional reward model for instruction-guided image editing and demonstrates value for both evaluation and data curation (Wu et al., 2025a). In the video domain, VideoScore, VideoReward, DenseDPO, WorldScore, and Pulse model human preferences or preference-driven alignment primarily for video generation (He et al., 2024; Liu et al., 2025a; Wu et al., 2025b; Duan et al., 2025; Gao et al., 2026b). VE-Bench also includes a video editing assessor, but it predicts only a single scalar score and is tied to an earlier benchmark setting (Sun et al., 2024). These methods do not explicitly reason over the relationship between the source video and the edited result. `VEFX-Reward` addresses this gap by jointly processing the original video, the editing instruction, and the edited result, and by predicting multi-dimensional quality scores tailored to video editing.

B Additional Training Details for `VEFX-Reward`

We provide the implementation details omitted from the main paper. `VEFX-Reward` is trained on the 4,200-example training split of `VEFX-Dataset` and evaluated on the 849-example test split, with the split stratified across editing categories and pipelines. The split is performed at the source-video

and prompt-group level, ensuring that no source video or editing prompt appears in both training and test sets.

Video Processing. For each example, we uniformly sample both the original and edited videos at 4 FPS and cap the frame resolution at 399,360 pixels, approximately 632×632, while preserving aspect ratio through Qwen3-VL's dynamic-resolution mechanism. The two videos are sampled with aligned temporal indices to support direct comparison. The maximum sequence length is 32,768 tokens.

Optimization. We use a two-stage training schedule for 50 epochs. In the first stage, lasting 1 epoch, we freeze all pretrained parameters and train only the newly introduced reward tokens and reward head. In the second stage, lasting 49 epochs, we unfreeze and fine-tune the language backbone and visual-language merger together with the reward head and reward tokens, while keeping the vision tower frozen. We use AdamW with learning rates of 1×10^{-5} for the language-side parameters and 5×10^{-5} for the reward tokens, cosine decay, and a 15% warmup ratio. Training is performed in bf16 on 8 GPUs with an effective batch size of 8; the checkpoint with the best validation loss is used for evaluation. The three reward dimensions are optimized jointly with equal loss weights.

Compute Resources. We used a pool of up to 64 H100 GPUs for open-source and agentic edited-video generation over approximately one month, with variable utilization because generation jobs differ by editing category and pipeline. Reward-model training used 8 GPUs with an effective batch size of 8 for both the 4B and 32B models.

C Additional Experimental Details

Model Variants. We evaluate two `VEFX-Reward` variants, `VEFX-Reward-4B` and `VEFX-Reward-32B`, which instantiate the same reward-model design on Qwen3-VL backbones at 4B and 32B scales.

Evaluation Setup. Both variants are evaluated on the same 849-example test split. For VLM-as-judge baselines, we use a shared prompt that presents the original video, editing instruction, and edited video, and asks the model to score IF, RQ, and EE according to the same 1–4 rubric used in human annotation. In the main paper, the human overall score is defined as the arithmetic mean of the three human dimension scores. For `VEFX-Reward` and VLM-as-judge baselines, the overall prediction is defined as the mean of the three predicted dimension scores.

External Reward Models. EditReward and VE-Bench are evaluated through their native outputs. For EditReward, the overall prediction is defined as the mean of its two native heads; for VE-Bench, the overall prediction is its native scalar output. SRCC and KRCC are computed directly from raw predictions, while PLCC and RMSE use the same logistic calibration protocol as in the main paper when applicable. Since EditReward lacks a dedicated EE head and VE-Bench predicts only a single overall score, unavailable entries are marked with –.

C.1 Comparison with a Textual SFT Judge

We fine-tune Qwen3-VL-4B on the same 4,200 training examples using ordinary autoregressive supervised fine-tuning. This baseline generates a JSON object containing IF, RQ, and EE scores, whereas `VEFX-Reward-4B` predicts the dimensions with reward tokens and an ordinal head. Both methods are evaluated on the same 849-example test split. Overall is the arithmetic mean of the three dimension scores, and Overall PairAcc follows the group-wise protocol in Section 5.3.

Table A1 shows that textual SFT achieves strong human alignment, with Overall SRCC of 0.735 and Overall PairAcc of 0.842. Using the same backbone and training examples, `VEFX-Reward-4B` improves the scores in all three dimensions and reaches 0.760 Overall SRCC and 0.863 Overall PairAcc. The gains of 0.025 in Overall SRCC and 0.021 in rounded Overall PairAcc support the dimension-token and ordinal formulation for supervised video-editing evaluation. The complementary REG/CLS/ORD study in Appendix D examines the loss function within this formulation.

Table A1: Same-backbone, same-training-data comparison on the 849-example test split. SRCC and Overall PairAcc are higher-is-better. Both methods use Qwen3-VL-4B and the same 4,200 training examples. PairAcc is rounded to three decimal places.

Method	SRCC↑				Overall
	IF	RQ	EE	Overall	PairAcc↑
Qwen3-VL-4B textual SFT	0.688	0.654	0.665	0.735	0.842
`VEFX-Reward-4B`	**0.714**	**0.690**	**0.693**	**0.760**	**0.863**

C.2 Generalization to Unseen Editing Systems

To test transfer beyond the editing systems used to construct `VEFX-Dataset`, we generate and human-annotate 200 outputs from Seedance 2.0 and Luma Ray 3. Both systems edit the same 100 pre-frozen source-video/instruction cases, and each output receives at least two independent human ratings. Seedance 2.0 tests a new editor family, while Luma Ray 3 tests a new version of a family represented by Ray 2 during dataset construction. We evaluate the existing `VEFX-Reward-32B` without editor-specific retraining and average the valid human ratings within each dimension.

Table A2: Human alignment on unseen editing systems (SRCC↑). The macro row averages the two editor-wise correlations. Gemini-3.1-Flash-Lite is reported separately on the pooled 200-output set; its pooled correlations and the macro row have different aggregation definitions.

Evaluated outputs / aggregation	N	IF	RQ	EE
Evaluator: `VEFX-Reward-32B`				
Original test split	849	0.754	0.681	0.717
Seedance 2.0 (new family)	100	0.758	0.688	0.721
Luma Ray 3 (new version)	100	0.747	0.676	0.709
Unseen-editor macro	200	0.753	0.682	0.715
Reference evaluator: Gemini-3.1-Flash-Lite				
Unseen editors, pooled	200	0.309	0.304	0.647

`VEFX-Reward-32B` achieves unseen-editor macro SRCC of 0.753/0.682/0.715 for IF/RQ/EE, compared with 0.754/0.681/0.717 on the original test split (Table A2). Human alignment remains similar across the two editor-wise evaluations, demonstrating transfer to a new editor family and a new model version with a fixed evaluator. The table also reports pooled correlations for Gemini-3.1-Flash-Lite as a reference judge.

C.3 Human Alignment in Complex Scenes

We use the non-exclusive source-video tags from Appendix H.1 to examine `VEFX-Reward-32B` on complex-scene subsets of the 849-example human-labeled test split. The deduplicated union of examples with at least one tagged condition contains 401 edited outputs. Table A3 reports both this union and the four individual scenario slices. Overall is the arithmetic mean of IF/RQ/EE for both the predictions and the human labels, as in Section 5.2.

Alignment on the hard-scene union is close to the full-test result, while the individual slices reveal scenario-specific differences. Extreme lighting yields RQ and Overall SRCC of 0.600 and 0.730, compared with 0.681 and 0.780 on the full test set, indicating a greater challenge for quality assessment under these conditions. The slice-wise results complement the coverage statistics in Appendix H.1; estimates for the less frequent handheld and extreme-lighting conditions provide descriptive diagnostics.

C.4 Validation of Key Design Choices

We keep the ablation study intentionally lightweight, since the main contribution is the benchmark and reward-model formulation rather than a complex architectural recipe. Controlled development

Table A3: `VEFX-Reward-32B` alignment with human judgments in complex scenes (SRCC↑). The full test set contains 849 edited outputs and the deduplicated hard-scene union contains 401. Individual scenario slices overlap; the union's correlation is computed on its examples, not by averaging the slice correlations.

Evaluation subset	IF	RQ	EE	Overall
Full test set	0.754	0.681	0.717	0.780
At least one hard factor	0.760	0.674	0.711	0.779
Handheld capture	0.770	0.660	0.730	0.780
Dynamic viewpoint	0.780	0.690	0.730	0.800
Multi-object interaction	0.754	0.675	0.705	0.777
Extreme lighting	0.740	0.600	0.680	0.730

experiments support the final configuration: ordinal regression best matches the ordered 1–4 label space, 4 FPS provides the best temporal sampling trade-off, and a per-frame budget of roughly 400K pixels balances local detail and efficiency. Detailed ablation tables are provided in Appendix D.

Analysis. Ordinal regression is consistently the strongest choice in development, which is well aligned with the ordered 1–4 label space. For video preprocessing, 4 FPS provides the best trade-off between temporal coverage and redundant frames. Spatially, around 400K pixels per frame is the most effective operating point: lower resolution removes subtle local editing cues, while higher resolution increases computation without yielding clear gains. These trends support the default configuration used in the final `VEFX-Reward` models.

D Additional Ablation Details

We provide the supporting development ablations behind the final `VEFX-Reward` configuration. These studies compare the loss formulation, temporal sampling rate, and per-frame resolution budget, reporting validation error as the common criterion; lower average error indicates better human-label alignment.

Table A4: Loss-function ablation. REG uses continuous regression, CLS treats the four score levels as independent classes, and ORD uses ordinal regression. Lower validation error is better.

Loss	Best Eval Loss	IF Error	RQ Error	EE Error	Avg Error
REG	0.761	1.104	**0.452**	0.666	0.741
CLS	1.196	1.099	0.459	0.689	0.749
ORD	**0.479**	**1.049**	0.468	**0.612**	**0.709**

Table A4 shows that ordinal regression provides the strongest overall validation behavior. This supports treating the 1–4 labels as ordered scores rather than unrelated classes or unconstrained continuous targets.

Table A5: Temporal sampling ablation under the ordinal objective. Lower average validation error is better.

FPS	IF Error	RQ Error	EE Error	Avg Error
1	1.186	0.517	0.836	0.846
2	1.225	**0.461**	0.873	0.853
4	1.230	0.464	**0.815**	**0.836**
8	1.251	0.472	0.929	0.884

Table A5 shows that 4 FPS yields the best average validation error, balancing temporal coverage against redundant frames.

Table A6 shows that the selected 400K-pixel budget gives the best overall validation error. Lower resolution loses local editing cues, while higher resolutions do not yield consistent gains.

Table A6: Per-frame resolution-budget ablation under the ordinal objective. Lower average validation error is better.

Max Frame Pixels	IF Error	RQ Error	EE Error	Avg Error
154K	1.314	0.694	0.841	0.950
400K	**1.158**	**0.450**	0.869	**0.826**
450K	1.230	0.464	**0.815**	0.836
920K	1.184	0.467	0.866	0.839

E Editing Pipeline Details

We describe the detailed procedures for each category of editing systems used in `VEFX-Dataset`.

E.1 Commercial Models

For generic instruction-guided video editing, we directly submit (source video, instruction) pairs to four commercial APIs: Grok Imagine (xAI, 2026), Kling Omni (Kling AI, 2025), Wan 2.6 (Alibaba Cloud, 2025), and Luma Ray2 (Luma AI, 2025a). These systems accept free-form text instructions and produce edited videos end-to-end.

E.2 Open-Source Specialized Models

Instance Removal. We first apply SAM 2 (Ravi et al., 2024) to segment the target instance across all video frames, with manual verification and correction of segmentation masks. The corrected masks are then fed to ROSE (Miao et al., 2025), using the removal model retrained on the ROSE dataset with the PISCO (Gao et al., 2026a) framework to support 720p resolution and 121-frame sequences.

Instance Insertion. We use NanoBanana-Pro to perform the desired edit on a single reference frame, then extract the newly inserted object via SAM 2 segmentation. The extracted single-frame instance serves as a spatial control signal for PISCO, which propagates the insertion consistently across all video frames while preserving the background.

Instance Repositioning and Resizing. The target instance is extracted using SAM 2 with manual correction. A VLM (Gemini-2.0-Flash (Google DeepMind, 2025a)) interprets the editing instruction and provides guidance for the required spatial transformation (translation, scaling, rotation) of the segmented instance. Simultaneously, the PISCO-finetuned removal model inpaints the vacated region. Finally, PISCO-14B performs instance insertion using the transformed instance as a spatial condition, composited onto the inpainted background video.

Human Motion Editing. We extract human pose keypoints using ViTPose (Xu et al., 2022) and modify them according to Gemini-2.0-Flash's interpretation of the editing instruction, with human-in-the-loop verification of pose correctness. The modified pose sequence, together with additional control signals (Canny edges, depth maps from Depth Anything V3 (Lin et al., 2025)), condition Wan-Animate to generate a new video from the original first frame.

Camera Motion and Angle Editing. Gemini-2.0-Flash maps the natural language instruction to predefined camera trajectory parameters (pan, tilt, zoom, dolly, arc, etc.). ReCamMaster (Bai et al., 2025a) and LightX (Liu et al., 2025b) then execute the specified camera transformation on the source video.

Style, Creative, Visual Effect, and Attribute Editing. For these categories, we apply NanoBanana-Pro to edit the first frame according to the instruction, then use VACE (Jiang et al., 2025) in first-frame-conditioned mode to propagate the edit temporally across all frames. Additionally, for a subset of samples, we employ UniVideo (Wei et al., 2025) for direct end-to-end text-conditioned video editing, similar to the commercial-model usage pattern.

F COMPLETE ANNOTATION GUIDE

This appendix provides the complete annotation guide used to train annotators for `VEFX-Dataset`. The guide was provided in both English and Chinese; we present the English version here. The guide was finalized through three calibration stages: an expert 100-example pilot to refine the initial rubric, a first annotator 100-example pilot reviewed item by item by video-editing experts, and a second annotator 100-example pilot used to identify annotators requiring additional training or reassignment when systematic errors are observed. Annotator feedback and expert review inform these calibration rounds. During production, 25 paid, trained annotators, including both English- and Chinese-native speakers, provide at least two independent ratings per example. Each assigned annotator scores IF, RQ, and EE under the shared rubric. An independent review pass flags high-disagreement cases, and examples with score differences larger than two points on the 1–4 scale are re-annotated. Final dimension labels are the means of the valid ratings, including those obtained after any required re-annotation. Video-editing experts supervise rubric refinement and annotator calibration.

Table A7: Summary of the 4-point scoring rubric for each annotation dimension.

	Score 4	Score 3	Score 2	Score 1
IF	All requested edits completed correctly	Core edit completed with minor deviation	Partial execution with major semantic deviation	Failure, contradiction, or unrelated edit
RQ	Clear, stable, and artifact-free	Minor but noticeable degradation	Clear quality failure with recurrent artifacts	Severe visual breakdown
EE	No clear non-target change	One clear non-target change	Two–three non-target changes or one large unintended change	Global or widespread over-editing

F.1 GENERAL INSTRUCTIONS

Annotators are presented with an original video, an editing instruction, and one or more edited videos produced by different models. For each edited video, annotators independently score three dimensions on a 4-point scale (1–4). The three dimensions must be scored independently: the score on one dimension must not influence the score on another. When a result matches multiple descriptions, annotators should assign the lowest applicable score. This rule is especially important when a video has both good aspects and one clear failure that crosses the boundary to a lower level.

F.2 DIMENSION 1: INSTRUCTION FOLLOWING (IF)

This dimension evaluates whether the edited content accurately reflects the semantic requirements of the instruction.

- **Score 4 — Complete and Correct Execution.** All requested edits are clearly completed, and no required element is missing or incorrect. The target object, attribute, action, style, or camera change matches the instruction without visible contradiction.
- **Score 3 — Mostly Correct Execution.** The core edit is completed, but one minor detail is wrong or missing. Typical cases include correct target and edit type but slight mismatch in fine-grained attribute, appearance, intensity, or local extent. The result should still be recognizably aligned with the instruction overall.
- **Score 2 — Partial Execution with Major Deviation.** The video shows some relationship to the instruction, but the main requirement is only partially satisfied or is satisfied with a major semantic error. Typical cases include editing the correct region but producing the wrong object or attribute, executing only one part of a multi-step instruction, or mixing the requested edit with an obviously incorrect alternative.
- **Score 1 — Failure or Contradiction.** The instruction is not executed, the edit is largely unrelated, or the result directly contradicts the instruction. Examples include no visible edit, editing the wrong target, or changing the scene in the opposite direction from the requested operation.

F.3 DIMENSION 2: RENDERING QUALITY (RQ)

This dimension evaluates the visual quality of the edited video, including naturalness, clarity, physical correctness of object movements, temporal consistency between frames, and the absence of artifacts.

- **Score 4 — High Visual Fidelity.** The video is clear, temporally stable, and visually natural throughout. Artifacts are absent or only barely perceptible, object structure remains intact, and motion follows plausible physical behavior.
- **Score 3 — Minor but Noticeable Degradation.** The video remains fully watchable, but there are visible quality issues such as slight blur, local flicker, mild temporal inconsistency, or small artifact regions. These issues are limited and do not damage the overall scene structure or object identity.
- **Score 2 — Clear Quality Failure.** Artifacts are obvious and recurrent, such as repeated flicker, deformation, ghosting, severe blur, unstable boundaries, or unnatural motion. The content is still recognizable, but the defects substantially reduce visual quality and viewing coherence.
- **Score 1 — Severe Visual Breakdown.** The result is visually unusable or close to unusable. Major regions are corrupted, object identity collapses, temporal coherence is lost, or motion becomes physically implausible to the point that the video no longer supports reliable evaluation of the intended edit.

F.4 Dimension 3: Edit Exclusivity (EE)

This dimension evaluates whether the model executed only the specified operation without unnecessary changes to unrelated areas. A non-target change is defined as any clearly visible modification to an object, region, or background element that is not required by the instruction. When counting non-target changes, multiple altered instances in different semantic regions should be counted separately.

- **Score 4 — Strict Preservation.** No clearly visible non-target change is introduced. All regions outside the intended edit remain visually unchanged, except for imperceptible pixel-level differences or negligible rendering noise.
- **Score 3 — One Clear Non-Target Change.** The intended target is edited, but exactly one additional non-target object or semantic region is also clearly altered. The overall scene layout is still preserved, and the error remains localized.
- **Score 2 — Two to Three Clear Non-Target Changes.** Two or three non-target objects or semantic regions are clearly altered, or one large unintended background change affects a substantial part of the scene. The result still resembles the original video, but over-editing is obvious.
- **Score 1 — Global or Widespread Over-Editing.** More than three non-target objects or semantic regions are clearly altered, or the scene is globally rewritten. The result looks like a substantially different video rather than a localized edit.

F.5 Dimension Decoupling Principle

The three dimensions must be scored independently. Consider the following example:

Instruction: "Turn the apple into a banana."

Result: The model completely fails and the apple remains unchanged.

- IF = 1 (complete failure to follow the instruction)
- RQ = score independently (if the video quality is excellent, this can still be 4)
- EE = score independently (if no unintended changes occurred, this can still be 4)

This principle ensures that each dimension captures a distinct aspect of editing quality.

F.6 Annotation Examples

We provide all example cases from the annotation guide. Each figure shows the first frame of the original video together with one or more edited results and their IF/RQ/EE scores. The scores are assigned from the full video rather than from the displayed frame alone. These examples cover attribute editing, creative editing, instance editing, visual effects, and style transfer, and illustrate how the same instruction can lead to different score patterns across dimensions.

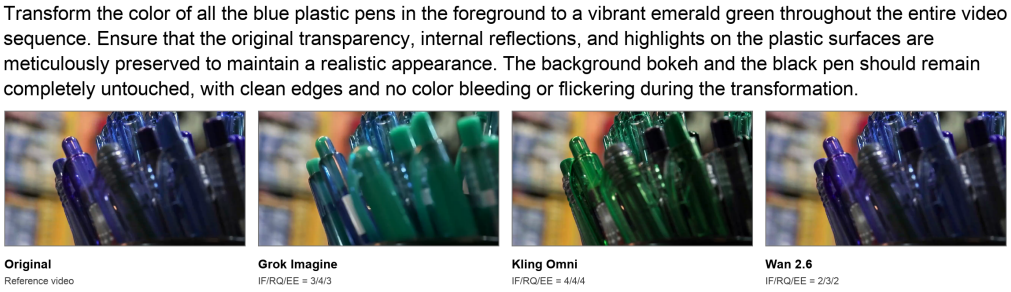

Figure A1: Annotation example 1. The instruction asks to turn only the blue foreground pens into emerald green while preserving transparency, reflections, highlights, the black pen, and the blurred background. Kling Omni receives IF/RQ/EE $= 4/4/4$ because it executes the requested color change cleanly, keeps the plastic appearance realistic, and leaves non-target content untouched. Grok Imagine receives $3/4/3$ because the target edit is mostly correct and visually clean, but the green conversion is less precise and some non-target regions are also affected, reducing both IF and EE. Wan 2.6 receives $2/3/2$ because the requested color transformation is incomplete, the result looks less stable and less realistic, and unintended color changes spill into regions that should have remained unchanged.

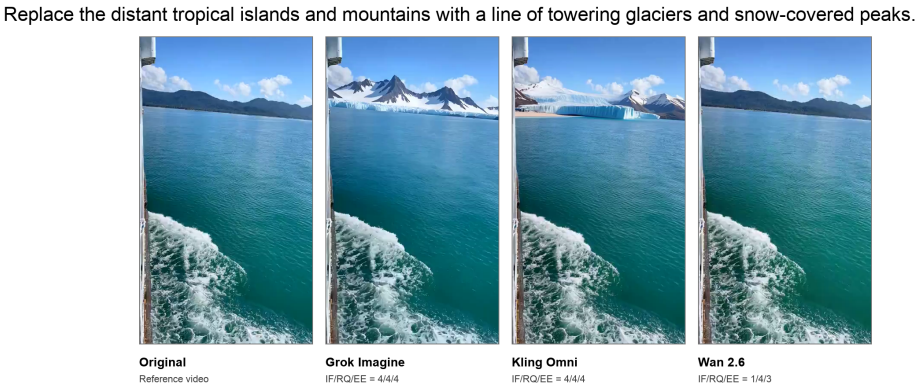

Figure A2: Annotation example 2. The instruction asks to replace the distant tropical islands and mountains with glaciers and snow-covered peaks while leaving the rest of the scene intact. Grok Imagine and Kling Omni both receive IF/RQ/EE $= 4/4/4$ because they fully carry out the requested background replacement, keep the water and boat-view foreground natural, and introduce no obvious non-target distortions. Wan 2.6 receives $1/4/3$ because it essentially fails to perform the requested semantic edit: the original tropical background remains, so IF is 1. Its RQ is still 4 because the video itself remains visually clean and artifact-free, which illustrates the intended decoupling between instruction following and rendering quality. EE is 3 rather than 4 because, although the edit does not heavily corrupt the scene, the output also does not faithfully realize the target modification.

G Inter-Annotator Agreement

G.1 Cross-Check Procedure

We randomly sample 550 examples from the annotated dataset and assign them to a separate pool of six paid, trained annotators who did not participate in the 25-person production pool. Each example receives at least two independent cross-check ratings. The second group follows the same annotation protocol and training procedure but has no access to the original annotations. Differences greater than two points on the 1–4 scale trigger re-annotation. Final dimension labels are computed by averaging the valid ratings. This independently re-annotated subset provides a focused consistency check.

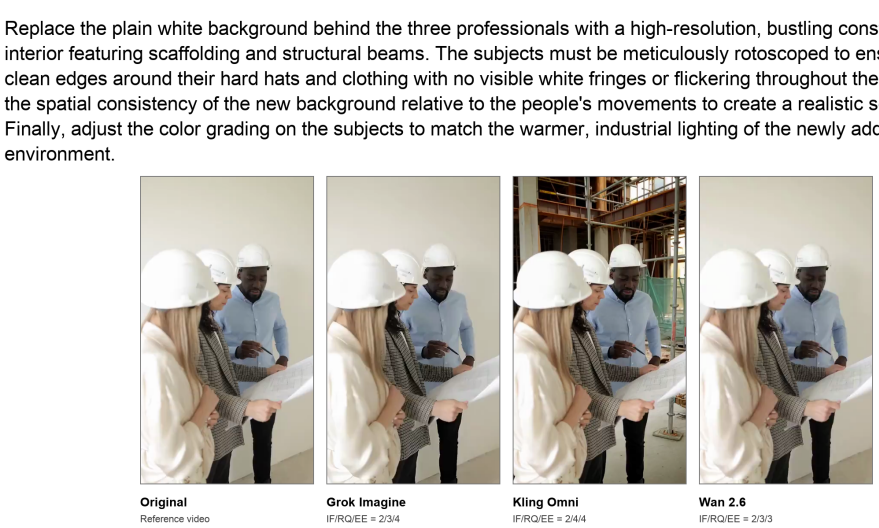

Figure A3: Annotation example 3. The instruction asks for a complete replacement of the white background with a bustling construction-site interior, together with clean subject boundaries, stable depth, and relit foreground subjects. All three edited results receive IF $= 2$ because they only partially satisfy the instruction: the construction-site replacement is introduced, but the compositing and relighting are not fully convincing, so the overall request is only partly achieved. Kling Omni receives the highest rendering score, RQ $= 4$, because its compositing is the cleanest and most visually coherent over time. Grok Imagine and Wan 2.6 receive RQ $= 3$ because the inserted environment looks less seamlessly integrated and shows weaker consistency. Grok Imagine and Kling Omni both receive EE $= 4$ because the three people remain largely intact, whereas Wan 2.6 receives EE $= 3$ because the foreground subjects are altered more noticeably during the edit.



Figure A4: Annotation example 4. The instruction asks for a heavy snowfall effect with visible accumulation on the mossy ground and bare branches. Grok Imagine receives IF/RQ/EE $= 4/4/3$ because it clearly adds snow and accumulation with strong visual quality, but it also alters parts of the scene beyond the requested effect. Kling Omni receives $3/4/4$ because the result is visually strong and preserves the scene structure well, but the snowfall effect is weaker than requested, so the instruction is not fully satisfied. Wan 2.6 receives $4/3/3$ because it does introduce the requested snow effect, but the rendering is less realistic and less stable, and some non-target structure is also modified.

G.2 Agreement by Dimension

Table A8 reports the agreement statistics used in the main paper.

Table A8: Inter-annotator agreement on the 550-sample double-annotation subset.

Metric	IF	RQ	EE
Exact Agreement (%)	75.2	87.2	72.2
Within-1 Agreement (%)	93.5	97.2	91.7

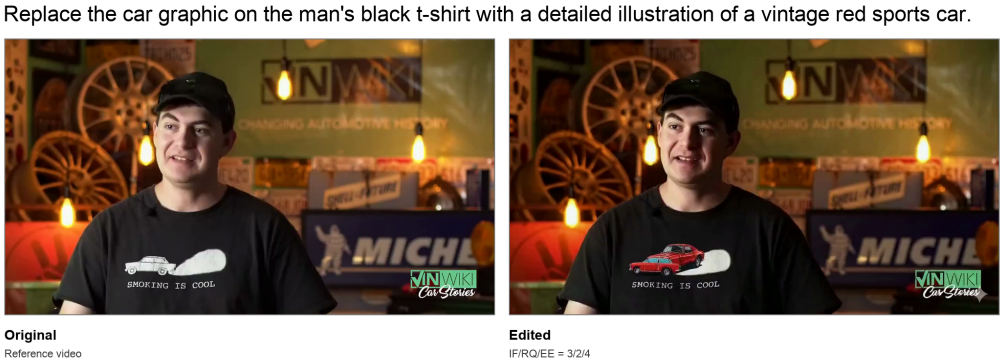

Figure A5: Annotation example 5. The instruction asks to replace the shirt graphic with a detailed vintage red sports car while keeping the rest of the person and scene unchanged. The edited result receives IF/RQ/EE $= 3/2/4$. IF is 3 because the shirt graphic is changed to a red car, so the main semantic request is met, but the inserted graphic is not fully convincing as a detailed vintage illustration across the clip. RQ is 2 because the edited graphic shows noticeable temporal instability and tracking inconsistency over time, even though the displayed frame looks acceptable. EE is 4 because the edit remains well localized to the shirt and does not introduce obvious unintended changes elsewhere in the video.

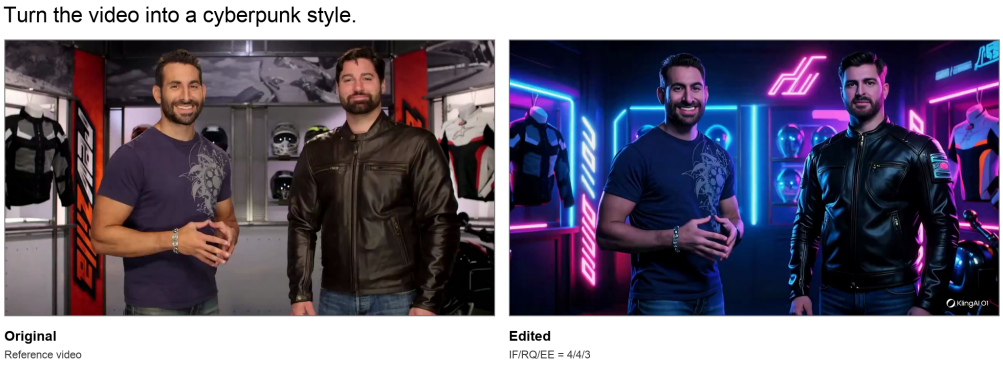

Figure A6: Annotation example 6. The instruction asks to convert the video into a cyberpunk style. The edited result receives IF/RQ/EE $= 4/4/3$. IF is 4 because the neon lighting, color palette, wardrobe styling, and overall atmosphere clearly match the requested cyberpunk aesthetic. RQ is 4 because the stylization is visually coherent and clean. EE is reduced to 3 because the transformation also modifies unrelated details, including text, facial appearance, and other local elements beyond the minimal style change needed to satisfy the instruction.

RQ achieves the strongest agreement, with 87.2% exact agreement and 97.2% within-1 agreement, indicating that rendering quality is relatively stable across annotators. IF also shows strong consistency, with 75.2% exact agreement and 93.5% within-1 agreement. EE remains the most challenging dimension, but still reaches 72.2% exact agreement and 91.7% within-1 agreement, which suggests that judgments about non-target changes are noisier yet still broadly consistent. Overall, these results support the reliability of the three-dimensional annotation protocol while also reflecting the inherently subjective nature of fine-grained video-editing assessment.

H Extended Dataset Analysis

We present additional analyses of `VEFX-Dataset` that complement the main text.

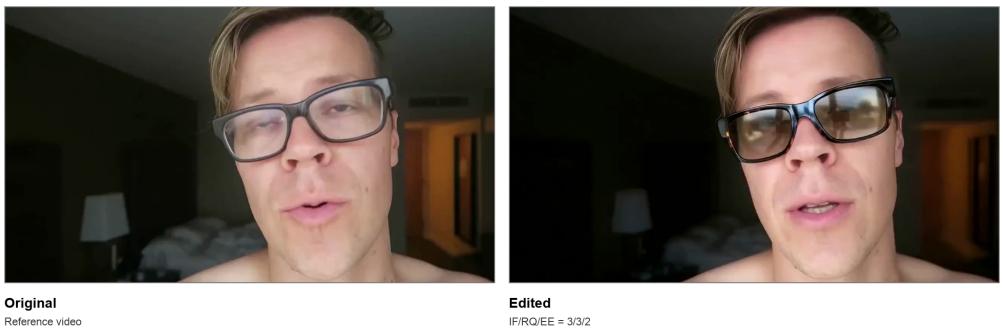

Figure A7: Annotation example 7. The instruction asks to replace the original black-rimmed glasses with gold-framed aviator sunglasses while preserving the man's facial features, expressions, head motion, reflections, and clean edges. The edited result receives IF/RQ/EE $= 3/3/2$. IF is 3 because the target object is indeed changed into sunglasses, but the replacement does not fully satisfy the requested appearance and realism. RQ is 3 because the local edit is usable but not fully clean, with only moderate realism in the eyewear integration. EE is 2 because the edit also changes other facial details, overall lighting, and background appearance, producing multiple unintended modifications outside the requested eyewear replacement.

H.1 Benchmark Composition and Complex-Scene Coverage

Table A9: Number of source-video/instruction cases per editing category in the fixed 300-case `VEFX-Bench` evaluation set.

Major category	Cases
Attribute Editing	33
Camera Angle Editing	34
Camera Motion Editing	33
Creative Edit	33
Instance Editing	33
Instance Motion Editing	33
Quantity Editing	34
Style Editing	34
Visual Effect Editing	33
Total	300

`VEFX-Dataset` and `VEFX-Bench` serve different purposes: the former provides human supervision and held-out evaluation for the reward model, whereas the latter is a fixed set of 300 unique source-video/instruction cases for repeated comparison of editing systems. The benchmark source videos are 720p, and the cases cover all nine major categories with 33–34 cases per category as shown in Table A9. This allocation gives major task categories approximately equal representation. Their complete counts are provided in Table A11.

Complex-scene audit. We apply the same Gemini-3.1-Flash classification prompt to the 1,419 `VEFX-Dataset` source videos and the 300 `VEFX-Bench` cases, assigning four non-exclusive tags: handheld capture, dynamic viewpoint, multi-object interaction, and extreme lighting. Table A10 reports the resulting model-assisted coverage audit. Tags may overlap, so the percentages are not additive. All four conditions are present, but handheld capture and extreme lighting remain underrepresented. These proportions characterize coverage; evaluator agreement within the observed conditions is assessed separately in Appendix C.3.

Table A10: Model-assisted complex-scene coverage audit. Dataset percentages use 1,419 source videos as the denominator; benchmark percentages use 300 cases. Tags are non-exclusive.

Scenario	VEFX-Dataset	VEFX-Bench
Handheld capture	5.1%	8.3%
Dynamic viewpoint	11.6%	12.0%
Multi-object interaction	42.1%	33.6%
Extreme lighting	3.2%	3.6%

H.2 Composition of the Fixed Evaluation Benchmark

Table A11 lists all 32 subcategories of the 300-case VEFX-Bench. These counts describe the fixed benchmark inputs and are distinct from task-frequency statistics of the 5,049-example VEFX-Dataset. Main-category totals match Table A9, but the subcategories are uneven: Arc shot and Ukiyo-e each have one case, Boom up has two, and Dolly out and Quantity Decrease each have three. These small groups support qualitative diagnostics of the corresponding editing capabilities.

H.3 Task Type Difficulty Ranking

Table A12 ranks the 9 task types by IF difficulty, and Figure A8 provides a finer-grained view across all 32 subcategories.

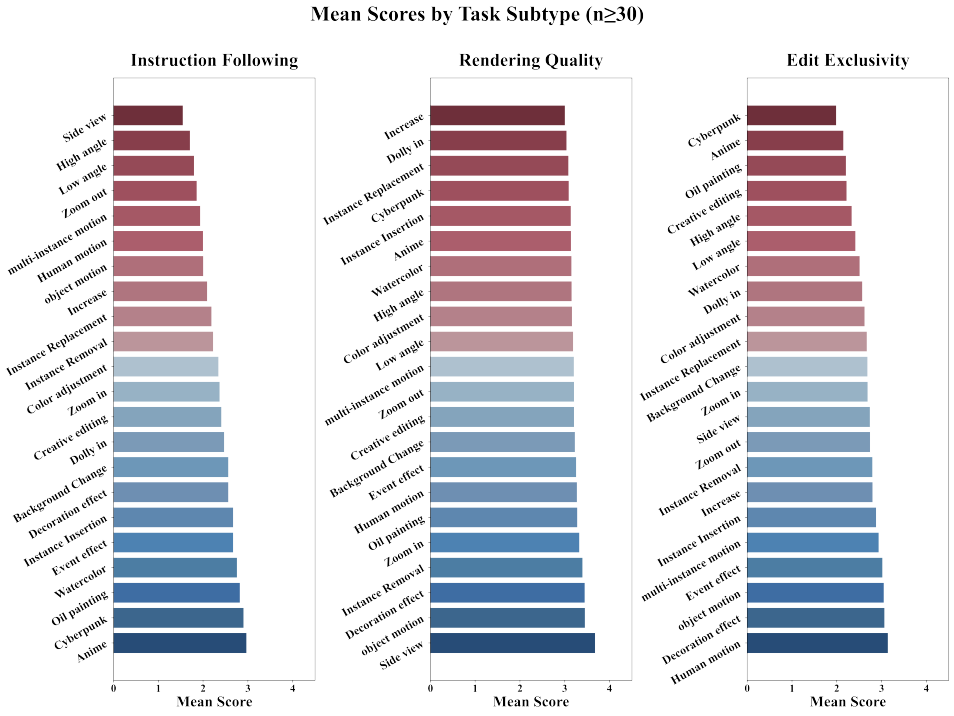

Figure A8: Score breakdown across all 32 subcategories grouped by 9 main categories. Fine-grained subcategory variation reveals which specific editing operations are most challenging.

Camera Angle editing is the hardest task for IF, as it requires geometric and 3D scene reasoning that current systems still handle poorly. Style Editing is the easiest for IF but has relatively low EE, reflecting the inherent tension between global style transformation and strict locality preservation. Notably, RQ varies much less across task types than IF or EE, which again suggests that current models find visual plausibility easier than precise semantic execution.

Table A11: Complete category and subcategory counts for the fixed 300-case VEFX-Bench. The 32 subcategories sum to 300 cases.

Major category	Subcategory	Cases
Attribute Editing	Color adjustment	33
Camera Angle Editing	Front	8
	High	9
	Low	9
	Side	8
Camera Motion Editing	Arc shot	1
	Boom up	2
	Dolly in	7
	Dolly out	3
	Tracking	6
	Zoom in	7
	Zoom out	7
Creative Edit	Creative editing	33
Instance Editing	Insertion	11
	Removal	11
	Replacement	11
Instance Motion Editing	Human motion	11
	Multi-instance motion	11
	Object motion	11
Quantity Editing	Decrease	3
	Increase	31
Style Editing	Anime	5
	Black-and-white	3
	Cyberpunk	5
	Ghibli	5
	Oil painting	5
	Ukiyo-e	1
	Watercolor	5
	Weather shift	5
Visual Effect Editing	Background change	11
	Decoration	11
	Event effect	11
Total	32 subcategories	300

Table A12: Task type ranking by editing difficulty. Tasks are sorted by IF score in ascending order, so lower values indicate harder semantic execution.

Task Type	IF	RQ	EE	N
Camera Angle	1.76	3.20	2.41	796
Instance Motion	2.00	3.39	3.06	450
Quantity	2.09	3.00	2.81	634
Camera Motion	2.31	3.28	2.69	383
Attribute	2.35	3.16	2.63	598
Creative	2.41	3.21	2.22	542
Instance	2.51	3.14	2.82	641
Visual Effect	2.58	3.32	2.89	520
Style	2.87	3.14	2.23	485

H.4 Video Difficulty and Training Signal Quality

Figure A9 shows the distribution of per-video difficulty, measured as the mean score across pipelines, against cross-pipeline score variance. Videos with high score variance are especially valuable for reward-model learning because they provide strong preference signals: different pipelines succeed or fail on the same input, enabling the reward model to learn discriminative features rather than a dataset-wide average.

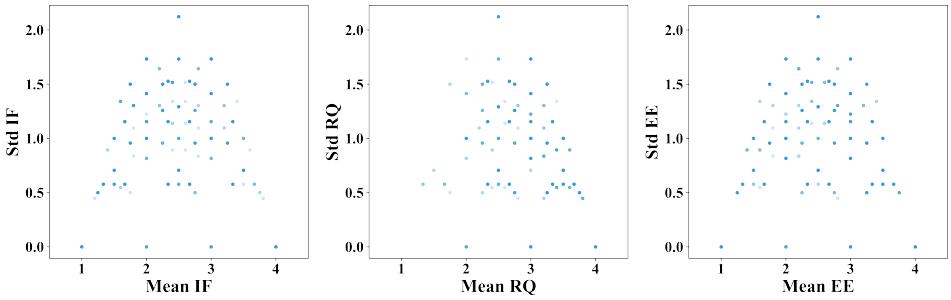

Figure A9: Per-video difficulty (mean score) versus cross-pipeline score variance. High-variance videos provide especially informative supervision for reward modeling.

I Per-Category Detailed Results

We provide finer-grained results for a representative subset of video editing models benchmarked in Section 6. Each heatmap-style table reports the mean `VEFX-Reward-32B` score for one dimension at the level of the 9 main editing categories, using the same model ordering as the main-text benchmark. These figures focus on comparative category-level behavior rather than distributional summaries.

Across the 9 main categories, IF shows the largest variation and remains the main source of separation between models. Grok Imagine and Kling Omni are strongest on many attribute, style, and instance-editing tasks, while camera-angle and camera-motion edits remain difficult for nearly all models. RQ is comparatively stable across categories, indicating that visually plausible outputs are often easier to produce than semantically correct ones. EE reveals the sharpest locality gap: Grok Imagine and UniVideo remain relatively strong on localized edits, whereas VACE and Luma Ray2 degrade more visibly when preserving non-target regions becomes difficult.

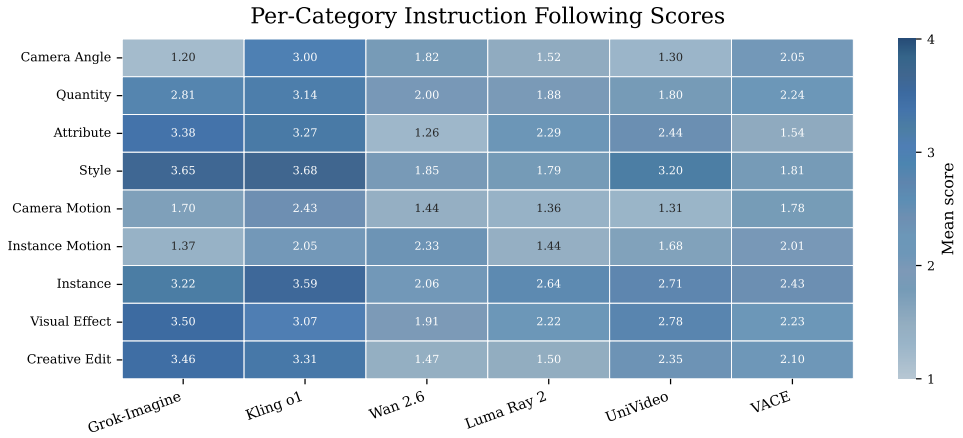

Figure A10: Heatmap table of per-category Instruction Following scores for the representative subset of benchmarked video editing models. Darker colors indicate higher `VEFX-Reward-32B` scores.

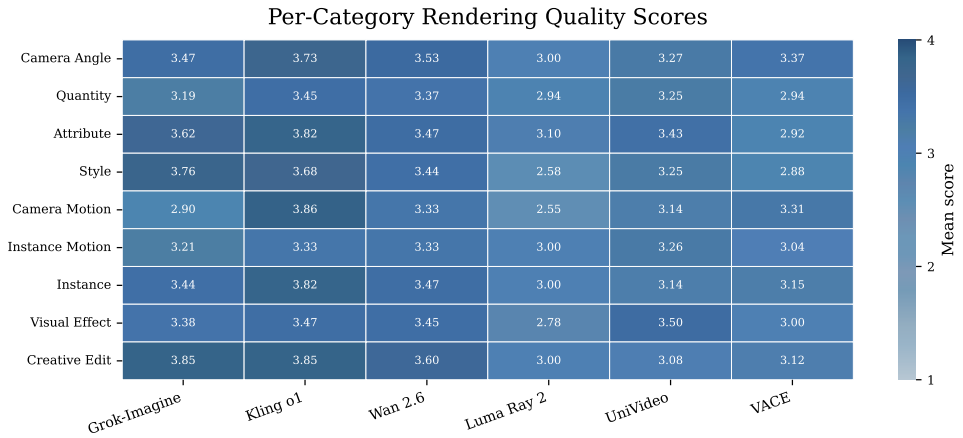

Figure A11: Heatmap table of per-category Rendering Quality scores for the representative subset of benchmarked video editing models.

J Additional Evaluation Details

J.1 Definitions of Standard IQA/VQA Metrics

In Section 5.2, we report four standard IQA/VQA-style metrics. We summarize their definitions here for completeness.

Spearman Rank-Order Correlation Coefficient (SRCC). SRCC measures monotonic agreement between predicted scores and human labels:

$$\mathrm{SRCC} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{A1}$$

where d_i is the rank difference between the i-th prediction and the corresponding human score.

Kendall Rank-Order Correlation Coefficient (KRCC). We use Kendall's τ-b to account for ties in the discrete 1–4 human scores:

$$\tau = \frac{N_c - N_d}{\sqrt{\left(\frac{n(n-1)}{2} - T_{\mathrm{pred}}\right)\left(\frac{n(n-1)}{2} - T_{\mathrm{human}}\right)}}, \tag{A2}$$

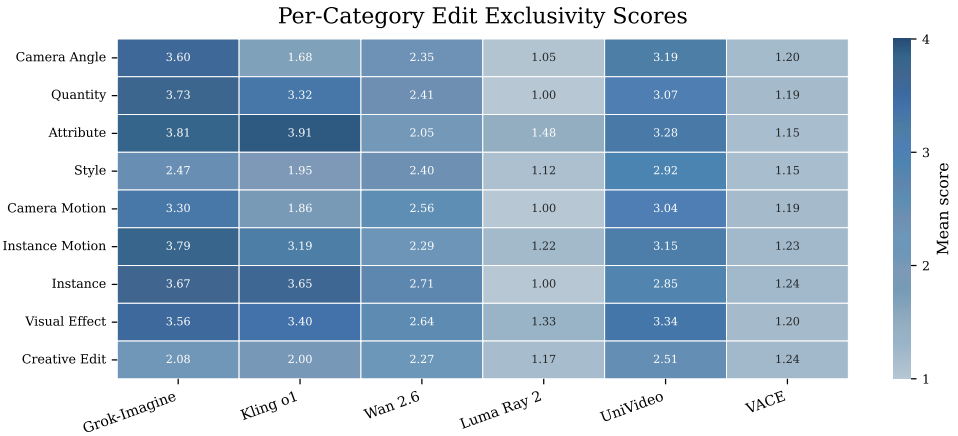

Figure A12: Heatmap table of per-category Edit Exclusivity scores for the representative subset of benchmarked video editing models.

where N_c and N_d denote the numbers of concordant and discordant pairs, and T_{pred} and T_{human} denote the numbers of tied pairs in the predicted and human rankings.

Pearson Linear Correlation Coefficient (PLCC). PLCC measures linear agreement after score calibration:

$$\text{PLCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{A3}$$

where x_i is the calibrated model prediction and y_i is the human score.

Root Mean Squared Error (RMSE). RMSE measures the calibrated absolute deviation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{A4}$$

Logistic calibration. Following common IQA/VQA protocol, we apply a four-parameter logistic mapping before computing PLCC and RMSE:

$$q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4, \tag{A5}$$

where x is the raw model score and $q(x)$ is the calibrated score. The parameters $\beta_1, \ldots, \beta_4$ are fitted by non-linear least squares on the evaluation set.

Supplementary scatter plot. Figure A13 directly compares `VEFX-Reward-4B` and `VEFX-Reward-32B` across IF, RQ, EE, and Overall. Here Overall is defined as the mean of IF, RQ, and EE for both predictions and human scores. The figure shows that scaling from 4B to 32B mainly improves IF, EE, and the overall score, where the 32B predictions form visibly tighter trends around the human annotations. By contrast, RQ remains relatively similar across scales, consistent with the main-text observation that rendering quality is already easier to model than semantic faithfulness and edit locality.

J.2 Additional Benchmark Methodology and Visualizations

Following prior work on multiplicative multi-attribute aggregation, we define *Overall (GeoAgg)* as a weighted geometric aggregate to reduce full compensability across dimensions and to penalize weak instruction following more strongly (Keeney & Raiffa, 1993; Keeney, 1974). For each evaluated sample i from model m, we first normalize the predicted scores to $[0, 1]$:

$$i_{m,i} = \frac{IF_{m,i} - 1}{3}, \qquad r_{m,i} = \frac{RQ_{m,i} - 1}{3}, \qquad e_{m,i} = \frac{EE_{m,i} - 1}{3}. \tag{A6}$$

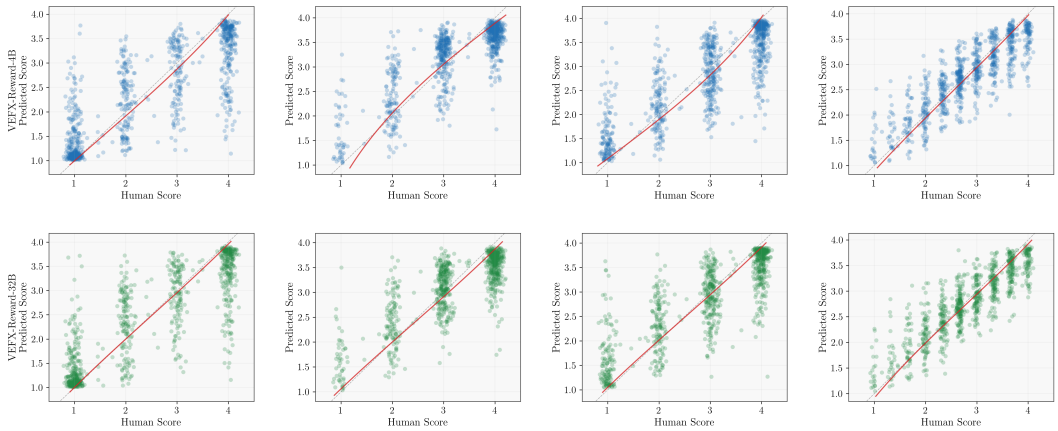

Figure A13: Side-by-side comparison between `VEFX-Reward-4B` and `VEFX-Reward-32B` across IF, RQ, EE, and Overall.

We then compute the sample-level aggregate and average it over the evaluated set Ω_m:

$$\text{Overall (GeoAgg)}_m = \frac{1}{|\Omega_m|} \sum_{i \in \Omega_m} \left[1 + 3 \left(i_{m,i}^{\alpha} r_{m,i}^{\beta} e_{m,i}^{\gamma} \right)^{\frac{1}{\alpha+\beta+\gamma}} \right]. \tag{A7}$$

We set $(\alpha, \beta, \gamma) = (2, 1, 1)$, so IF receives twice the weight of RQ and EE. We compute GeoAgg before averaging because the geometric aggregate is nonlinear; applying it after averaging IF, RQ, and EE would overestimate models with high variance or unbalanced per-sample behavior.

Adjusting incomplete model coverage. Some commercial models impose inference constraints, resulting in incomplete benchmark coverage for models such as Runway Gen-4.5 and Seedance 2.0. Rather than reporting a naive mean over each observed subset, which can be biased when coverage correlates with item difficulty, we treat incomplete coverage as a missing-data problem (Rubin, 1976). Our adjustment follows the standard inverse-propensity weighting principle (Horvitz & Thompson, 1952; Robins et al., 1994; Seaman & White, 2013). Let $R_{m,i} = 1$ indicate that model m has a valid evaluated output for benchmark item i, and let $\mathbf{x}_i$ denote item-level covariates such as task type, prompt length, and constraint count. We estimate the observation propensity $\hat{p}_{m,i} = \Pr(R_{m,i} = 1 \mid m, \mathbf{x}_i)$ and weight each observed score by a clipped inverse-propensity weight $w_{m,i} = 1/\hat{p}_{m,i}$. For each dimension d, we then fit a weighted linear mixed-effects model

$$y_{m,i,d} = \mu_{m,d} + u_i + \epsilon_{m,i,d}, \qquad u_i \sim \mathcal{N}(0, \sigma_u^2), \tag{A8}$$

where u_i captures item difficulty. The reported IF, RQ, and EE scores are coverage-adjusted model-level estimates $\hat{\mu}_{m,d}$ under the assumption that coverage is explainable by observed item covariates.

We also provide two additional distributional views of the ten-model evaluation. The strip plot in Figure A16 shows individual `VEFX-Reward-32B` scores, while the violin plot in Figure A17 emphasizes score-density shape for each model and dimension.

J.3 Human Validation of GeoAgg

Overall editing success depends on executing the requested edit while maintaining rendering quality and preserving unrelated content. GeoAgg captures this joint requirement through a weighted geometric aggregate. For example, an unchanged output with $(IF, RQ, EE) = (1, 4, 4)$ receives Mean $= 3$ and GeoAgg $= 1$ under the normalization in Equation (A7). We evaluate these aggregation rules against direct human ratings of overall editing success, complementing the dimension-derived Overall target in Section 5.2.

The overall-rating study uses 100 pre-frozen edited videos balanced across editing categories and systems. Three independent, blinded annotators rate each video's overall editing success on a 1–4 scale without seeing model identities or the predicted IF/RQ/EE scores. We average their ratings per

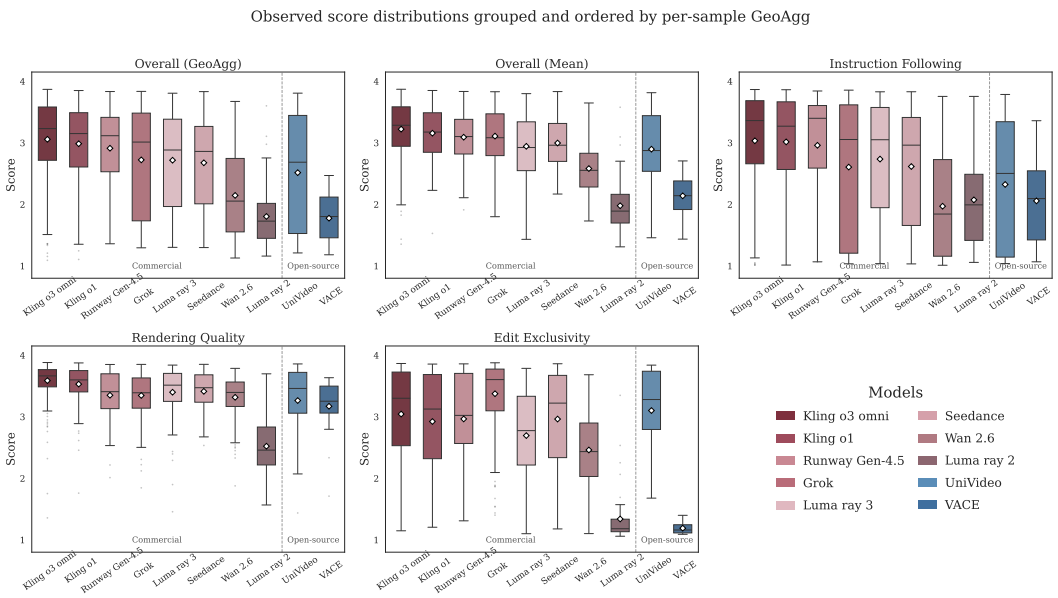

Figure A14: Observed soft-score `VEFX-Reward-32B` distributions for the benchmarked video editing models across *Overall (GeoAgg)*, *Overall (Mean)*, IF, RQ, and EE. Models are grouped by availability and ordered by per-sample *Overall (GeoAgg)*.

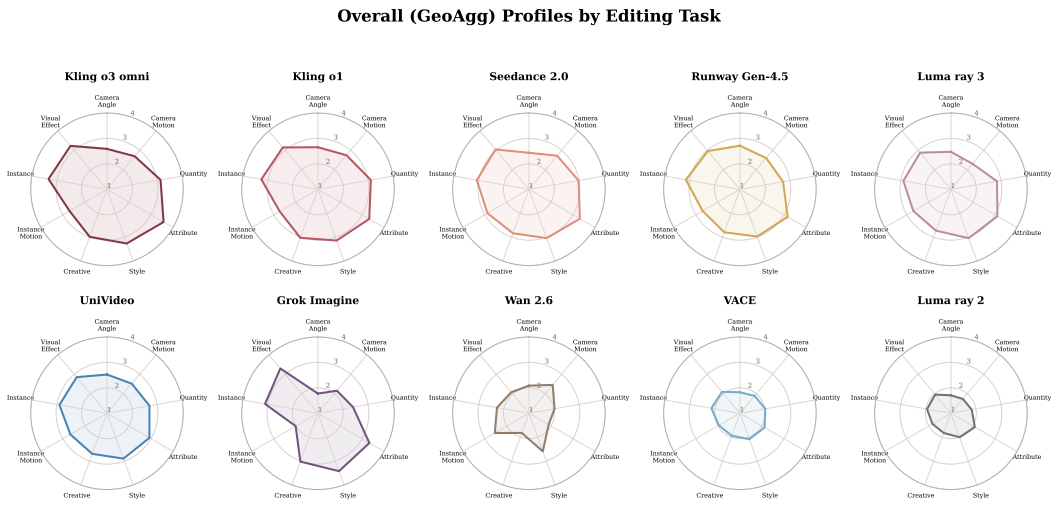

Figure A15: Task-wise *Overall (GeoAgg)* profiles of the benchmarked video editing models. Each radar plot uses the same radial scale, allowing the profile shape and absolute score level of each model to be compared across editing tasks.

video and compute SRCC with three aggregations of the same `VEFX-Reward` dimension predictions: arithmetic Mean, equal-weight geometric aggregation, and GeoAgg with weights $(2, 1, 1)$. Both geometric variants use the same normalization and per-video aggregation defined in Equation (A7).

GeoAgg achieves the strongest alignment with human overall ratings, with SRCC of 0.814 versus 0.752 for equal-weight geometric aggregation and 0.681 for Mean (Table A13). Its improvement of 0.133 over Mean has a paired video-bootstrap 95% confidence interval of $[0.058, 0.207]$ and $p = 0.002$. These results support geometric aggregation for overall editing success and the emphasis on instruction following in this study. The leaderboard reports IF/RQ/EE and Mean alongside GeoAgg to show each system's quality profile.

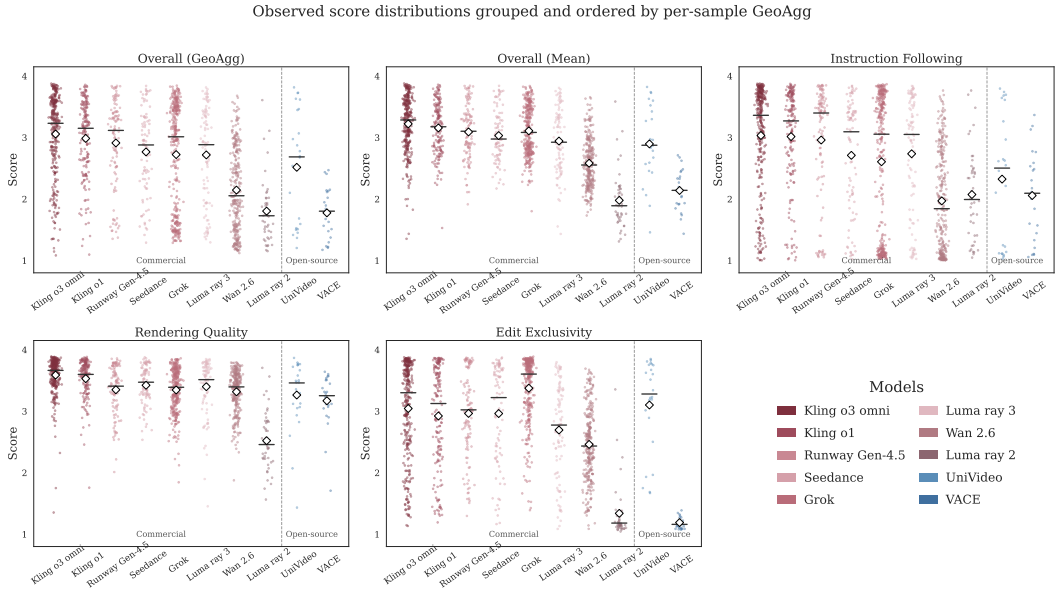

Figure A16: Strip plot of the ten-model evaluation, showing individual `VEFX-Reward-32B` scores for IF, RQ, and EE.

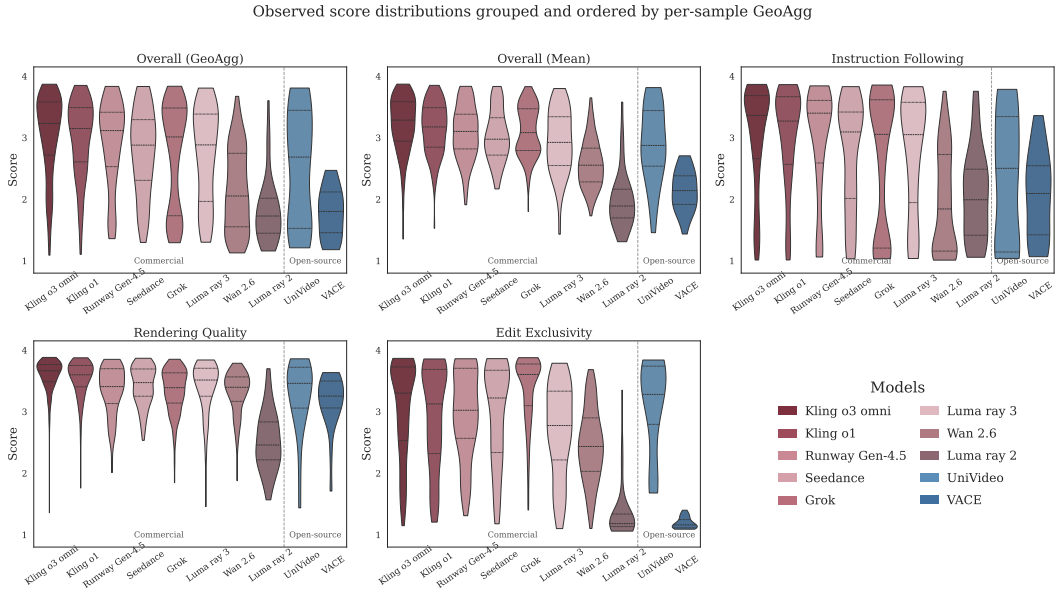

Figure A17: Violin plot of the ten-model evaluation, showing `VEFX-Reward-32B` score density for each model and dimension.

Table A13: Agreement with independent human overall editing-success ratings on 100 videos. The same dimension predictions are used for every aggregation; higher SRCC is better. Intervals are 95% confidence intervals from paired video bootstrap. The human target is the mean of three direct overall editing-success ratings.

Aggregation	SRCC↑	95% CI
Arithmetic Mean	0.681	[0.559, 0.774]
Equal-weight geometric mean	0.752	[0.652, 0.826]
GeoAgg $(2, 1, 1)$	**0.814**	[0.735, 0.871]