V2V-Bench: A Comprehensive Benchmark for Video-to-Video Generation Evaluation

Tao Liu [*1] **Leela Krishna** [*1] **Gouti Pavan Kumar** [1] **Sreeja K** [1] **Vishav Garg** [1]

Abstract

Video-to-video (V2V) generation is difficult to evaluate because outputs must both follow editing instructions and preserve frame-level correspondence with the source video, which existing T2V and I2V metrics do not capture. We introduce V2V-Bench, a 11-dimension benchmark organized into five categories: temporal alignment, structural fidelity, transformation quality, video quality, and semantic alignment. V2V-Bench pairs diverse source videos with challenging editing tasks and evaluates two commercial models, Grok Imagine and Gemini Veo3.1, and one open-source model, Open Sora 2. Results show complementary model strengths: Grok performs better on editing fidelity, while Veo3.1 achieves stronger visual quality. On six V2V-specific dimensions, V2V-Bench reaches a Spearman correlation of 0.905 with human judgments.

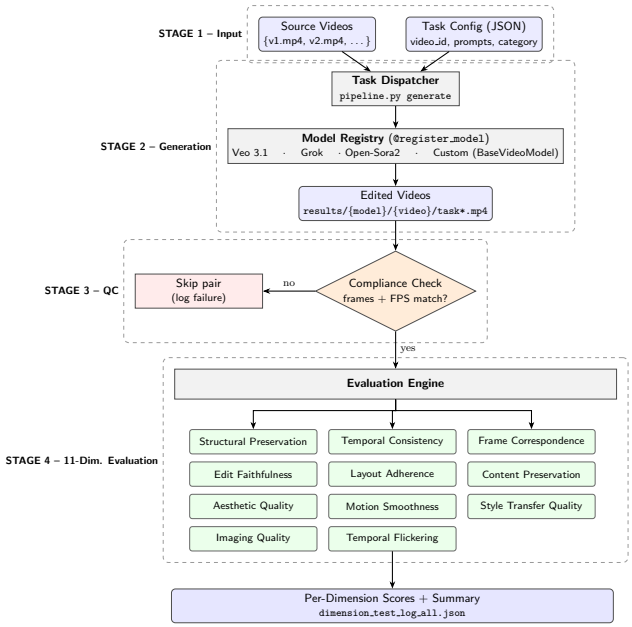

Figure 1. Overview of the V2V-Bench framework.

1. Introduction

Video-to-video generation has become an important paradigm for controllable video editing, where a model is required to transform a source video according to a target instruction while preserving its temporal structure, scene dynamics, and spatial relationships (Wang et al., 2018; Jeong et al., 2025; Hu & Xu, 2023). Although recent diffusion-based and autoregressive video models have made substantial progress in generating realistic motion and high-fidelity visual content, evaluating video-to-video generation remains challenging (HaCohen et al., 2024; Hu & Xu, 2023). Existing benchmarks largely emphasize perceptual quality, semantic relevance, or general video realism, but these criteria do not fully capture the core requirement of video-to-video transformation: maintaining fine-grained correspondence with the source video while faithfully applying the intended edit (Huang et al., 2024; Han et al., 2025; Zheng et al., 2025; Liu et al., 2023).

To address this gap, we introduce the V2V-Bench, a comprehensive benchmark for video-to-video generation. V2V-Bench provides a hierarchical and disentangled evaluation framework covering 11 fine-grained dimensions across temporal alignment, structural fidelity, transformation quality, video quality, and semantic alignment. Instead of reporting only aggregate scores, our benchmark offers interpretable diagnostics that reveal how models succeed or fail under different transformation requirements. V2V-Bench includes 81 curated source videos that span diverse scenes, motions, and visual content. Each video is paired with carefully designed editing tasks, including appearance editing, style transfer, scene modification, and content adaptation, enabling controlled comparison between shared source conditions and various transformation types. To validate the reliability of the benchmark, we further collect human judgment annotations and analyze their correlation with the proposed

[1]Centific Global Solutions Inc.. Correspondence to: Tao Liu <tao.liu@centific.com>, Leela Krishna <leela.krishna@centific.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

evaluation dimensions.

Using V2V-Bench, we systematically evaluate current video generation models across multiple editing scenarios and content categories. Our results show that while recent models often achieve strong perceptual quality, they still struggle to preserve source-video fidelity, maintain temporal consistency, and execute transformations with robust frame-level correspondence. These findings highlight the need for evaluation protocols that go beyond realism and semantic relevance to measure transformation correctness under source-video constraints.

Our contributions are as follows:

1) We introduce V2V-Bench, a benchmark specifically designed for video-to-video generation, which includes a hierarchical evaluation framework with 11 disentangled dimensions covering temporal alignment, structural fidelity, transformation quality, video quality, and semantic alignment.

2) We collect human judgment annotations to validate the alignment between benchmark scores and human perception.

2. Related Work

Video generation benchmarks. VBench (Huang et al., 2024) introduced multi-dimensional evaluation for T2V generation across 16 dimensions including subject consistency, motion smoothness, and aesthetic quality. VBench-I2V extended this to image-conditioned generation with 2 additional dimensions for first-frame preservation. EvalCrafter (Liu et al., 2024) proposed a similar multi-metric evaluation for T2V. However, all these benchmarks assume a *single-input* paradigm (text or image prompt) and lack the source-output temporal alignment constraints fundamental to V2V evaluation.

Video editing methods. The V2V generation landscape spans a diverse spectrum of approaches. Early methods include instruction-based editing (InstructPix2Pix (Brooks et al., 2023)), attention manipulation (Prompt-to-Prompt), and one-shot tuning (Tune-A-Video (Wu et al., 2023)). More recent advances focus on improving temporal consistency and controllability, such as feature propagation methods (TokenFlow (Geyer et al., 2023)) and control-signal guided frameworks (ControlVideo (Zhao et al., 2023)).

Beyond single-shot editing, recent works have begun to explore *frame-to-story* or narrative-driven video generation paradigms, where generation is conditioned on structured keyframes or story representations. For example, STAGE (Zhang et al., 2025) formulates multi-shot video generation as storyboard-anchored frame pair prediction to enforce long-range consistency, while StoryAnchors (Wang et al., 2025a) introduces a bidirectional framework for generating temporally coherent story frames across multiple scenes. These approaches emphasize global narrative coherence and cross-shot consistency, representing an emerging direction complementary to traditional V2V editing.

Commercial video generation models such as Grok, Runway, Gen-3, Kling, Pika, and Sora further demonstrate the practical importance of V2V capabilities (Wang et al., 2025b; Team et al., 2025). Despite these advances, existing methods make different trade-offs between edit faithfulness, temporal consistency, and source preservation, and a unified evaluation protocol for systematically comparing them remains lacking.

3. V2V-Bench Framework

The proposed V2V-Bench framework consists of four sequential stages, as illustrated in Figure 1.

Stage 1: Input Preparation. The pipeline takes (i) a set of source videos $\{v_1, v_2, \ldots\}$ and (ii) a JSON task configuration specifying video IDs, editing prompts, and task categories. A Task Dispatcher parses the inputs and schedules editing jobs.

Stage 2: Video Generation. Tasks are routed through a Model Registry that enables plug-and-play integration of heterogeneous video editing models (e.g., Veo-3.1, Grok, Open-Sora2, and user-defined models).

Stage 3: Quality Control. Prior to evaluation, we first compare the generated video with its corresponding source video in terms of frame count and frame rate (FPS). If either the frame count or FPS is inconsistent, the sample is marked as a failure, indicating that the model is unable to produce outputs with the required temporal length and frame consistency. These cases are still recorded for analysis. For all remaining compliant video pairs, we proceed with evaluation across the 11 defined dimensions.

Stage 4: Multi-Dimensional Evaluation. The 11 fine-grained evaluation dimensions are organized into five categories: temporal alignment, structural fidelity, transformation quality, video quality, and semantic alignment. Among these, six dimensions are specifically designed for video-to-video (V2V) evaluation: Frame Correspondence, Temporal Consistency, Structural Preservation, Layout Adherence, Edit Faithfulness, and Style Transfer Quality.

3.1. Compliance Check

Given a source video $\mathcal{V}_s = \{s_1, \ldots, s_T\}$ with frame rate r and a transformation prompt p, a video-to-video (V2V) model generates an output video $\mathcal{V}_o = \{o_1, \ldots, o_{T'}\}$. Unlike T2V or I2V settings, V2V requires preserving structural

correspondence between input and output over time.

We define three constraints: (i) **temporal length preservation**, $T' = T$; (ii) **frame rate consistency**, $\mathrm{FPS}(\mathcal{V}_o) = r$; and (iii) **frame-level correspondence**, enforcing a mapping $o_t \leftrightarrow s_t$ for all $t \in \{1, \ldots, T\}$. The first two are enforced via a pre-evaluation compliance check, while the third is assessed through downstream metrics such as temporal alignment and structural fidelity.

3.2. Evaluation Dimensions

V2V-Bench evaluates 11 dimensions across 5 categories: temporal alignment, structural fidelity, transformation quality, visual quality, and semantic alignment, with 6 novel dimensions (Frame Correspondence, Temporal Consistency, Structural Preservation, Layout Adherence, Edit Faithfulness, and Style Transfer Quality) specifically designed for V2V and 4 dimensions (Motion Smoothness, Aesthetic Quality, Imaging Quality, Temporal Flickering) reused from VBench.

3.2.1. Temporal Alignment

This category captures the core V2V requirement: each output frame must align temporally with its source counterpart while preserving motion dynamics.

Frame Correspondence. For each frame pair (s_t, o_t), we combine DINO ViT-B/16 (Caron et al., 2021) semantic features with SSIM:

$$S_{\mathrm{fc}} = \frac{1}{T} \sum_{t=1}^{T} \left[\alpha \cdot \cos(f_s^t, f_o^t) + (1-\alpha) \cdot \mathrm{SSIM}(s_t, o_t) \right] \quad (1)$$

where $f_s^t = \mathrm{DINO}(s_t)$, $f_o^t = \mathrm{DINO}(o_t)$, and $T = 8$ is the number of video frames used for evaluation. We set $\alpha = 0.7$ to prioritize semantic correspondence over pixel-level similarity, as V2V generation should preserve high-level content consistency rather than exact reconstruction, particularly for style transfer and appearance editing. DINO features (70%) capture semantic/object-level alignment, while SSIM (30%) provides complementary structural consistency. Empirically, rankings remain stable for $\alpha \in [0.65, 0.8]$, with $\alpha = 0.7$ offering a balanced trade-off between semantic robustness and structural sensitivity.

Temporal Consistency. We evaluate temporal consistency by measuring whether the generated video preserves the motion pattern of the source video. We compare the optical flow fields of the source and generated videos using relative endpoint error:

$$S_{\mathrm{temp}} = \exp\left(-\frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{E}\left[\frac{\|F_t^s - F_t^o\|_2}{\|F_t^s\|_2 + 1} \right] \right), \quad (2)$$

where F_t^s and F_t^o denote optical flow between adjacent frames in the source and generated videos.

3.2.2. Structural Fidelity

This category evaluates whether the output preserves the geometric structure and scene layout of the source.

Structural Preservation. Structural preservation evaluates whether object boundaries and spatial structures are retained after editing. We extract Canny edge maps from source and generated frames and compute an edge-level F1 score:

$$S_{\mathrm{struct}} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{F1}\left(E(I_t^s), E(I_t^o) \right), \quad (3)$$

where $E(\cdot)$ denotes edge extraction with spatial tolerance, and I_t^s, I_t^o are the source and generated video frames at time t, respectively.

Layout Adherence. Layout adherence measures whether the global spatial arrangement of the source video is preserved. We compute frame-level structural similarity between source and generated videos:

$$S_{\mathrm{layout}} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{SSIM}(I_t^s, I_t^o). \quad (4)$$

3.2.3. Transformation Quality

This category assesses whether the V2V model correctly executed the intended transformation.

Edit Faithfulness. Edit faithfulness measures how well the generated video follows the textual editing instruction. We compute CLIP image-text similarity between sampled generated frames and the prompt:

$$S_{\mathrm{edit}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\cos\left(f_{\mathrm{CLIP}}(I_t^o), f_{\mathrm{CLIP}}(p) \right) + 1}{2}. \quad (5)$$

Style Transfer Quality. For style-transfer tasks, we evaluate both the magnitude and direction of the style change. The magnitude is measured by VGG-19 Gram-matrix distance between source and generated frames, while the direction is measured by directional CLIP similarity:

$$S_{\mathrm{style}} = \lambda \cdot M_{\mathrm{Gram}} \cdot G_{\mathrm{dir}} + (1 - \lambda) \cdot D_{\mathrm{CLIP}}, \quad (6)$$

where M_{Gram} captures whether the style has changed and D_{CLIP} measures whether the change follows the target style. We assign a larger weight $\lambda = 0.6$ to Gram-based magnitude because style-transfer tasks should first exhibit a perceptible

style change, while directional CLIP ensures that the change follows the requested target style. The term G_{dir} penalizes edits moving away from the requested direction, defined as

$$G_{\text{dir}} = \begin{cases} 1.0, & \text{if } \cos(\Delta I, \Delta T) \geq 0, \\ 0.5, & \text{if } \cos(\Delta I, \Delta T) < 0, \end{cases} \tag{7}$$

where $\Delta I = I_{\text{out}} - I_{\text{src}}$ denotes the visual change direction in CLIP space, and $\Delta T = T_{\text{target}} - T_{\text{source}}$ denotes the desired textual style-change direction.

3.2.4. Video Quality

We reuse four well-validated VBench (Huang et al., 2024) dimensions that assess intrinsic video quality independent of the source: imaging quality, temporal flickering, aesthetic quality, and motion smoothness.

- **Motion Smoothness:** Optical flow acceleration magnitude; lower acceleration indicates smoother motion trajectories.
- **Aesthetic Quality:** LAION aesthetic predictor (Schuhmann et al., 2022) score reflecting color harmony, composition, and sharpness.
- **Imaging Quality:** BRISQUE no-reference quality assessment detecting blur, noise, and compression artifacts.
- **Temporal Flickering:** Mean inter-frame pixel difference; regions with natural fast motion are masked to avoid false positives.

3.2.5. Semantic Alignment

Content Preservation. Content preservation evaluates whether the primary visual content remains consistent with the source video. We use color-distribution similarity as a lightweight proxy:

$$S_{\text{content}} = \frac{1}{T} \sum_{t=1}^{T} \text{HistSim}(I_t^s, I_t^o), \tag{8}$$

where HistSim computes the average correlation between RGB-channel histograms.

3.3. Task Suite

The curated task suite contains 81 valid task instances spanning five edit video generation categories: object editing, appearance editing, style transfer, motion editing, and identity preservation. This design provides relatively even coverage of different editing objectives within a visually coherent set of human-motion videos.

In terms of video duration, the dataset mainly consists of short clips, with an average duration of approximately 8 seconds. For each video, we use Gemini-2.5-Flash to generate structured prompts. All prompts are grounded in the visual content of the source video.

Table 1. Compliance check results for the three evaluated V2V models.

Model	Pass	Pass Rate	Main Failure Mode
Grok	0 / 41	0.0%	Frame: 192 → 185
Veo-3.1	41 / 41	100.0%	None
Open-Sora-2	0 / 41	0.0%	Frame: 192 → 129

4. Experiments

4.1. Settings

We evaluate two commercial models, Veo-3.1 and Grok-Imagine-Video, and one open-source model, Open-Sora2, on 81 editing prompts spanning style transfer, appearance editing, environment editing, identity preservation, and related tasks.

Given the broad range of possibilities with video world model testing, we also constructed a focused subset of four videos with 10 detailed task prompts. For this subset, all video generation, human judgment, and VLM-based evaluations have been completed. All generation and evaluation experiments are conducted on a single NVIDIA H100 GPU.

4.2. Experiment 1

4.2.1. Compliance Check

The compliance check is critical for video-to-video evaluation because V2V generation requires frame-level correspondence between the source and generated videos. As shown in Table 1, Veo-3.1 satisfies this requirement for all 41 evaluated samples, while Grok and Open-Sora-2 fail all cases due to frame count mismatch. This failure is primarily attributed to their inability to consistently generate videos of sufficient temporal length, which further reflects their limitations in handling long-duration video synthesis tasks. As a result, the compliance check also serves as an indicator of a model's capability in maintaining temporal consistency over longer video generation horizons.

For Grok and Open-Sora-2, which do not pass the compliance check, we still include their outputs in the evaluation by performing frame-level alignment-based comparison, where metrics are computed over the overlapping generated frames only. This ensures a fair comparison under non-compliant generation settings while preserving evaluability across all models.

This highlights that compliance is not merely a formatting constraint, but a prerequisite for faithful V2V evaluation and for preserving the temporal structure of the input video.

Table 2. Three-model comparison over all 11 evaluation dimensions. Scores are averaged over 41 common video-editing tasks. Bold indicates the best and underlining indicates the second-best.

Dimension	Grok	Veo	Open-Sora
Imaging Quality	0.4979	**0.6522**	0.3031
Temporal Flickering	0.9836	**0.9856**	0.9814
Structural Preservation	**0.5726**	0.2926	0.2225
Temporal Consistency	**0.5289**	0.1752	0.3464
Frame Correspondence	**0.7895**	0.7118	0.6697
Edit Faithfulness	**0.6187**	0.6161	0.6105
Aesthetic Quality	**0.4981**	0.4976	0.4931
Motion Smoothness	0.9657	**0.9865**	0.9645
Layout Adherence	**0.7822**	0.6564	0.6692
Style Transfer Quality	**0.8660**	0.6903	0.6141
Content Preservation	0.6086	**0.7489**	0.4633
Mean	**0.7011**	0.6376	0.5762
Dimensions Won	**7 / 11**	4 / 11	0 / 11

Table 3. Comparison on the six V2V-Bench-specific dimensions.

V2V-Specific Dimension	Grok	Veo	Open-Sora
Structural Preservation	**0.5726**	0.2926	0.2225
Temporal Consistency	**0.5289**	0.1752	0.3464
Frame Correspondence	**0.7895**	0.7118	0.6697
Edit Faithfulness	**0.6187**	0.6161	0.6105
Layout Adherence	**0.7822**	0.6564	0.6692
Style Transfer Quality	**0.8660**	0.6903	0.6141
Mean	**0.6937**	0.5237	0.5221
Dimensions Won	**6 / 6**	0 / 6	0 / 6

4.2.2. Benchmark Results

Table 2 shows that Grok achieves the highest overall mean score and wins the most dimensions, despite similar performance among models on several general metrics such as temporal flickering, aesthetic quality, and motion smoothness. In contrast, Table 3 reveals larger gaps on the six V2V-specific dimensions, where Grok consistently outperforms the other models in preserving source-video structure, temporal correspondence, and edit faithfulness. This suggests that V2V-specific metrics are more discriminative than general video-quality metrics for evaluating video-to-video generation.

4.3. Experiment 2

We conduct a controlled study on four representative videos using 10 diverse editing tasks per video and evaluating across 11 fine-grained dimensions. Human judgments from three independent annotators assess how well the benchmark aligns with human preferences.

4.3.1. Human and VLM Preference Alignment

To validate that the proposed evaluation method can faithfully reflect human perception, we performed human annotation for each dimension. We show the correlation between V2V-Bench evaluation results and human preference annotations and two VLM models in Table 4, which shows that V2V-Bench aligns well with human judgments, achieving a Spearman correlation of 0.688 across all 11 dimensions and a much stronger correlation of 0.905 on the V2V-core subset. This indicates that the proposed V2V-specific dimensions better capture human preferences for source preservation, temporal correspondence, and edit fidelity than the full mixed dimensions. The two VLM judges are highly correlated with each other, but their agreement with humans is lower than that of V2V-Bench on the V2V-core dimensions.

Table 4. Spearman correlation between Benchmark/Human/GPT-4o/Gemini 2.5 Pro.

Pair	All 11 dims	V2V-core 6
Human $\leftrightarrow$ Bench	0.688	**0.905**
Human $\leftrightarrow$ Gemini 2.5 Pro	0.713	0.899
Human $\leftrightarrow$ GPT-4o	**0.737**	0.816
GPT-4o $\leftrightarrow$ Gemini 2.5 Pro	0.943	0.912
Bench $\leftrightarrow$ Gemini 2.5 Pro	0.578	0.826
Bench $\leftrightarrow$ GPT-4o	0.578	0.823

Due to space limitations, we present the main results here and defer additional experimental results and detailed analyses to the appendix.

5. Conclusion

We introduce V2V-Bench, a benchmark for video-to-video generation that decomposes quality into hierarchical, disentangled dimensions, each paired with tailored prompts and dedicated evaluation methods. The benchmark spans 11 fine-grained dimensions organized into 5 categories: temporal alignment, structural fidelity, transformation quality, video quality, and semantic alignment. We curate a diverse set of source videos that cover varied scenes, motions, and visual content, each paired with editing tasks that include appearance editing, style transfer, and other challenging transformations. To validate alignment with human perception, we collect human preference annotations on prompt-based tasks across two commercial models (Grok Imagine, Gemini Veo3) and one open-source model (Open Sora 2). Benchmark results show complementary strengths: Grok Imagine achieves higher editing fidelity, while Gemini Veo3 delivers stronger visual quality. In the six dimensions specific to V2V, our benchmark achieves a Spearman correlation of 0.905 with human judgments.

References

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Geyer, M., Bar-Tal, O., Bagon, S., and Dekel, T. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Deng, Y., Leong, C. T., Du, H., Fu, J., Li, Y., et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Hu, Z. and Xu, D. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jeong, H., Lee, S., and Ye, J. C. Reangle-a-video: 4d video generation as video-to-video translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11164–11175, 2025.

Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., and Hou, L. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36: 62352–62387, 2023.

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022.

Team, K., Chen, J., Ding, Y., Fang, Z., Gai, K., Gao, Y., He, K., Hua, J., Jiang, B., Lao, M., et al. Klingavatar 2.0 technical report. *arXiv preprint arXiv:2512.13313*, 2025.

Wang, B., Huang, H., Lu, Z., Liu, F., Ma, G., Yuan, J., Zhang, Y., Duan, N., and Jiang, D. Storyanchors: Generating consistent multi-scene story frames for long-form narratives. *arXiv preprint arXiv:2505.08350*, 2025a.

Wang, H., Zhang, G., and Yan, K. Based on runway gen-4: A dynamic video generation method for optimizing movie vfx workflows. In *2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC)*, pp. 146–151. IEEE, 2025b.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2023.

Zhang, P., Jia, Z., Liu, K., Weng, S., Li, S., and Shi, B. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*, 2025.

Zhao, M., Wang, R., Bao, F., Li, C., and Zhu, J. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2(3), 2023.

Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.-S., et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Table 5. Per-dimension scores on V2V-Bench (**higher is better**).

Dimension	**Veo-3.1**	**Grok**	**Open-Sora2**
Imaging Quality	0.346	**0.578**	0.248
Temporal Flickering	0.983	**0.987**	0.984
Aesthetic Quality	**0.607**	0.508	0.503
Motion Smoothness	**0.983**	0.976	0.970
Structural Preservation	0.435	**0.674**	0.305
Frame Correspondence	0.711	**0.829**	0.638
Layout Adherence	0.565	**0.743**	0.548
Content Preservation	0.540	**0.603**	0.362
Temporal Consistency	0.644	**0.675**	0.517
Edit Faithfulness	0.618	**0.623**	0.618
Style Transfer Quality	0.620	**0.638**	0.634
Mean (all 11)	0.639	**0.705**	0.595
Dimensions Won	2 / 11	9 / 11	0 / 11

Table 6. Inter-human agreement among the three annotators.

Raters	SA	Strict κ	DA	Decisive κ
A1 vs A2	0.554	0.362	0.783	0.571
A1 vs A3	0.546	0.351	0.786	0.576
A2 vs A3	0.960	0.940	0.982	0.964
Average	0.687	0.551	0.850	0.704

A. Additional Experimental Results

A.1. Benchmark Result

Table 5 shows per-dimension results on four videos with carefully designed prompt-based tasks in V2V-Bench. Grok still achieves the highest mean score and wins 9 of 11 dimensions, showing strong performance in structure preservation, temporal alignment, and edit fidelity. Veo-3.1 remains competitive on general video-quality metrics, while Open-Sora2 generally lags behind.

A.2. Inter-human Agreement

Table 6 shows pairwise agreement among the three annotators. Strict agreement (SA) treats ties as a separate label, while decisive agreement (DA) considers only cases where both annotators choose a non-tie model; Cohen's κ measures chance-corrected agreement. Decisive agreement is substantially higher, averaging 0.850 with an average κ of 0.704, indicating reliable human judgments when annotators make committed preferences.

A.3. Win Ratio

Given the human preference and VLM annotations, we compute the win ratio for each model on all the dimensions. In each pairwise comparison, a model receives a score of 1 if its generated video is preferred by annotators, while the other model receives a score of 0. In the case of a tie, both models receive a score of 0.5. The win ratio of each model is then defined as the total score accumulated across all pairwise comparisons divided by the total number of comparisons in which the model participates.

Table 7. Win ratios for every dimension. V2V-core dimensions are marked with †.

Dimension	Model	BENCH	Gemini 2.5 Pro	GPT-4o	Human
SP†	Veo	0.450	0.487	0.550	0.388
	Grok	**0.963**	**0.650**	**0.588**	**0.896**
	OpenSora	0.087	0.362	0.362	0.217
TC†	Veo	0.662	0.463	0.525	0.446
	Grok	**0.800**	**0.650**	**0.600**	**0.840**
	OpenSora	0.037	0.388	0.375	0.215
FC†	Veo	0.425	0.487	0.537	0.404
	Grok	**0.900**	**0.637**	**0.600**	**0.898**
	OpenSora	0.175	0.375	0.362	0.198
EF†	Veo	0.388	0.537	0.537	0.528
	Grok	**0.650**	**0.594**	**0.575**	**0.734**
	OpenSora	0.463	0.369	0.388	0.237
LA†	Veo	0.338	0.500	**0.562**	0.381
	Grok	**0.850**	**0.637**	0.550	**0.923**
	OpenSora	0.312	0.362	0.388	0.196
CP†	Veo	0.512	0.525	0.550	0.392
	Grok	**0.713**	**0.625**	**0.588**	**0.879**
	OpenSora	0.275	0.350	0.362	0.229
AQ	Veo	**1.000**	0.475	0.537	**0.789**
	Grok	0.463	**0.625**	**0.575**	0.475
	OpenSora	0.037	0.400	0.388	0.236
MS	Veo	**0.706**	0.450	0.525	**0.665**
	Grok	0.381	**0.662**	**0.600**	0.640
	OpenSora	0.412	0.388	0.375	0.196
STQ	Veo	0.000	0.475	0.537	**0.527**
	Grok	0.625	**0.637**	**0.575**	0.502
	OpenSora	**0.875**	0.388	0.388	0.471
IM	Veo	0.412	0.438	0.537	**0.843**
	Grok	**0.850**	**0.650**	**0.575**	0.406
	OpenSora	0.237	0.412	0.388	0.250
TF	Veo	0.338	0.388	0.525	0.415
	Grok	**0.675**	**0.675**	**0.588**	**0.588**
	OpenSora	0.487	0.438	0.388	0.498

Table 7 reports the pairwise win ratios for each model across evaluation dimensions and annotation sources. On the V2V-core dimensions, both BENCH and human judgments consistently favor Grok, especially for structural preservation, frame correspondence, layout adherence, and content preservation, indicating stronger source-video fidelity and temporal alignment. In contrast, the general quality dimensions show more source-dependent variation: Veo performs strongly in human judgments for aesthetic quality, motion smoothness, and imaging quality, while Grok remains preferred by the automatic benchmark on several perceptual metrics. The VLM judges generally produce

Table 8. Prompts for the four task types.

#	Task Type	Prompt
2	Replace Object	Replace the large yellow surfboard held by the man with an identical-shaped deep ocean blue surfboard.
4	Change Lighting	Transform the office lighting from bright neutral daylight to warm golden-hour evening lighting.
7	Change Background	Replace the dark, dimly lit office background with a bright modern open-plan office environment.
5	VFX (Visual Effects)	Apply a vivid bioluminescent ocean glow effect to the wave in the surf video.

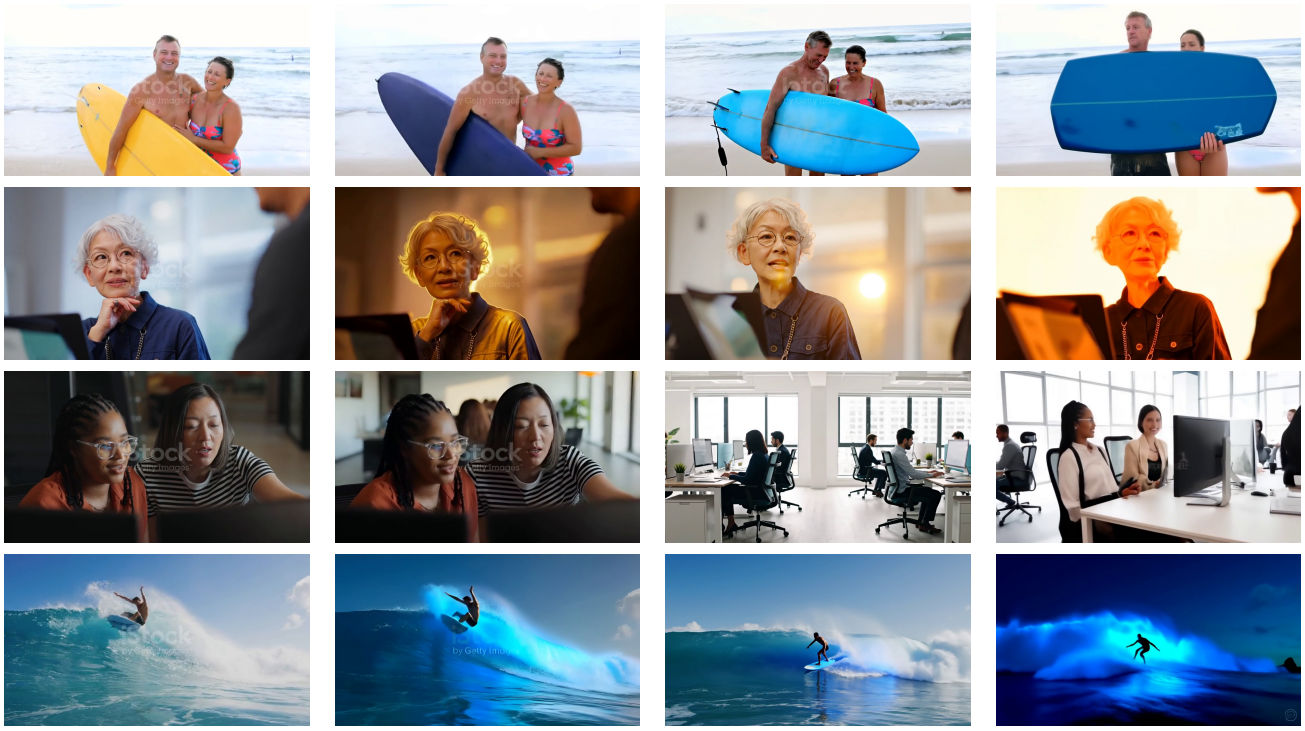

Figure 2. Qualitative comparison across different tasks. Columns from left to right show Raw, Grok, Veo-3.1, and Open-Sora2. Rows from top to bottom correspond to Tasks 2, 4, 7, and 5.

more compressed win-ratio distributions than human judgments and BENCH, suggesting that they are less discriminative in separating model quality. Overall, the table shows that V2V-Bench closely reflects human preferences on the V2V-specific dimensions where source preservation and edit fidelity are most critical.

A.4. Visual Result

Figure 2 shows qualitative comparisons across tasks. From left to right, we present the 120th frame of the source video and the outputs from Grok, Veo-3.1, and Open-Sora-2. Rows correspond to Replace Object, Change Lighting, Change Background, and VFX. The prompts for each task are in Table 8.

Grok consistently preserves structural fidelity, maintaining spatial relationships and human pose while applying edits. In contrast, Veo-3.1 and Open-Sora-2 exhibit structural drift in background and VFX tasks, often deviating significantly from the source. For object replacement, Veo-3.1 modifies the object but alters pose, while Open-Sora-2 introduces larger deviations in both pose and expression.