

LOVE: Benchmarking and Evaluating Text-to-Video Generation and Video-to-Text Interpretation

Jiarui Wang¹, Huiyu Duan^{1,2}, Ziheng Jia¹, Yu Zhao¹, Woo Yi Yang¹, Zicheng Zhang¹,

Zijian Chen¹, Juntong Wang¹, Yuke Xing¹, Guangtao Zhai^{1,2}, Xiongkuo Min^{1*}

¹Institute of Image Communication and Network Engineering,

²MoE Key Lab of Artificial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China

Abstract

Recent advancements in large multimodal models (LMMs) have driven substantial progress in both text-to-video (T2V) generation and video-to-text (V2T) interpretation tasks. However, current AI-generated videos (AIGVs) still exhibit limitations in terms of perceptual quality and text-video alignment. Therefore, a reliable and scalable automatic model for AIGV evaluation is desirable, which heavily relies on the scale and quality of human annotations. To this end, we present **AIGVE-60K**, a comprehensive dataset and benchmark for AI-Generated Video Evaluation, which features **(i) comprehensive tasks**, encompassing 3,050 extensive prompts across 20 fine-grained task dimensions, **(ii) the largest human annotations**, including 120K mean-opinion scores (MOSS) and 60K question-answering (QA) pairs annotated on 58,500 videos generated from 30 T2V models, and **(iii) bidirectional benchmarking and evaluating** for both T2V generation and V2T interpretation capabilities. Based on AIGVE-60K, we propose **LOVE**, a LMM-based metric for AIGV Evaluation from multiple dimensions including perceptual preference, text-video correspondence, and task-specific accuracy in terms of both instance level and model level. Comprehensive experiments demonstrate that LOVE not only achieves state-of-the-art performance on the AIGVE-60K dataset, but also generalizes effectively to a wide range of other AIGV evaluation benchmarks. These findings highlight the significance of the AIGVE-60K dataset. Database and codes are available at <https://github.com/IntMeGroup/LOVE>.

1 Introduction

The rapid advancement of large multimodal models (LMMs) has revolutionized the fields of both text-to-video (T2V) generation [1–3] and video-to-text (V2T) interpretation [4–6], leading to high-quality video generation and comprehensive multimodal video understanding capabilities. However, state-of-the-art T2V models may still produce videos with degraded **perceptual quality** and limited **text-video correspondence**, thus may fail to meet human preferences [7–9]. Given the high cost and inefficiency of human evaluation, it is of great significance to develop a reliable and scalable evaluation metric that aligns well with human preferences for AI-generated videos (AIGVs) and corresponding T2V models.

To fairly and effectively evaluate T2V models and AIGVs, many T2V model benchmarks and AIGV evaluation datasets [7–18] have been constructed as shown in Table 1, and many AIGV evaluation metrics have been proposed [19–22]. However, these efforts face the following limitations that may

*Corresponding author

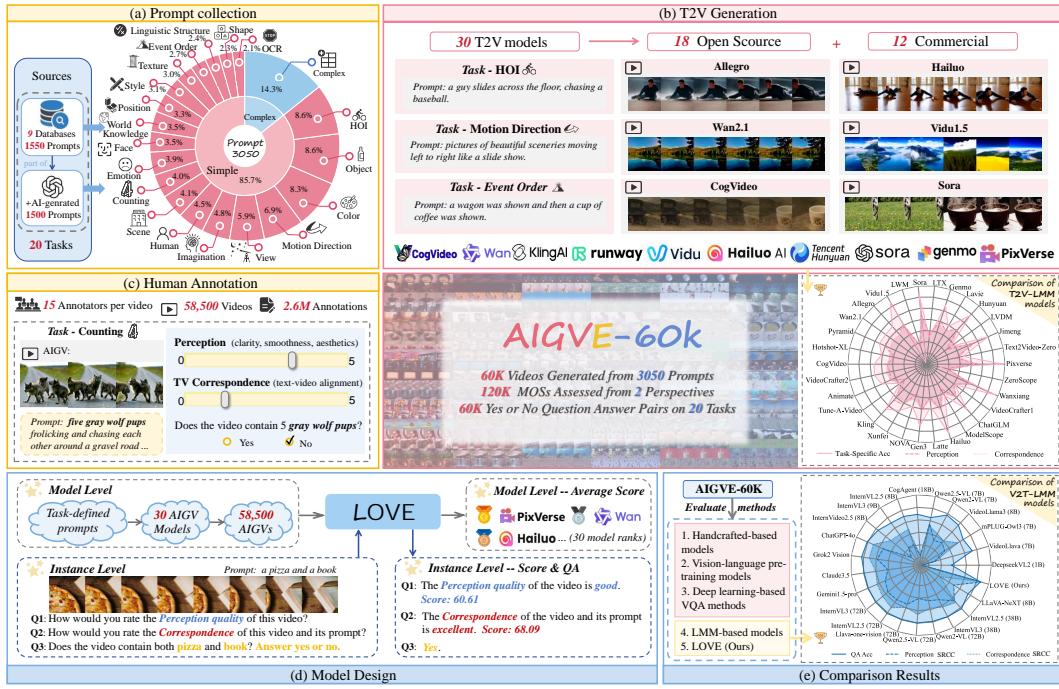


Figure 1: We present the largest AIGV evaluation database (**AIGVE-60K**) and a novel model (**LOVE**). (a) We collect 3,050 prompts across 20 fine-grained tasks. (b) 30 T2V models are applied to generate 60K videos. (c) 120K MOSs from 2 perspectives and 60K question-answer pairs are acquired from annotators. (d) We design a model to evaluate **T2V generation** at both the instance and model levels. (e) Comparison results of LMM’s **V2T interpretation** ability based on AIGVE-60K.

affect their effectiveness in diverse applications. (1) **Most benchmarks or datasets only consider either perception, correspondence, or task-specific accuracy dimensions, while comprehensive subjective evaluation works are still lacking.** Since high-quality AIGVs may exhibit poor text-video alignment, well-aligned AIGVs may suffer from low perceptual quality [7], and we also need a binary true-or-false metric in some scenarios (such as number-based generation scenes) [10, 23], an extensive evaluation is important. However, some existing metrics only focus on one dimension [8] or use the fused overall evaluation [12, 16]. (2) **The scale of the datasets remains small and annotations remain coarse.** Some datasets or benchmarks include only a limited number of T2V models [8–10, 12] or AIGVs [11, 14–16], which constrains the ability to validate the effectiveness and scalability of evaluation methods across diverse models and outputs, and some works use coarse-MOSs [11, 13], which do not meet ITU standards [24] and may produce invalid MOSs. (3) **Current evaluation metrics for T2V generation consider from either the instance level or the model level, while a comprehensive study integrates both perspectives remains lacking.** For example, Inception Score (IS) [25], Fréchet Video Distance (FVD) [26], VBench [9], etc., are mainly designed to evaluate T2V models, *i.e.*, from a model level perspective, while LGVQ [15], AIGV-Assessor [7], etc., are mainly designed for the individual AIGV evaluation or multiple AIGV comparison, *i.e.*, from a instance level perspective. To the best of our knowledge, no existing work has jointly considered both perspectives towards a comprehensive evaluation.

To address these challenges, we present **AIGVE-60K**, a large-scale dataset and benchmark for **AI-Generated Video Evaluation**, which includes 58,500 videos generated by 30 state-of-the-art T2V models using 3,050 diverse prompts across 20 task-specific challenges. As shown in Figure 1, we collect **2.6M** human annotations from the perception, text-video correspondence, and task-specific accuracy, respectively, and finally obtain 120K mean opinion scores (MOSS) and 60K question-answering (QA) pairs. Based on AIGVE-60K, we propose **LOVE**, a LMM-based metric for AIGV Evaluation from multiple dimensions at both instance level and model level, which integrates: (1) dual encoders for vision-temporal feature extraction, (2) a large-language model (LLM) backbone for ***all-in-one*** video quality assessment, and (3) instruction tuning techniques [27] for accurate response generation. Through extensive experimental validation, we demonstrate that LOVE achieves state-of-the-art performance on the AIGVE-60K dataset and manifests strong zero-shot generalization ability on other benchmarks. Our contributions are summarized as follows:

Table 1: Comparison of T2V model evaluation benchmarks and AIGV quality evaluation databases.

Database	Annotation Type (People per Sample)	Videos	Prompts	Annotations	Models	Evaluation Concern			
						T2V Tasks	Perception	T2V Correspondence	QA Acc
VBench [9]	Pairs (1)	3,200	800	24,000	4	16	✓	✓	✗
VBench2.0 [10]	Pairs (1)	100,800	1,260	151,200	4	18	✗	✓	✓
EvalCrafter [11]	Coarse-MOS (7)	2,500	700	70,000	8	17	✓	✓	✗
FETV [12]	Coarse-MOS (3)	2,476	619	7,428	4	5	<i>Overall</i>		
GenAI-Bench [8]	Coarse-MOS (3)	32,000	800	9,600	4	8	✗	✓	✗
Q-Eval [13]	Coarse-MOS (3)	40,000	2,500	384,000	16	16	✓	✓	✓
MQT [14]	Fine-MOS (24)	1,005	201	48,240	5	5	✓	✓	✓
LGVQ [15]	Fine-MOS (20)	2,808	468	168,480	6	6	✗	✓	✗
T2VQA-DB [16]	Fine-MOS (27)	10,000	1,000	270,000	9	9	✗	✓	✗
AIGVQA-DB [7]	Fine-MOS (20) & Pairs (3)	36,576	2,048	371,520	15	15	✗	✓	✓
AIGVE-60K (Ours)	Fine-MOS (15)	58,500	3,050	2,632,500	30	20	✓	✓	✓

- We present **AIGVE-60K**, the largest **text-to-video evaluation dataset** so far that contains 58,500 generated videos with 2.6M subjective ratings from the perception, text-video correspondence, and task-specific accuracy, respectively.
- We introduce a **bidirectional benchmarking and evaluation strategy**. Based on AIGVE-60K, we can benchmark the **T2V generation ability** of 30 T2V models, and the **V2T interpretation ability** of 23 LMMs and 24 VQA metrics, respectively.
- We propose **LOVE**, a novel LMM-based evaluation model capable of assessing both the perceptual quality and T2V alignment for AIGVs. Extensive experimental results on AIGVE-60K and other AIGV benchmarks manifest the state-of-the-art performance and strong generalization ability of LOVE.

2 Related Works

2.1 Benchmarks for T2V Generation

As shown in Table 1, the development of T2V generation has spawned many T2V model evaluation benchmarks and VQA databases, which can be categorized into pairs, coarse MOS, and fine-MOS based on the annotation method and granularity. VBench [9] and VBench2.0 [10] focus on video pairs comparison, but are limited in T2V comparison model number and lack precise quality assessment for each AIGV. Fine-MOS databases offer more reliable assessments derived from more than 15 annotators, following the guidelines of ITU-R BT.500 [24]. MQT [14], LGVQ [15] collect fine-grained MOSs but the number of AIGVs is limited. AIGVQA-DB [7] considers both perceptual quality and T2V correspondence, however, it mainly focuses on the pair comparison and lacks task-specific QA pairs, limiting their ability to assess T2V generation across diverse tasks. AIGVE-60K stands out by its largest scale of annotations, providing fine-grained MOSs for both perceptual quality and T2V correspondence, and answer annotations for task-specific questions.

2.2 Evaluation Metrics for T2V Generation

Many quality assessment models have been proposed in the literature [8, 28–34], including hand-crafted models (*e.g.*, NIQE [28], QAC [35], BRISQUE [36]) and deep learning-based VQA models (*e.g.*, BVQA [33], FAST-VQA [21], DOVER [22]). These models characterize quality-aware information to predict perception quality scores but can not evaluate T2V correspondence, which is crucial for assessing the relationship between the generated video and its corresponding text prompt. PickScore [34] and VQAScore [8] improve the evaluation of the T2V correspondence, but they struggle to assess the perception quality of AIGV. LMMs with visual understanding capabilities perform well in QA tasks, but their ability to assess image perceptual quality remains limited and often fail to give precise quality scores. VBench [9] employs various detection models for task-specific accuracy, but this approach is quite complex. To address this gap, our proposed LMM-based model complies with an *all-in-one* framework, which can evaluate quality scores and task-specific accuracies in one model.

3 AIGVE-60K Dataset & Benchmark

In this section, we introduce the construction process of AIGVE-60K and **benchmark T2V models** based on the dataset.



Figure 2: Video examples generated by 30 T2V models using prompt: “*a cat is drinking a cup of water*”, annotated with MOSs from 2 dimensions: **perceptual quality** and **text-video correspondence**.

3.1 Data Collection

Prompts of the AIGVE-60K are primarily sourced from 9 existing open-domain text-video pair datasets and some are refined using DeepSeek R1 [37] to expand and modify them, ensuring clarity and diversity. Our prompt design focuses on 20 different tasks as shown in Figure 1(a). The complex tasks are designed by combining simpler task components, such as motion direction, event order, and counting, into more complex challenges. In total, we collect 3,050 prompts, each corresponding to a specific task. To generate the AIGVs, we utilize 30 of the latest T2V models, as shown in Figure 1(b). We leverage open-source website APIs or the default weights of these models to generate videos. For the training set, we employ 2,750 distinct prompts, each processed by 18 open-source models. The test set consists of 300 unique prompts generated using all 30 models. With 3,050 distinct prompts, this process results in a total of 58,500 videos (2,750 prompts \times 18 open-source models + 300 prompts \times 30 open-source and close-source models). The imbalance in the number of T2V models used between the training and testing sets is attributed to two factors: (1) generating videos using close-source tools is costly, and (2) we aim to evaluate the scalability of evaluation metrics on training-set unseen generation models. More details in Appendix Sections C and D.

3.2 Subjective Experiment Setup and Procedure

Due to the unique distortions in AIGVs and varying elements determined by different text prompts, relying solely on an overall score for evaluation is inadequate. In this paper, we propose to evaluate AIGVs across two dimensions, as shown in Figure 2. (1) **Perceptual quality** focuses on visual perception, evaluating factors such as detail richness, motion smoothness, color vibrancy, and distortion levels. (2) **Text-video correspondence** evaluates how accurately the generated video reflects the objects, scenes, styles, and details described in the text prompt. We use a 1-5 Likert scale to score the videos based on the perception and T2V correspondence. For the correspondence evaluation, in addition to the rating, annotators are instructed to answer task-specific yes/no questions to determine whether the video consistently aligns with the prompt. Finally, we obtain a total of 2,632,500 human annotations including 1,755,000 reliable score ratings (15 annotators \times 2 dimensions \times 58,500 videos), and 877,500 task-specific QA pairs (15 annotators \times 58,500 videos).

3.3 Subjective Data Processing

In order to obtain the MOS for an AIGV, we first convert the raw ratings into Z-scores, and then linearly scale them to the range [0, 100] as follows:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \quad \mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} r_{ij}, \quad \sigma_i = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (r_{ij} - \mu_{ij})^2}, \quad (1)$$

where r_{ij} is the raw rating given by the i -th subject to the j -th video. N_i is the number of videos judged by subject i . Next, the MOS of the j -th video is computed by averaging the rescaled z-scores across all subjects as follows:

$$\text{MOS}_j = \frac{1}{M} \sum_{i=1}^M z'_{ij}, \quad (2)$$

where MOS_j indicates the MOS for the j -th AIGV, M is the number of subjects, and z'_{ij} are the rescaled z-scores. The task-specific yes/no answer is determined by the most votes. Therefore, a total of 117,000 MOSs (2 dimensions \times 58,500 videos) and 58,500 question answering pairs are obtained.

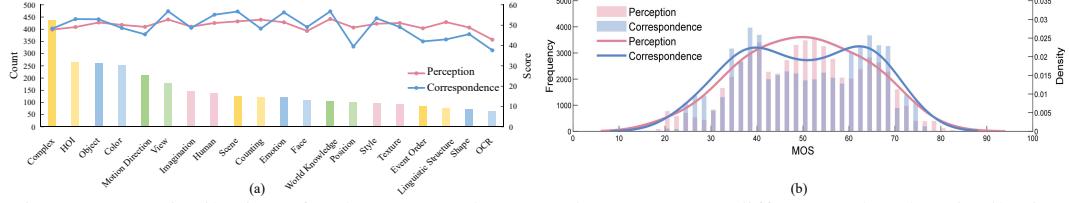


Figure 3: (a) Distribution of task counts and averaged scores across different tasks. (b) Distribution of perception and correspondence MOSS.

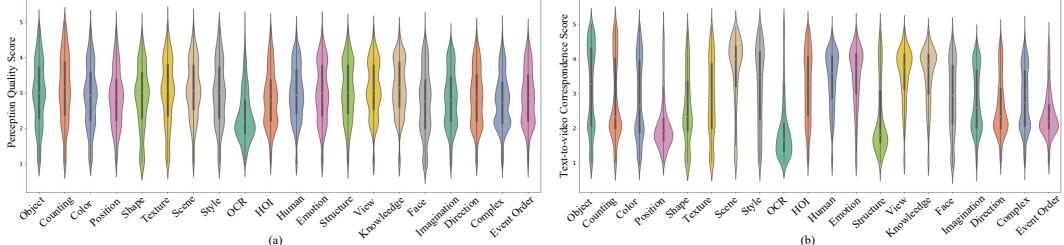


Figure 4: The MOS distribution in terms of different task contents. (a) MOS distribution of perception quality (b) MOS distribution of T2V correspondence.

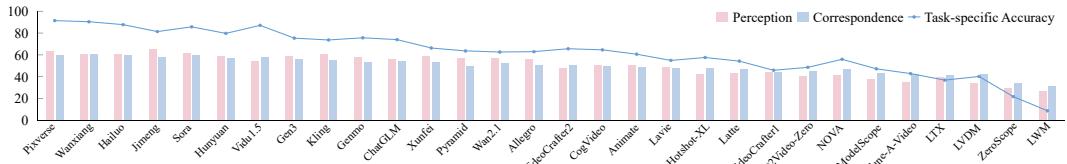


Figure 5: Comparison of T2V generation models regarding the perception MOSSs, correspondence MOSSs, and task-specific accuracy.

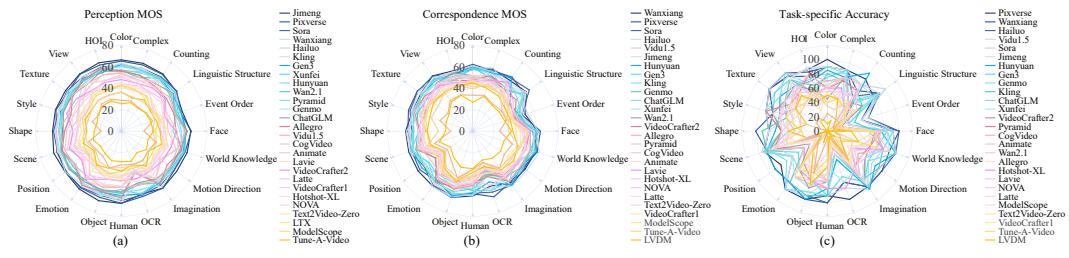


Figure 6: Comparison of MOSSs and task-specific accuracy of 30 T2V generation models across 20 tasks with descending order arranged in legend. (a) Results in terms of perception MOSSs. (b) Results in terms of correspondence MOSSs. (c) Results in terms of task-specific accuracy.

3.4 Subjective Data Analysis & T2V Model Benchmark

The distribution of task counts and averaged scores is shown in Figure 3(a). It can be observed that the correspondence score fluctuates more than the quality score, which means that the text-video alignment is more sensitive to the tasks. The distribution of MOSSs for both T2V correspondence and perceptual quality is shown in Figure 3(b), which approximately follows the Gaussian distributions. Figure 4 displays the MOS distribution for each task, demonstrating notable differences in model capabilities between tasks, especially for T2V correspondence. Moreover, we also launch comparisons for T2V generation models across different tasks based on perceptual quality MOSSs, T2V correspondence MOSSs, and task-specific accuracy, as shown in Figure 5. Jimeng [38] achieves a higher perceptual quality than correspondence, while Videl1.5 [39] shows the inverse trend. We further analyze the MOSSs and task-specific accuracies across 20 different tasks. As shown in Figure 6(a), perception MOS shows particular sensitivity to OCR tasks, where video clarity directly impacts character recognition accuracy. Figure 6(b) and (c) display similar trends, with task-specific accuracy results exhibiting sharper distinctions, which manifests the higher discriminative of the task-specific accuracy perspective. While task-specific accuracy delivers binary (0/1) evaluations, MOS offers continuous scoring, enabling more granular evaluation of T2V correspondence. Moreover, for the task of event order, most models show poor performance, indicating fundamental limitations in current models' ability to generate temporally coherent narratives. Finally, the top-performing models vary across perceptual quality, text-video correspondence, and task-specific accuracy, underscoring the importance of evaluating AIGVs from these three perspectives.

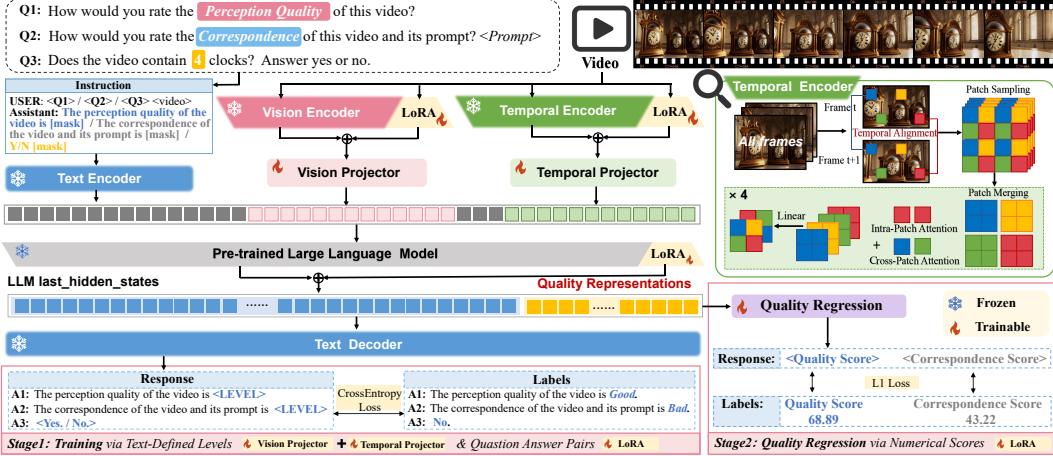


Figure 7: Overview of the LOVE architecture. The model includes two functions: (1) text-defined quality level and score prediction, (2) task-specific visual question answering. The training process consists of two stages: training via text-defined levels, and quality regression via numerical scores. The model incorporates a vision encoder, a temporal encoder, and a text encoder for extracting visual and textual features, which are fed into a pre-trained LLM to generate results. LoRA [40] weights are introduced to the pre-trained image encoder and the LLM to adapt the models to perception quality and T2V correspondence evaluation, and 20 task-specific visual question-answering tasks.

4 The LOVE Method

In this section, we present an *all-in-one* video quality assessment method **LOVE** to identify quality degradation levels, predict perception and T2V correspondence scores, and deliver visual question answers within a unified model.

4.1 Model Structure

Visual and Temporal Encoding. Figure 7 illustrates that the visual encoding component comprises a vision encoder and a temporal encoder for feature extraction, along with two projectors for aligning the video features with the input of the large language model (LLM). The vision encoder is constructed upon a pre-trained vision transformer (ViT), *i.e.*, InternViT [41]. To improve the scalability of processing high-resolution videos, we use a pixel unshuffle method that decreases the number of visual tokens to a quarter of their initial size. For the temporal encoder, we decompose all video frames $\{F^{(n)}\}_{n=1}^{N_v}$ into temporally aligned mini-patch map through grid mini-patch sampling, following [21]. For each video frame $F^{(n)}$, we first split it into uniform $L \times L$ grids, the set of grids $G^{(n)}$ can be described as:

$$G^{(n)} = \{g_{0,0}^{(n)}, \dots, g_{i,j}^{(n)}, \dots, g_{L,L}^{(n)}\}, \quad g_{i,j}^{(n)} = F^{(n)}[\frac{i \times H}{L} : \frac{(i+1) \times H}{L}, \frac{j \times W}{L} : \frac{(j+1) \times W}{L}] \quad (3)$$

where $g_{i,j}^{(n)} \in \mathbb{R}^{\frac{H}{L} \times \frac{W}{L} \times 3}$ denotes the grid in the i -th row and j -th column of $F^{(n)}$. Then we sample the mini-patches from each $g_{i,j}^{(n)}$ and splice all the selected mini-patches to get the mini-patch map $M \in \mathbb{R}^{H \times W \times 3}$, which are then fed into a Swin-T [42] with four hierarchical self-attention layers as the backbone. To align the extracted features with the input space of the LLM, a vision projector and a temporal projector with two multilayer perceptron (MLP) layers are applied.

Feature Fusion and Quality Regression. We utilize the InternLM3-9B-instruct [43] to integrate the visual tokens and text instruction tokens to perform the following two tasks. (1) Prediction of text-defined quality level: the model produces an evaluation of the input video’s quality level, such as “*The perception quality of the video is (bad, poor, fair, good, excellent).*” A preliminary sense of the video quality is provided by text-defined categorization, which is useful for directing later quality regression tasks because LLMs comprehend textual data better than numerical data. (2) Regression score output: the model takes the quality representations from the last hidden states of the LLM to perform regression through a quality regression module, outputting numerical quality scores.

Limitations and Broader Impact.

The current rankings are based on data we obtained from random-selected professional annotators, and we do not intend to offend the developers of these excellent T2V and V2T models. Although our model shows promising scalability in evaluating AIGVs generated by new prompts and previously unseen T2V models, the effectiveness in real-world applications remains an open question. We expect our benchmark and dataset will contribute to the advancement of T2V generation, T2V evaluation, and V2T interpretation.

References

- [1] A. Singh, “A survey of ai text-to-image and ai text-to-video generators,” in *Proceedings of the International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pp. 32–36, IEEE, 2023.
- [2] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, “A survey on video diffusion models,” *ACM Computing Surveys (Acm Comput Surv)*, vol. 57, no. 2, pp. 1–42, 2024.
- [3] M. Liao, Q. Ye, W. Zuo, F. Wan, T. Wang, Y. Zhao, J. Wang, X. Zhang, et al., “Evaluation of text-to-video generation models: A dynamics perspective,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 109790–109816, 2024.
- [4] P. Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, W. Zhang, H. Yan, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang, “Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition,” *arXiv preprint arXiv:2309.15112*, 2023.
- [5] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, “mplug-owl3: Towards long image-sequence understanding in multi-modal large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [6] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, “Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models,” *arXiv preprint arXiv:2407.07895*, 2024.
- [7] J. Wang, H. Duan, G. Zhai, J. Wang, and X. Min, “Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [8] B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, X. Xia, P. Zhang, G. Neubig, and D. Ramanan, “Evaluating and improving compositional text-to-visual generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5290–5301, 2024.
- [9] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, “VBench: Comprehensive benchmark suite for video generative models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [10] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, Y. Zhang, J. He, W.-S. Zheng, Y. Qiao, et al., “Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness,” *arXiv preprint arXiv:2503.21755*, 2025.
- [11] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, “Evalcrafter: Benchmarking and evaluating large video generation models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024.
- [12] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou, “Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.
- [13] Z. Zhang, T. Kou, S. Wang, C. Li, W. Sun, W. Wang, X. Li, Z. Wang, X. Cao, X. Min, et al., “Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content,” *arXiv preprint arXiv:2503.02357*, 2025.
- [14] I. Chivileva, P. Lynch, T. E. Ward, and A. F. Smeaton, “Measuring the quality of text-to-video model outputs: Metrics and dataset,” *arXiv preprint arXiv:2309.08009*, 2023.
- [15] Z. Zhang, X. Li, W. Sun, J. Jia, X. Min, Z. Zhang, C. Li, Z. Chen, P. Wang, Z. Ji, et al., “Benchmarking aigc video quality assessment: A dataset and unified model,” *arXiv preprint arXiv:2407.21408*, 2024.

- [16] T. Kou, X. Liu, Z. Zhang, C. Li, H. Wu, X. Min, G. Zhai, and N. Liu, “Subjective-aligned dataset and metric for text-to-video quality assessment,” in *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pp. 7793–7802, 2024.
- [17] Z. Chen, W. Sun, Y. Tian, J. Jia, Z. Zhang, J. Wang, R. Huang, X. Min, G. Zhai, and W. Zhang, “Gaia: Rethinking action quality assessment for ai-generated videos,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 40111–40144, 2024.
- [18] W. Y. Yang, J. Wang, S. Wu, H. Duan, Y. Zhu, L. Yang, K. Fu, G. Zhai, and X. Min, “Lmme3dhf: Benchmarking and evaluating multimodal 3d human face generation with lmms,” *arXiv preprint arXiv:2504.20466*, 2025.
- [19] D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” in *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2019.
- [20] W. Sun, X. Min, W. Lu, and G. Zhai, “A deep learning based no-reference quality assessment model for ugc videos,” in *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pp. 856–865, 2022.
- [21] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, “Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 538–554, Springer, 2022.
- [22] H. Wu, E. Zhang, L. Liao, C. Chen, J. H. Hou, A. Wang, W. S. Sun, Q. Yan, and W. Lin, “Exploring video quality assessment on user generated contents from aesthetic and technical perspectives,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.
- [23] D. Ghosh, H. Hajishirzi, and L. Schmidt, “Geneval: An object-focused framework for evaluating text-to-image alignment,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 52132–52152, 2023.
- [24] B. Series, “Methodology for the subjective assessment of the quality of television pictures,” *Recommendation ITU-R BT*, pp. 500–13, 2012.
- [25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [26] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018.
- [27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 34892–34916, 2023.
- [28] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters (SPL)*, vol. 20, no. 3, pp. 209–212, 2012.
- [29] L. Yang, H. Duan, L. Teng, Y. Zhu, X. Liu, M. Hu, X. Min, G. Zhai, and P. L. Callet, “Aigcoiq2024: Perceptual quality assessment of ai generated omnidirectional images,” in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 1239–1245, 2024.
- [30] H. Duan, Q. Hu, J. Wang, L. Yang, Z. Xu, L. Liu, X. Min, C. Cai, T. Ye, X. Zhang, et al., “Finevq: Fine-grained user generated content video quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [31] J. Wang, H. Duan, G. Zhai, and X. Min, “Quality assessment for ai generated images with instruction tuning,” *arXiv preprint arXiv:2405.07346*, 2025.
- [32] Z. Xu, H. Duan, G. Ma, L. Yang, J. Wang, Q. Wu, X. Min, G. Zhai, and P. L. Callet, “Harmonyiq: Pioneering benchmark and model for image harmonization quality assessment,” *arXiv preprint arXiv:2501.01116*, 2025.
- [33] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, “Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 9, pp. 5944–5958, 2022.
- [34] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, “Pick-a-pic: An open dataset of user preferences for text-to-image generation,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 36652–36663, 2023.

- [35] W. Xue, L. Zhang, and X. Mou, “Learning without human scores for blind image quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 995–1002, 2013.
- [36] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [37] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [38] B. Team, “Jimeng ai.” <https://jimeng.jianying.com/>, 2024. Accessed: 2025-03-08.
- [39] V. A. Team, “Vidu ai.” <https://www.vidu.studio/zh>, 2024. Accessed: 2025-03-08.
- [40] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., “Lora: Low-rank adaptation of large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, vol. 1, p. 3, 2022.
- [41] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al., “Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling,” *arXiv preprint arXiv:2412.05271*, 2024.
- [42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 10012–10022, 2021.
- [43] W. Wang, Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, J. Zhu, X. Zhu, L. Lu, Y. Qiao, and J. Dai, “Enhancing the reasoning ability of multimodal large language models via mixed preference optimization,” *arXiv preprint arXiv:2411.10442*, 2024.
- [44] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, “Blind image quality estimation via distortion aggravation,” *IEEE Transactions on Broadcasting (TBC)*, vol. 64, no. 2, pp. 508–517, 2018.
- [45] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, “Blind quality assessment based on pseudo-reference image,” *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 8, pp. 2049–2062, 2017.
- [46] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, “Blind image quality assessment based on high order statistics aggregation,” *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 9, pp. 4444–4457, 2016.
- [47] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” *arXiv preprint arXiv:2104.08718*, 2021.
- [48] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proceedings of the International conference on machine learning (ICML)*, pp. 12888–12900, 2022.
- [49] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., “Laion-5b: An open large-scale dataset for training next generation image-text models,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 25278–25294, 2022.
- [50] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, “Imagereward: Learning and evaluating human preferences for text-to-image generation,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 15903–15935, 2023.
- [51] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, “Human preference score: Better aligning text-to-image models with human preference,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2096–2105, 2023.
- [52] S. Han, H. Fan, J. Fu, L. Li, T. Li, J. Cui, Y. Wang, Y. Tai, J. Sun, C. Guo, and C. Li, “Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation,” *arXiv preprint arXiv:2412.18150*, 2024.
- [53] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, Z. Xie, Y. Wu, K. Hu, J. Wang, Y. Sun, Y. Li, Y. Piao, K. Guan, A. Liu, X. Xie, Y. You, K. Dong, X. Yu, H. Zhang, L. Zhao, Y. Wang, and C. Ruan, “Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding,” *arXiv preprint arXiv:2412.10302*, 2024.

- [54] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, “Video-llava: Learning united visual representation by alignment before projection,” *arXiv preprint arXiv:2311.10122*, 2023.
- [55] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, “Videollama 3: Frontier multimodal foundation models for image and video understanding,” *arXiv preprint arXiv:2501.13106*, 2025.
- [56] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, *et al.*, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [57] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, “Qwen2.5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [58] A. Meta, “Llama 3.2: Revolutionizing edge ai and vision with open, customizable models,” *Meta AI Blog. Retrieved December*, vol. 20, p. 2024, 2024.
- [59] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Zhang, J. Li, B. Xu, Y. Dong, M. Ding, and J. Tang, “Cogagent: A visual language model for gui agents,” *arXiv preprint arXiv:2312.08914*, 2024.
- [60] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao, M. Dou, K. Chen, W. Wang, Y. Qiao, Y. Wang, and L. Wang, “Internvideo2.5: Empowering video mllms with long and rich context modeling,” *arXiv preprint arXiv:2501.12386*, 2025.
- [61] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, “Llava-onevision: Easy visual task transfer,” *arXiv preprint arXiv:2408.03326*, 2024.
- [62] G. Team, “Gemini1.5-pro.” <https://gemini.google.com/>, 2024. Accessed: 2025-03-08.
- [63] A. Team, “Claude3.5.” <https://claude.ai/>, 2024. Accessed: 2025-03-08.
- [64] xAI Team, “Grok2 vision.” <https://grok.com/>, 2024. Accessed: 2025-03-08.
- [65] O. Team, “Chatgpt-4o.” <https://chatgpt.com/>, 2024. Accessed: 2025-03-08.
- [66] P. AI, “Pixverse: Ai video creation platform.” <https://pixverse.ai/>, 2024. Accessed: 2025-03-08.
- [67] A. Cloud, “Wanxiang.” <https://tongyi.aliyun.com/wanxiang/>, 2024. Accessed: 2025-03-08.
- [68] M. Team, “Hailuo ai.” <https://hailuoai.video/>, 2024. Accessed: 2025-03-08.
- [69] O. Team, “Sora.” <https://openai.com/research/video-generation-models-as-world-simulators>, 2024. Accessed: 2025-03-08.
- [70] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao, *et al.*, “Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding,” *arXiv preprint arXiv:2405.08748*, 2024.
- [71] Runway, “Introducing gen-3 alpha: A new frontier for video generation.” <https://runwayml.com/research/introducing-gen-3-alpha>, 2024. Accessed: 2025-03-08.
- [72] K. Team, “Kling ai.” <https://klingai.io/>, 2024. Accessed: 2025-03-08.
- [73] G. Team, “Gemo.” <https://www.genmo.ai>, 2024. Accessed: 2025-03-08.
- [74] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang, “Chatglm: A family of large language models from glm-130b to glm-4 all tools,” *arXiv preprint arXiv:2406.12793*, 2024.
- [75] X. Team, “Xunfei.” <https://typemovie.art/>, 2024. Accessed: 2025-03-08.
- [76] Y. Jin, Z. Sun, N. Li, K. Xu, K. Xu, H. Jiang, N. Zhuang, Q. Huang, Y. Song, Y. Mu, and Z. Lin, “Pyramidal flow matching for efficient video generative modeling,” *arXiv preprint arXiv:2410.05954*, 2024.

- [77] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu, “Wan: Open and advanced large-scale video generative models,” *arXiv preprint arXiv:2503.20314*, 2025.
- [78] Y. Zhou, Q. Wang, Y. Cai, and H. Yang, “Allegro: Open the black box of commercial-level video generation model,” *arXiv preprint arXiv:2410.15458*, 2024.
- [79] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, “Videocrafter2: Overcoming data limitations for high-quality video diffusion models,” *arXiv preprint arXiv:2401.09047*, 2024.
- [80] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al., “Cogvideox: Text-to-video diffusion models with an expert transformer,” *arXiv preprint arXiv:2408.06072*, 2024.
- [81] J. Xu, X. Zou, K. Huang, Y. Chen, B. Liu, M. Cheng, X. Shi, and J. Huang, “Easyanimate: A high-performance long video generation method based on transformer architecture,” *arXiv preprint arXiv:2405.18991*, 2024.
- [82] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al., “Lavie: High-quality video generation with cascaded latent diffusion models,” *arXiv preprint arXiv:2309.15103*, 2023.
- [83] J. Mullan, D. Crawbuck, and A. Sastry, “Hotshot-XL.” <https://github.com/hotshotco/hotshot-xl>, 2023. Accessed: 2025-03-08.
- [84] X. Ma, Y. Wang, X. Chen, G. Jia, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, “Latte: Latent diffusion transformer for video generation,” *Transactions on Machine Learning Research*, 2025.
- [85] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan, “Videocrafter1: Open diffusion models for high-quality video generation,” *arXiv preprint arXiv:2310.19512*, 2023.
- [86] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, “Text2video-zero: Text-to-image diffusion models are zero-shot video generators,” *arXiv preprint arXiv:2303.13439*, 2023.
- [87] H. Deng, T. Pan, H. Diao, Z. Luo, Y. Cui, H. Lu, S. Shan, Y. Qi, and X. Wang, “Autoregressive video generation without vector quantization,” *arXiv preprint arXiv:2412.14169*, 2024.
- [88] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, “Modelscope text-to-video technical report,” *arXiv preprint arXiv:2308.06571*, 2023.
- [89] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7623–7633, 2023.
- [90] Y. HaCohen, N. Chirrut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi, “Ltx-video: Realtime video latent diffusion,” *arXiv preprint arXiv:2501.00103*, 2024.
- [91] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, “Latent video diffusion models for high-fidelity long video generation,” *arXiv preprint arXiv:2211.13221*, 2022.
- [92] Z. Team, “Zeroscope.” https://huggingface.co/cerspense/zeroscope_v2_XL, 2024. Accessed: 2025-03-08.
- [93] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, “World model on million-length video and language with blockwise ringattention,” *arXiv preprint arXiv:2402.08268*, 2024.
- [94] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar, “Responsible research with crowds: pay crowdworkers at least minimum wage,” *Communications of the ACM*, vol. 61, no. 3, pp. 39–41, 2018.
- [95] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, “Toward verifiable and reproducible human evaluation for text-to-image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14277–14286, 2023.

- [96] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, *et al.*, “Internvid: A large-scale video-text dataset for multimodal understanding and generation,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [97] J. Xu, T. Mei, T. Yao, and Y. Rui, “Msr-vtt: A large video description dataset for bridging video and language,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [98] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in time: A joint video and image encoder for end-to-end retrieval,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1728–1738, 2021.
- [99] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, “Tgif: A new dataset and benchmark on animated gif description,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4641–4650, 2016.
- [100] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, “T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 78723–78747, 2023.
- [101] J. Wang, H. Duan, Y. Zhao, J. Wang, G. Zhai, and X. Min, “Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms,” *arXiv preprint arXiv:2504.08358*, 2025.
- [102] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, “Scaling autoregressive models for content-rich text-to-image generation,” *arXiv preprint arXiv:2206.10789*, 2022.
- [103] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, “Musiq: Multi-scale image quality transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021.
- [104] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, “Perceptual quality assessment of smartphone photography,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3677–3686, 2020.
- [105] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv preprint arXiv:2303.05499*, 2023.
- [106] W. Sun, X. Min, W. Lu, and G. Zhai, “A deep learning based no-reference quality assessment model for ugc videos,” in *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, p. 856–865, 2022.
- [107] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.
- [108] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, *et al.*, “Minicpm-v: A gpt-4v level mllm on your phone,” *arXiv preprint arXiv:2408.01800*, 2024.
- [109] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.

Part I

Appendix

Table of Contents

A Overview	16
B Broader Impact and Ethical Discussions	16
B.1 Broader Impact of Our Research	16
B.2 Ethical Discussions of Data Collection	17
C Prompts and Task-specific Challenges	17
D Detailed Information of T2V Models	20
E More Details of Subjective Experiment	23
E.1 Annotation Dimension and Criteria	23
E.2 Significance of the Evaluation Perspectives	23
E.3 Annotation Interface	24
E.4 Annotation Management	24
F More Analysis of AIGVE-60K Database	26
F.1 MOS Distribution across 20 Challenges	26
F.2 T2V Model Performance across 20 Challenges	26
G Details of Loss Function	28
H Implementation Details	29
H.1 Detailed Information of Evaluation Criteria	29
H.2 Detailed Information of Evaluation Methods	29
H.3 Question design for LLM-based models	32
I More Results Comparisons	33

A Overview

In this Appendix, we provide additional details on the Ethical Discussions, data collection, methodology, experiments, and results discussed in the main paper. We detail the broader impact and ethical discussions in Section **B**. In data collection, we detail the 20 distinct tasks in Section **C** and the overview of the 30 T2V models in Section **D**. We then elaborate on the subjective experiments in Section **E**, including the annotation dimension, criteria, interface and management. In addition, we provide an in-depth analysis of the AIGVE-60K database, including MOS distributions and model performance comparisons across the 20 tasks in Section **F**. We outline the loss functions used in the training process for the LOVE model in Section **G**. Details on the evaluation criteria and algorithms are also included in Section **H**. Finally, we provide more performance comparisons between our model and other metrics in Section **I**.

B Broader Impact and Ethical Discussions

B.1 Broader Impact of Our Research

We discuss how our work can be applied to benefit the community through three key contributions. **Firstly**, we resolve the long-standing limitation of evaluation subjectivity and scalability by constructing **AIGVE-60K**, the **largest** AIGV evaluation dataset with 58,500 videos and 2.6M

human annotations explicitly disentangled into *perceptual quality* (visual artifacts, temporal consistency), *text-video correspondence* (prompt-video alignment), and *task-specific accuracy*. Unlike prior datasets with coarse merged scores [12] or narrow expert judgments (Table 1), our fine-grained annotations eliminate hidden bias in human preference modeling while supporting both instance-level and model-level analysis. **Secondly**, we confront the methodological fragmentation in current evaluation ecosystems where isolated metrics like FVD [26] lack prompt awareness and VBench’s threshold-dependent pipelines [9] introduce inconsistent standards. Our **LOVE** framework establishes ethical transparency through an ***all-in-one*** LMM architecture with dual vision-temporal encoders and instruction tuning, unifying perceptual/correspondence/accuracy evaluation under reproducible criteria without arbitrary thresholds. **Thirdly**, we pioneer ***bidirectional benchmarking*** that benchmarking and evaluating 30 T2V video generation models and 48 V2T interpretation models on shared standards. By open-sourcing annotations and maintaining minimal baseline tuning (LoRA adapters with 2-epoch training), we ensure community accessibility for: (1) engaging more T2V video generation models and V2T interpretation models for comparison, (2) developing more effective models based on our framework and training strategies.

B.2 Ethical Discussions of Data Collection

We detail the ethical issues that may emerge during the dataset collection process. All participants in the subjective evaluation are clearly informed of the contents in our experiments. Specifically, we addressed the ethical problems by getting a written and informed permission from each person featured in the dataset stating that they approved their subjective ratings being used for non-commercial research, thus equipping it with such legal and ethical qualities. The experiments do not contain any visually improper or NSFW content (both *textual* and *visual*), because we used extensive manual review during the AIGV generation stage. A total of 15 annotators, all postgraduate students in related fields from our laboratory, contributed to this process over 1.5 months, dedicating 3–4 hours daily to annotation tasks. We grouped the 58,500 analyzed videos into 30 sessions due to their vast volume. Each participant received \$20 per session in accordance with the current ethical standard [94, 95]. The experiment took more than a month to complete, with each participant contributing an average of 60 hours. All associated AIGVs and their corresponding prompts in the **AIGVE-60K** dataset are released under the **CC BY 4.0** license.

C Prompts and Task-specific Challenges

In this study, we systematically investigate the capabilities of text-to-video generation models through a comprehensive evaluation framework. We focus on 20 distinct tasks that vary in complexity and require diverse compositional skills, as detailed in Table 6 with their corresponding subcategories, keywords, and example prompts. Prompts of the AIGVE-60K are primarily sourced from 9 existing open-domain real-world text-video pair datasets and AIGC datasets, including InternVid [96], MSRVTT [97], WebVid [98], TGIF [99], FETV [12], AIGVQA-DB [7], GenEval [23], Sora[69] and Runaway [71] website. The prompts are classified into 20 tasks according to the subcategories and keywords. The tasks are adapted from GenEval [23] (6 tasks), T2I-CompBench [100] (3 tasks), EvalkMi-50K [101] (9 tasks), and defined by common AIGC prompt categories (e.g., PartiPrompts [102]). To avoid overfitting to template-like input, DeepSeek R1 [37] to expand and modify the prompts to ensure clarity and diversity, while human verification ensured that task integrity and alignment with evaluation goals were preserved. These tasks are carefully designed to assess different aspects of model performance, ranging from basic object rendering to complex spatial and attribute understanding, as shown in Figures 14–18. Below, we provide an overview of the main task categories and their associated challenges.

- **Object:** evaluates a model’s ability to generate a specified object class. The challenge lies in generating a single object or multiple objects. The model may be required to generate a single object from a predefined category, such as a “cat”, or to combine several objects from distinct categories, such as “dog” and “tree” into a coherent scene.
- **Color:** evaluates a model’s proficiency in associating specific color attributes with generated objects. The challenge involves not only generating an object with a single color applied but also handling more complex cases where multiple colors are used across multiple objects.

Table 6: Prompt categories with corresponding keywords and examples.

Category	Subcategory / Keywords	Prompt examples
Object	clock, person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, ...	A video of a clock
Color	red, orange, yellow, green, blue, purple, pink, brown, black, white, ...	A video of a green clock
Counting	zero, one, two, three, four, five, six, seven, eight, nine, ten, ...	A video of four clocks
Texture	glass, cement, stone, rubber, fabric, ceramics, leather, metallic, wooden, plastic, ...	A video of a wooden forks
Position	left of, right of, above, below, ...	A video of a cow below an airplane
HOI	hold a stop sign, operate an oven, peel an apple, lie on a bench, carry a book, ...	A video of people reading a book
Face	hair, mouth, emotion, eyes, necklace, cheeks, nose, skin, ...	A face video with green eyes and green hair
Emotion	happy, sadness, love, fear, surprise, anger, worry, neutrality	A dog is smiling with happy emotion
Human	human, cloth, cloth-color, hair, hair-color	A man in blue shirt smiles warmly, his curly black hair framing his face
OCR	“HELLO”, “STOP”, “SUCCESSFUL”, “Let it go”, “Thank you”, ...	A video of phrase “Good luck”
Scene	kitchen, living room, street, swimming pool, playground, waterfall, forest, ...	A video in the forest
Style	cartoon, realistic, oil painting, vintage, watercolor, line drawing, ...	A watercolor video of a backpack
Shapes	circle, cylinder, sphere, star, triangle, rectangle, irregular, oval, linear, cone, ...	A video of a circle chair
View	close-up, ground view, aerial view, first-person view, wide-angle view, ...	Aerial view of a harbor
World Knowledge	Great Wall, Great Pyramid, Ha Long Bay, Machu Picchu, Eiffel Tower, Grand Canyon, ...	Boats in Ha Long Bay
Linguistic Structure	without, no, not, ...	The garden has no flowers blooming.
Imagination	imaginative objects and scenes, impossible scenarios in the real world, ...	A panda is flying in the sky
Motion Direction	left-to-right, up-to-down, rotate, zoom in, zoom out	A balloon rises from the bottom to the top
Event Order	and then, ...	The flowers first wilt and then bloom again
Complex	Counting + Color + Shapes + Scene, Style + Color + Position, Human + Emotion, ...	A video of four blue birds playing on a circle playground

The evaluation focuses on the model’s ability to accurately bind color properties to target objects while preserving the integrity and consistency of the objects themselves.

- **Counting:** evaluates a model’s ability to generate a specific number of objects in a scene. The challenge includes numerical understanding and managing multiple instances without overlap or spatial issues, especially for larger numbers.
- **Texture:** evaluates a model’s capability to render objects with specific surface textures and material properties (*e.g.*, metallic, wooden, glass). The challenge lies in creating realistic textures that match the object’s properties and lighting conditions.
- **Position:** evaluates a model’s capability to render two objects with specified positional relationships. The challenge encompasses not only object generation but also the accurate representation of specific spatial relationships (*e.g.*, above, below, left of, right of). This requires precise control over object arrangement while maintaining their identities.
- **HOI (Human-Object Interaction):** evaluates a model’s ability to generate realistic interactions between humans and objects, ensuring the actions are physically plausible. The challenge is to create recognizable humans and objects while maintaining natural spatial and logical relationships.
- **Face:** evaluates a model’s ability to generate human faces with specific features (*e.g.*, face shape, nose structure, hairstyle). The challenge is to create realistic and diverse facial representations while maintaining feature consistency, and testing the model’s understanding of facial anatomy.
- **Emotion:** evaluates a model’s ability to convey specific emotions or moods, either through human facial expressions (*e.g.*, happiness, sadness) or through the overall atmosphere of a scene (*e.g.*, serene, love). This evaluates the model’s understanding of emotional cues and its ability to translate abstract emotions into visuals.
- **Human:** evaluates a model’s ability to generate human figures with specific occupational attire, unique accessories, and hairstyles. The challenge lies in creating realistic and coherent human representations while maintaining consistency across these attributes.
- **OCR (Optical Character Recognition):** evaluates a model’s capability to generate readable text within videos, such as words or short sentences. The challenge is to make the text visually coherent with the video and machine-readable by OCR systems, testing the model’s understanding of typography and text integration.
- **Scene:** evaluates a model’s ability to create complex scenes with multiple naturally composed elements in a specific environment (*e.g.*, beach, forest, kitchen). The challenge is to ensure all objects and backgrounds are contextually relevant and spatially consistent, evaluating the model’s holistic scene understanding.
- **Style:** evaluates a model’s proficiency in generating videos in specific artistic styles (*e.g.*, watercolor, oil painting, cartoon). The challenge is to mimic the style’s visual characteristics

while keeping objects and scenes recognizable, testing the model’s ability to apply abstract stylistic concepts consistently.

- **Shape:** evaluates a model’s ability to generate objects with specific geometric shapes (e.g., spherical, rectangular, triangular, star) while preserving their recognizability. This tests the ability to abstract representations of real-world objects and express them in other shapes.
- **View:** evaluates a model’s ability to generate videos from specific viewpoints (e.g., first-person, third-person, side view). The challenge is to maintain correct spatial orientation, scale, and proportion across perspectives, testing the model’s understanding of spatial geometry.
- **World Knowledge:** evaluates a model’s knowledge of real-world landmarks, historical sites (e.g., the Great Wall, Eiffel Tower, Great Pyramid), and the physical appearances of famous individuals (e.g., Albert Einstein). The challenge lies in creating content that accurately aligns with people’s perceptions of famous landmarks and the physical appearances of well-known individuals.
- **Linguistic Structure:** evaluates a model’s ability to interpret and render linguistic structures involving negation (e.g., “without,” “no”). The challenge is to generate videos that accurately reflect the absence of specified objects or features (e.g., a “classroom without people”) while maintaining scene integrity. This tests the model’s comprehension of negative constructs.
- **Imagination:** evaluates a model’s ability to generate imaginative scenes that combine elements from different categories or depict impossible scenarios in the real world (e.g., “a dog is driving a car”). The challenge is to balance creativity with visual plausibility, evaluating the model’s capacity for creative thinking and novel concept synthesis.
- **Motion Direction:** evaluates a model’s ability to generate or simulate motion accurately, ensuring that objects move in the intended direction. This can include simple motion along a single axis, such as left-to-right or up-to-down, as well as more complex trajectories, such as circular paths. It also involves assessing camera motion, where the model must generate realistic and coherent movement of the camera itself, whether it’s panning, zooming in/out, or rotating, to maintain consistent perspective and spatial relationships within the scene.
- **Event Order:** evaluates a model’s ability to correctly sequence events in a logical order. This typically involves understanding the chronological relationship between two events, where one occurs before the other. For example, the model may be tasked with recognizing that a certain action happens first and then follows another action. The challenge lies in ensuring that the events are correctly ordered.
- **Complex:** is designed by combining simpler task components, such as color recognition, object counting, and shape identification, into more intricate and multifaceted challenges. These tasks require models to integrate and execute multiple simple tasks simultaneously within a single video. Below are some combined forms of complex tasks along with corresponding examples:
 - (1) **Counting + Color + Shapes + Scene:** A video of [number] [color] [class] [action] in a [shape] [scene]. **Example:** *A video of two white dogs swimming in a triangle-shaped swimming pool.*
 - (2) **Counting + Color + Shapes + Texture:** A video of [number] [color] [texture] [shape] [class]. **Example:** *A video of two brown wooden rectangular books.*
 - (3) **HOI + Color + Shape + Texture:** A video of [human action] a [color] [texture] [shape] [object]. **Example:** *A video of people opening a yellow wooden triangle box.*
 - (4) **Style + Color + Position:** A [style] video of a [color1] [class1] [position] a [color2] [class2]. **Example:** *A cartoon video of a yellow dog to the left of a white cat.*
 - (5) **Style + OCR + Color:** A [style] video of [color] text “[content]”. **Example:** *An oil painting of red text “CONGRATULATIONS”.*
 - (6) **OCR + Color + Object:** A video of [color1] text “[content]” on a [color2] [class]. **Example:** *A video of green text “Happy Birthday” on a pink cake.*
 - (7) **Counting + Shapes + Object:** A video of [number1] [shape1] [class1] and [number2] [shape2] [class2]. **Example:** *A video of six spherical balls and three rectangular cups.*
 - (8) **Counting + Color + Object:** A video of [number1] [color1] [class1] and [number2] [color2] [class2]. **Example:** *A video of six red books and four blue pens.*

Sora [69] has a deep understanding of language, enabling it to accurately interpret prompts and generate compelling characters that express vibrant emotions. Sora can also create multiple shots within a single generated video that accurately preserve characters and visual style and generate complex scenes with multiple characters, specific types of motion, and accurate details of the subject and background.

Hunyuan [70] represents parameter-rich and high-performance text-to-video model currently available in the open-source domain. With 13 billion parameters, it is capable of generating videos that exhibit high physical accuracy and scene consistency, thereby actualizing conceptual visions and fostering creative expression.

Vidu1.5 [39] Vidu 1.5 introduces the world’s first Multiple-Entity Consistency capability, which seamlessly integrates people, objects, and environments to create stunning video effects. With Vidu’s Multiple-Entity Consistency feature, images with no relation to each other, be it characters, objects, or even environments, can be integrated into a single video featuring all three characteristics. Moreover, the resulting video from Vidu 1.5 is capable of ensuring visual consistency even with complex inputs that require the processing of multiple subjects or environments.

Gen3 [71] is the first of the next generation of foundation models trained by Runway on a new infrastructure built for large-scale multimodal training. Trained jointly on videos and images, Gen-3 Alpha powers Runway’s text-to-video, video-to-video and text-to-video tools, existing control modes such as motion brush, advanced camera controls, director mode as well as upcoming tools for more fine-grained control over structure, style, and motion.

Kling [72] is the new generation of AI creative productivity tools, developed by Kuaishou’s Large Model Algorithm Team, providing a wealth of AI videos, AI images, and related controllable editing capabilities.

Genmo [73] is a creative AI co-pilot developed by Genmo AI, designed for video generation and editing. It enables users to produce animations and videos from text prompts or images, and to restyle existing video clips, offering a versatile suite of tools for artistic expression and iterative creation on an accessible platform.

ChatGLM [74] is a family of large language models developed by Zhipu AI and Tsinghua University’s KEG lab, showcasing an evolution from the foundational GLM-130B to the advanced GLM-4. These models are engineered for strong bilingual (Chinese and English) capabilities in text understanding and generation, with recent iterations like GLM-4 integrating an “All Tools” framework to enhance their capacity for complex task execution and interaction with various external functionalities.

Xunfei [75], developed by iFlytek, is an AI-powered platform designed for the rapid creation of videos from textual input. It streamlines video production by enabling users to effortlessly generate diverse short video content, offering a range of visual styles and templates to facilitate quick and accessible movie-making.

Pyramid [76] is an innovative video generative model, distinguished by its novel “Pyramidal flow matching” technique. This method is designed to significantly enhance the efficiency of video synthesis by progressively matching data distributions across multiple scales, enabling high-quality video generation with potentially reduced computational resources.

Wan2.1 [77] is an advanced large-scale video generative model, presented as part of the “Wan” open research. This model is engineered for high-fidelity video synthesis, aiming to push the boundaries of generative capabilities while promoting transparency and accessibility in advanced video AI.

Allegro [78] is a video generation model, with the ambitious goal of matching commercial-level quality while aiming to “open the black box” of such advanced systems. Allegro seeks to demystify high-fidelity video synthesis by offering a transparent framework or model that endeavors to replicate sophisticated generative capabilities found in commercial offerings.

VideoCrafter2 [79] addresses the challenge of training high-quality video models without high-quality video data. By disentangling motion from appearance at the data level, it uses low-quality videos to maintain motion consistency and high-quality images to ensure visual quality.

CogVideo X1.5 [80], part of the CogVideoX series, is a text-to-video diffusion model distinguished by its integration of an “Expert Transformer.” This specialized architecture is designed to enable more

efficient scaling and task specialization, aiming for enhanced performance in generating complex and high-fidelity videos from textual descriptions.

Animate [81] is a high-performance method specifically designed for long video generation utilizing a transformer-based architecture. This approach focuses on creating extended and temporally consistent video narratives by harnessing the long-range dependency modeling strengths of transformers to maintain coherence over time.

Lavie [82] is a video generation model engineered to produce high-quality outputs through the use of “cascaded latent diffusion models.” This multi-stage architecture progressively generates and refines video data in the latent space, enabling the creation of detailed and coherent visual sequences by breaking down the complexity of direct high-resolution synthesis.

Hotshot-XL [83] is a text-to-gif model trained to work alongside Stable Diffusion XL. We adopt its official code with default parameters and change the output format from GIF to MP4.

Latte [84] is a latent diffusion Transformer model for video generation. It adopts a video Transformer as the backbone and leverages a pretrained variational autoencoder (VAE) to encode input videos into latent space representations. From these latent features, spatial-temporal tokens are extracted and processed by a series of Transformer blocks. Latte introduces four effective Transformer-based architectural variants by decomposing the input video’s spatial and temporal dimensions.

VideoCrafter1 [85] introduces two diffusion models for high-quality video generation: a T2V model for text-to-video and an I2V model for image-to-video. The T2V model enhances SD 2.1 with temporal attention layers for consistent video generation, trained on large datasets.

NOVA [87] is a model that enables autoregressive image/video generation with high efficiency. It reformulates the video generation problem as non-quantized autoregressive modeling of temporal frame-by-frame prediction and spatial set-by-set prediction. It generalizes well and enables diverse zero-shot generation abilities in one unified model.

ModelScope [88] is a decomposed diffusion probabilistic model for video generation. Unlike traditional methods that add independent noise to each frame, it separates noise into shared base noise and residual noise, improving spatial-temporal coherence. This approach leverages pretrained video-generation models for efficient frame content prediction while maintaining motion dynamics.

Text2Video-Zero [86] is a zero-shot text-to-video synthesis model without any further fine-tuning or optimization, which introduces motion dynamics between the latent codes and cross-frame attention mechanism to keep the global scene time consistent.

Tune-A-Video [89] is a one-shot text-to-video generation model that extends text-to-video models to the spatio-temporal domain. It uses sparse spatio-temporal attention to maintain consistent objects across frames, overcoming computational limitations. It can synthesize novel videos from a single example compatible with personalized and conditional pretrained T2V models.

LTX [90] is a latent diffusion model engineered for real-time video generation. This system’s core innovation lies in its ability to synthesize video sequences at interactive speeds, significantly reducing the typical generation latency associated with diffusion-based video models.

LVDM [91] is an efficient video diffusion model operating in a compressed latent space, designed to address the computational challenges of video synthesis. It uses a hierarchical framework to extend video generation beyond training lengths, effectively mitigating performance degradation via conditional latent perturbation and unconditional guidance techniques.

ZeroScope [92] is specifically designed for upscaling content made with zeroscope_v2_576w using vid2vid in the text2video extension by kabachuha. Leveraging this model as an upscaler allows for superior overall compositions at higher resolutions, permitting faster exploration in 576x320 (or 448x256) before transitioning to a high-resolution render.

LWM [93] is a multimodal autoregressive model trained on extensive video and language data. Using RingAttention, it efficiently handles long-sequence training, expanding context size up to 1M tokens, enabling strong language, video, and video understanding and generation.

E More Details of Subjective Experiment

E.1 Annotation Dimension and Criteria

To comprehensively assess the performance of AI-generated videos (AIGVs), we propose a dual-dimensional evaluation framework that examines both perceptual quality and text-to-video (T2V) correspondence. This approach enables a thorough analysis of different aspects of video generation, providing a holistic understanding of a model’s capabilities and limitations.

- **Perceptual quality** evaluates the visual characteristics and aesthetic appeal of generated videos. This dimension focuses on multiple aspects of video quality, including **visual clarity** (the sharpness and resolution of video details), **naturalness** (the degree to which the video appears realistic and free from artifacts), **aesthetic appeal** (the composition, color harmony, and overall visual attractiveness), **temporal smoothness** (the temporal consistency of the video frames and the smoothness of visual elements transition over time), and **authenticity** (whether the generated video is realistic). High-scoring videos are characterized by exceptional clarity, vivid and well-balanced colors, and meticulous attention to detail, offering an immersive and visually striking experience. In contrast, low scores reflect videos with blurriness, unnatural color tones, faded visuals, and a lack of clarity or detail. This dimension captures the foundational visual attributes that make a video aesthetically pleasing or distracting. For detailed criteria, refer to Figure 12.
- **Text-video correspondence** assesses the semantic alignment between the generated video and the input text prompt, including **content accuracy** (the presence and correct representation of described objects and elements), **contextual relevance** (the appropriate depiction of scenes and relationships between objects), **attribute fidelity** (the accurate representation of specific characteristics mentioned in the prompt), and **semantic consistency** (the logical coherence between visual elements and textual descriptions). Videos with high scores perfectly match the descriptions in the prompt, accurately reflecting all elements with high fidelity. These videos effectively translate textual information into visual content without mismatches. In contrast, videos with lower scores exhibit inconsistencies, missing elements, or mismatched content. For detailed criteria, refer to Figure 13.

E.2 Significance of the Evaluation Perspectives

The dual-dimensional evaluation framework, which combines perception quality and T2V correspondence, is essential for addressing the inherent trade-offs and complementary aspects of AIGVs. While perception quality emphasizes the visual characteristics that contribute to a video’s appeal and realism, T2V correspondence ensures that the generated content remains semantically faithful to the original textual description. A high perception quality score alone does not guarantee semantic accuracy. For example, a video may exhibit exceptional visual quality, characterized by high resolution, vibrant colors, and meticulous detail, yet fail to accurately represent the specific objects, relationships, or attributes described in the text prompt. Conversely, a video may perfectly align with the textual description in terms of content and context but suffer from poor visual quality, such as low resolution, unnatural textures, or inconsistent lighting, which detracts from its overall appeal and usability. The integration of both dimensions ensures that generated videos achieve a balance between visual excellence and semantic fidelity. This holistic approach not only enhances the evaluation of generative models but also aligns with real-world applications where both video quality and content accuracy are critical. By considering both dimensions, the framework provides a more nuanced understanding of a model’s strengths and weaknesses, facilitating targeted improvements in video generation systems.

While high-level scores (perception quality and T2V correspondence) provide a general overview of model performance, they fall short when it comes to accurately evaluating tasks that require a deeper level of understanding and precision. For simple tasks, such as identifying the color “white” in an image while the prompt states “blue”, high-level scores might vary slightly depending on the annotators’ interpretation of the content, but for such straightforward tasks, the use of task-specific yes/no questions provides a unified and clear evaluation criterion. A simple yes/no answer is sufficient to confirm whether the model has correctly interpreted the task. In contrast, for more complex tasks such as “3 red pens in the box”, yes/no answer is not adequate. These tasks require a more detailed and precise assessment, as the model needs to recognize and quantify multiple tasks and their accuracy. A simple yes/no response would fail to capture the depth of understanding needed

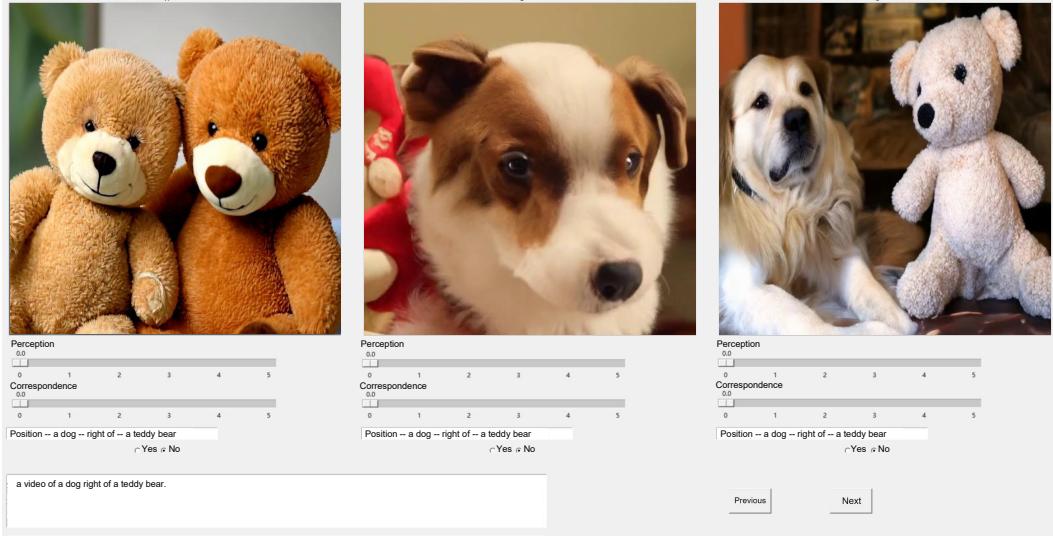


Figure 9: An example of the simple task annotation interface for human evaluation. The subjects are instructed to rate two dimensions of AI-generated videos, *i.e.*, perception and text-video correspondence, and provide a binary (yes/no) response for a task-specific challenge. Each trial presents three videos generated from 30 models for the same prompt, with absolute scoring applied independently to each video.

to evaluate whether the model has correctly identified the number of objects and their characteristics. For such tasks, a more detailed and nuanced assessment is necessary, which is why high-level scores are employed. Together, high-level scores and task-specific QA pairs allow for a comprehensive evaluation framework that ensures both general alignment and task-specific accuracy, enabling the model’s performance to be measured across a wide range of complexity.

E.3 Annotation Interface

To ensure a comprehensive and efficient video quality evaluation, we design two custom annotation interfaces tailored for different assessment tasks: simple task annotation and complex task annotation. The simple task annotation interface, shown in Figure 9, is a manual evaluation platform that was developed using the Python tkinter package to facilitate MOS assessments. The experiment involves evaluating videos based on two independent dimensions and answering a binary question related to a specific task-specific challenge. There are 20 task-specific challenges, including categories such as human, shape, scene, color, etc. Each trial presents three videos that correspond to the same prompt. These videos are randomly selected from 30 different models. Importantly, participants are instructed to assign absolute scores to each video on the two predefined dimensions, rather than making relative comparisons between the videos. For each video, participants provide: (1) Two separate scores representing the two evaluation dimensions. (2) A binary response (yes/no) to indicate whether the video meets the specified challenge criterion. Meanwhile, the complex task annotation interface is illustrated in Figure 10. The complex tasks are composed of multiple subtasks, such as Color, Shape, and Scene. Each subtask is evaluated independently with a yes/no response. The complex task is considered correct only if all its sub-tasks are correct. If any sub-task is incorrect, the entire complex task is marked as incorrect. To ensure uniformity and minimize resolution-related biases in video quality evaluation, all videos displayed in this interface are cropped to a spatial resolution of 1030×1030 pixels. Navigation options, such as “Previous” and “Next” streamline the workflow, enabling efficient annotation.

E.4 Annotation Management

The annotation process is structured into two primary components: Mean Opinion Score (MOS) annotation and task-specific question-answering (QA) annotation. Each component is designed to evaluate videos across 20 task-specific challenges, including color, position, shapes, view, and *etc.* The MOS annotation task involves 15 participants to rate each video on a 0-5 Likert scale,

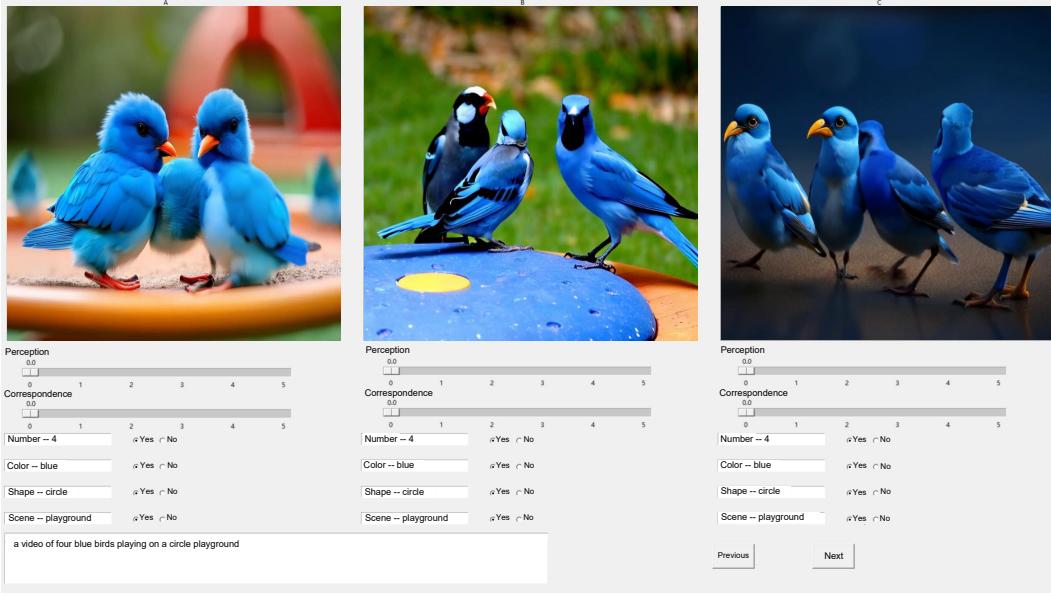


Figure 10: An example of the complex task annotation interface, which extends the simple task evaluation by incorporating multiple sub-tasks (*e.g.*, Number, Color, Shape, and Scene). The subjects are instructed to rate two dimensions of AI-generated videos, *i.e.*, perception and text-video correspondence, based on the given video and its prompt. Each sub-task is judged independently with a yes/no response. The complex task is considered correct only if all sub-tasks are correctly identified; if any sub-task is incorrect, the entire complex task is marked as incorrect.

assessing both perception quality and T2V correspondence. The question-answering annotation task is similarly conducted with 15 participants, ensuring consistency in the evaluation process. In this task, participants are presented with a series of yes/no questions across the 20 task-specific challenges. To determine the final answer for each question, a majority voting mechanism is employed. This approach ensures that the final decision reflects the collective judgment of the participants, minimizing the impact of individual biases or errors.

Prior to engaging in the annotation tasks, all participants undergo a rigorous training process. As illustrated in Figures 12-13, they are provided with detailed instructions and multiple standardized examples. To ensure a high level of understanding and consistency, a pre-test is conducted to evaluate participants’ comprehension of the criteria and their alignment with the standard examples. Participants who do not meet the required accuracy threshold are excluded from further participation, ensuring that only well-prepared individuals contribute to the final dataset. During the experiment, all evaluations are conducted in a controlled laboratory environment under normal indoor lighting conditions. Participants are seated at a comfortable viewing distance of approximately 60 cm from the screen to minimize visual strain and ensure consistent evaluation conditions. While individual preferences may naturally vary, the use of detailed explanations and standardized annotation criteria ensures a high degree of agreement among participants. This consensus is particularly evident in question-answering annotations, where majority voting effectively captures group preferences.

To ensure the reliability of subjective ratings, we follow the ITU-R BT.500 [24] and apply a two-step outlier rejection procedure based on statistical properties of the collected scores. $S_{i,j}$ denote the raw score given by subject j to item i , and let μ_i , σ_i be the mean and standard deviation across raters for item i .

1. Subject-level Outlier Rejection. We compute the kurtosis β_i of the score distribution for each item i , and use it to determine a dynamic threshold coefficient k_i :

$$k_i = \begin{cases} 2, & \text{if } 2 \leq \beta_i \leq 4 \\ \sqrt{20}, & \text{otherwise} \end{cases}$$

For each subject j , we count how many times their score deviates from the mean by more than $k_i\sigma_i$:

$$P_j = |\{i \mid S_{i,j} \geq \mu_i + k_i\sigma_i\}|, \quad Q_j = |\{i \mid S_{i,j} \leq \mu_i - k_i\sigma_i\}|$$

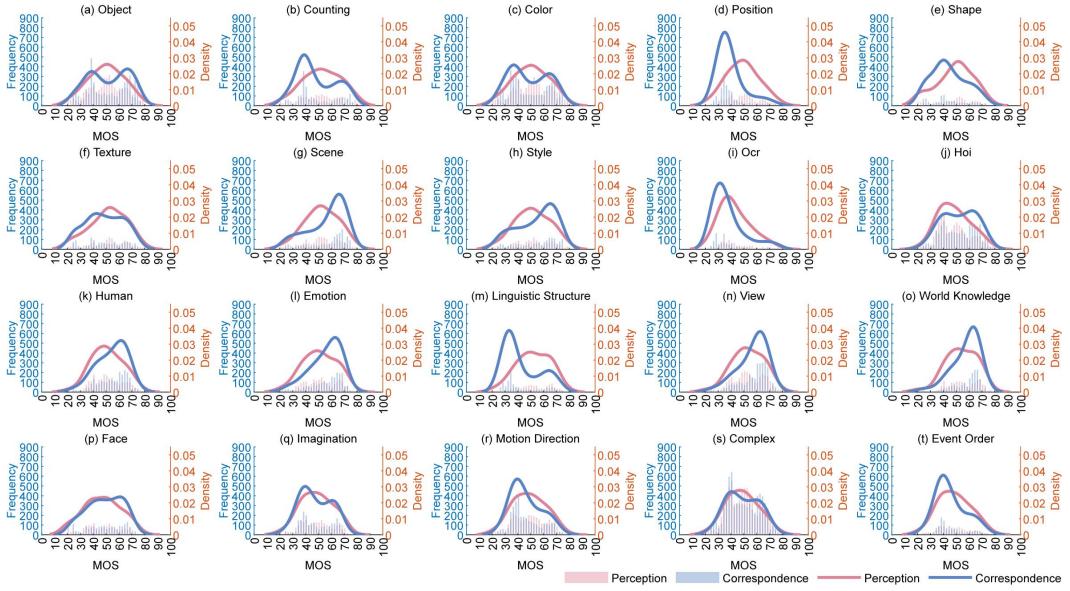


Figure 11: Mean Opinion Score (MOS) distribution histograms and kernel density curves of AIGVE-60K dataset. It includes two dimensions: perception MOS and correspondence MOS. Each dimension contains a total of 58,500 MOS values.

Subject j is considered an outlier and rejected if:

$$\frac{P_j + Q_j}{N} > 0.05 \quad \text{and} \quad \left| \frac{P_j - Q_j}{P_j + Q_j} \right| < 0.3$$

where N is the total number of items.

2. Score-level Outlier Rejection. For each accepted subject, their score $S_{i,j}$ is retained only if it falls within the interval:

$$\mu_i - k_i \sigma_i \leq S_{i,j} \leq \mu_i + k_i \sigma_i$$

Otherwise, the score is excluded from the MOS calculation.

With a subject rejection rate of 3% and a per-score rejection rate of approximately 2%, this rigorous and ethically sound annotation management strategy establishes AIGVE-60K as a robust and reliable resource for advancing research in video quality assessment.

F More Analysis of AIGVE-60K Database

F.1 MOS Distribution across 20 Challenges

As mentioned in the main text, we process and compute the valid subjective evaluation results, obtaining a total of 117,000 Mean Opinion Scores (MOSS) across two dimensions, along with QA accuracy. To better illustrate the generative capabilities of current T2V models in different prompt challenges, we categorize the computed MOSSs data into 20 task categories and used the categorized data to plot histograms and kernel density curves (KDC) graphs, as shown in Figure 11. We can observe that the 30 T2V models we tested exhibit relatively poor text-video alignment in prompt challenges related to position, OCR, linguistic structures, and complexity, with MOSSs primarily clustering around 30. In contrast, their performance in other prompt challenges is relatively better. The overall perception MOSSs does not show significant differences across different prompt challenges, with scores generally concentrated at a higher level. However, models perform slightly worse in OCR, HOI, and Face-related prompt challenges, where lower MOSSs appear more frequently compared to other prompt challenges.

F.2 T2V Model Performance across 20 Challenges

Tables 8-10 provide detailed performance comparisons of the 30 T2V models across 20 task-specific challenges on three types of human annotations: perception MOS, T2V correspondence MOS, and

task-specific accuracy. For perception quality, as shown in Table 8, Jimeng [38] stands out with the highest MOS and performs particularly well in the category “Complex”. This is likely because complex combinations tend to be visually appealing and often involve the harmonious integration of various components, which makes them more engaging and immersive. For T2V correspondence, as demonstrated in Table 9, Wanxiang [67] leads the way, demonstrating strong alignment between the generated videos and the textual descriptions, but has a relatively lower performance in perception quality. In contrast, models such as Kling [72] excel in perception quality, delivering high MOS scores, but cannot perform as well in terms of T2V correspondence. The contrasting trends in performance between perception quality and T2V correspondence emphasize the importance of evaluating both dimensions independently. While perception quality focuses on the visual aspects of the generated videos, T2V correspondence measures how well the video aligns with the content described in the text prompt. This dual evaluation ensures a more comprehensive understanding of a model’s abilities, where one dimension evaluates aesthetic quality, and the other checks the accuracy of the video-text alignment. Most models perform well in the “Object” category but fall short in the “OCR” category. This suggests that while the T2V models excel at generating visually coherent objects, they struggle with accurately interpreting and rendering textual elements. In terms of task-specific accuracy, the differences between models become particularly evident. Some models achieve near-perfect accuracy, with scores reaching as high as 100%, demonstrating their exceptional ability to handle specific tasks with precision. On the other hand, certain models perform poorly, with accuracy scores as low as 0%, indicating significant limitations in their ability to correctly complete the designated tasks. This stark contrast highlights the varying capabilities of the models and underscores the importance of task-specific evaluation in assessing the true performance of text-to-video models across different challenges.

G Details of Loss Function

The training process for LOVE is divided into two progressive stages, each utilizing a specific loss function to target distinct objectives: language loss for training, aligning visual and language features to give visual question answers across the 20 task-specific challenges, L1 loss for quality regression to generate accurate perception and correspondence scores.

(1) Training with language loss. In the first stage, we train the projector to align visual and language features using the standard language loss. This involves ensuring that the visual tokens extracted from the vision encoder and temporal encoder correspond effectively to the language representations from the LLM. The language loss, calculated using a cross-entropy function, measures the model’s ability to predict the correct token given the prior context:

$$\mathcal{L}_{\text{language}} = -\frac{1}{N} \sum_{i=1}^N \log P(y_{\text{label}}|y_{\text{pred}}) \quad (5)$$

where $P(y_{\text{label}}|y_{\text{pred}})$ represents the probability assigned to the correct token y_{label} by the model, y_{pred} is the predicted token, and N is the total number of tokens. By minimizing this loss, the model learns to generate coherent textual descriptions of video content, laying the foundation for subsequent stages.

(2) Quality Regression with L1 loss. Once the model can produce coherent descriptions of video content, the focus shifts to fine-tuning the quality regression module to output stable and precise numerical quality scores. The quality regression module takes the aligned visual tokens as input and predicts a quality score that reflects the overall video quality. Using the AIGVE-60K, which contains human-annotated MOS for each video, the model is trained to align its predictions with human ratings. The training objective minimizes the difference between the predicted quality score Q_{predict} and the ground-truth MOS Q_{label} using the L1 loss function:

$$\mathcal{L}_{\text{MOS}} = \frac{1}{N} \sum_{i=1}^N |Q_{\text{predict}}(i) - Q_{\text{label}}(i)| \quad (6)$$

where $Q_{\text{predict}}(i)$ is the score predicted by the regressor i and $Q_{\text{label}}(i)$ is the corresponding ground-truth MOS derived from subjective experiments, and N is the number of videos in the batch. This loss function ensures that the predicted scores remain consistent with human evaluations, enabling the model to accurately assess the quality of AI-generated videos in numerical form.

H Implementation Details

H.1 Detailed Information of Evaluation Criteria

We adopt the widely used metrics in VQA literature [20–22]: Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and Kendall’s Rank Correlation Coefficient (KRCC) as our evaluation criteria. SRCC quantifies the extent to which the ranks of two variables are related, which ranges from -1 to 1. Given N action videos, SRCC is computed as:

$$SRCC = 1 - \frac{6 \sum_{n=1}^N (v_n - p_n)^2}{N(N^2 - 1)}, \quad (7)$$

where v_n and p_n denote the rank of the ground truth y_n and the rank of predicted score \hat{y}_n respectively. The higher the SRCC, the higher the monotonic correlation between ground truth and predicted score. Similarly, PLCC measures the linear correlation between predicted scores and ground truth scores, which can be formulated as:

$$PLCC = \frac{\sum_{n=1}^N (y_n - \bar{y})(\hat{y}_n - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^N (y_n - \bar{y})^2} \sqrt{\sum_{n=1}^N (\hat{y}_n - \bar{\hat{y}})^2}}, \quad (8)$$

where \bar{y} and $\bar{\hat{y}}$ are the mean of ground truth and predicted score respectively.

We also adopt the Kendall Rank Correlation Coefficient (KRCC) as an evaluation metric, which measures the ordinal association between two variables. For a pair of ranks (v_i, p_i) and (v_j, p_j) , the pair is concordant if:

$$(v_i - v_j)(p_i - p_j) > 0, \quad (9)$$

and discordant if < 0 . Given N AIGVs, KRCC is computed as:

$$KRCC = \frac{C - D}{\frac{1}{2}N(N - 1)}, \quad (10)$$

where C and D denote the number of concordant and discordant pairs, respectively.

H.2 Detailed Information of Evaluation Methods

BMPRI [44], **BPRI** [45], **BRISQUE** [36], **HOSA** [46], **NIQE** [28], **QAC** [35] are conventional handcrafted IQA methods. These methods primarily rely on the extraction of natural image features, which are designed to capture perceptual characteristics such as texture, structure, and color information. For our video input, we employ a strategy of uniformly sampling 8 frames from each video. This ensures that we capture key frames of the videos, while maintaining computational efficiency. The scores obtained for each frame are then averaged to produce a final score for the video.

V-Aesthetic Quality [9], **V-Imaging Quality** [9], **V-Overall Consistency** [9], **V-Subject Consistency** [9], **V-Temporal Flickering** [9] are metrics derived from VBench [9], which incorporate a wide range of smaller, specialized sub-metrics designed to evaluate various aspects of video quality. **V-Aesthetic** Quality evaluates the visual appeal of the generated video, considering factors such as layout, color richness, and the artistic quality of the subjects. Using the LAION [49] aesthetic predictor, each video frame is rated on a scale from 0 to 10, normalized to a 0-1 range. The final aesthetic score for the video is obtained by averaging the scores of all frames, with higher scores indicating better aesthetic quality. **V-Imaging Quality** measures low-level distortions like noise and blur, using the MUSIQ [103] predictor trained on the SPAQ [104] dataset to compute frame-wise quality scores. **V-Overall Consistency** measures the overall consistency between the video and the corresponding text prompt, focusing on both semantic and style alignment. This is assessed using ViCLIP [96], a model that evaluates how well the text and generated video align in terms of meaning and stylistic elements. **V-Subject Consistency** is measured using DINO [105] features, with the subject consistency score calculated based on cosine similarity between frames. **V-Temporal Flickering** evaluates temporal consistency by focusing on flickering caused by lighting or shaky camera motions, using static video scenes to isolate this issue and calculating flickering on a frame-by-frame basis.

VSFA [19] is an objective no-reference video quality assessment method by integrating two eminent effects of the human visual system, namely, content-dependency and temporal-memory effects into a deep neural network.

Table 11: An overview and URLs of the adopted 30 V2T interpretation models.

Methods	URL
♣VSFA [19]	https://github.com/lidq92/VSFA
♣BVQA [33]	https://github.com/vztu/BVQA_Benchmark
♣SimpleVQA [20]	https://github.com/Raykshj/SimpleVQA
♣FAST-VQA [21]	https://github.com/VQAssessment/FAST-VQA-and-FasterVQA
♣DOVER [22]	https://github.com/VQAssessment/DOVER
♡CLIPScore [47]	https://github.com/jmhessel/clipscore
♡BLIPScore [48]	https://github.com/salesforce/BLIP
♡AestheticScore [49]	https://github.com/sorekdj60/AestheticScore
♡ImageReward [50]	https://github.com/THUDM/ImageReward
♡PickScore [34]	https://github.com/yuvalkirstain/PickScore
♡HPSv2 [51]	https://github.com/tgxs002/HPSv2
♡VQAScore [8]	https://github.com/linzhiqu/t2v_metrics
♡FGA-BLIP2 [52]	https://github.com/DYEvaLab/EvalMuse
★DeepseekVL2 [53]	https://github.com/deepseek-ai/DeepSeek-V2
★VideoLlava [54]	https://github.com/PKU-YuanGroup/Video-LLAVA
★VideoLlama3 [55]	https://github.com/DAMO-NLP-SG/VideoLLaMA3
★mPLUG-OWL3 [5]	https://github.com/X-PLUG/mPLUG-Owl
★Qwen-VL [57]	https://github.com/QwenLM/Qwen2.5-VL
★Llama3.2-Vision [58]	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision
★CogAgent [59]	https://github.com/THUDM/CogAgent
★LLaVA-NEXT [6]	https://github.com/LLaVA-VL/LLaVA-NeXT
★InternVideo2.5 [60]	https://github.com/OpenGVLab/InternVideo
★InternVL [41]	https://github.com/OpenGVLab/InternVL
△Gemini1.5-pro [62]	https://gemini.google.com
△Claude3.5 [63]	https://claude.ai
△Grok2 Vision [64]	https://grok.com
△ChatGPT-4o [65]	https://chatgpt.com

BVQA [33] leverages the transferred knowledge from IQA databases with authentic distortions and large-scale action recognition with rich motion patterns for better video representation.

SimpleVQA [106] adopts an end-to-end spatial feature extraction network to directly learn the quality-aware spatial feature representation from raw pixels of the video frames and extract the motion features to measure the temporal-related distortions. A pre-trained SlowFast model is used to extract motion features.

FAST-VQA [21] proposes a grid mini-patch sampling strategy, which allows consideration of local quality by sampling patches at their raw resolution and covers global quality with contextual relations via mini-patches sampled in uniform grids. It overcomes the high computational costs when evaluating high-resolution videos.

DOVER [22] is a disentangled objective video quality evaluator that learns the quality of videos based on technical and aesthetic perspectives.

CLIPScore [47] is an image captioning metric and passes both the image and the candidate caption through their respective feature extractors, then computing the cosine similarity between the text and video embeddings.

BLIPScore [48] provides more advanced multi-modal feature extraction capabilities. Using the same methodology as CLIPScore [47], it computes the cosine similarity between the text and visual embeddings, but benefits from enhanced pre-training strategy, which is designed to better capture fine-grained relationships between text and visual content.

ImageReward [50] builds upon the BLIP model [48] by introducing an additional MLP layer on top of BLIP’s output. Instead of directly computing a similarity score, the MLP generates a scalar value representing the preference for one video over another in comparative settings.

AestheticScore [49] is given by an aesthetic predictor introduced by LAION [49].

PickScore [34] is trained by fine-tuning CLIP-H on human preference data, aiming to maximize the probability of a preferred video being selected. PickScore exhibits strong correlation with human rankings, outperforming traditional metrics like FID and aesthetics predictors.

HPSv2 [51] is designed for better aligning their outputs with human preferences. HPS is based on a fine-tuned CLIP model that accurately predicts human preferences over generated contents.

VQAScore [8] is designed to assess the alignment between generated videos and text prompts, particularly for compositional text-to-visual generation tasks. It can be used in a black-box manner, requiring no fine-tuning or additional prompt decomposition.

FGA-BLIP2 [52] utilizes vision-language models to jointly fine-tune image-text alignment scores and element-level annotations. This approach enables the model to generate overall scores while determining whether the generated images match the elements specified in the prompt.

DeepseekVL2 [53] is an advanced series of mix-of-experts (MoE) vision language models. It introduces a dynamic tiling vision encoding strategy, allowing efficient processing of high-resolution videos with varying aspect ratios, enhancing tasks like visual grounding and document analysis. It also leverages the Multi-head Latent Attention (MLA) mechanism for the language component, which reduces computational costs and improves inference efficiency.

VideoLlava [107] Video-LLaVA binds visual signals to the language feature space, unifying visual representations, and proposes a solution to align before projection. It enables LLM to perform visual reasoning capabilities on both images and videos simultaneously.

VideoLlama3 [108] is a vision-language model trained through a four-stage paradigm. Key innovations include using Rotary Position Embedding (RoPE) for dynamic image resolution and compressing video tokens for more efficient representation, enabling the model to achieve good performance in both image and video understanding benchmarks.

LLaVA-NeXT [6] improves on LLaVA-1.5 [107] by increasing input video resolution and enhances visual detail, reasoning, and OCR capabilities. It also improves world knowledge and logical reasoning while maintaining LLaVA’s minimalist design and data efficiency, using under 1M visual instruction tuning samples.

mPLUG-Owl3 [5] is a versatile multi-modal large language model designed to handle long video sequences, interleaved video-text, and lengthy video inputs. It introduces Hyper Attention blocks that efficiently integrate vision and language into a shared semantic space, allowing for the processing of extended multi-video scenarios.

Qwen2-VL [109] is an advanced large vision-language model designed to process videos, images, and text with dynamic resolution handling and multimodal rotary position embedding (M-RoPE). The model features strong capabilities in OCR, video comprehension, multilingual support, and robust agent functionalities for device operations.

Qwen2.5-VL [57] is the latest flagship model in the Qwen vision-language series, featuring significant improvements in visual recognition, object localization, document parsing, and long-video comprehension. Building on the Qwen2-VL architecture, it introduces key enhancements such as dynamic resolution processing for videos and images, absolute time encoding for temporal dynamics, and window attention to optimize inference efficiency.

Llama3.2-Vision [58] excels in video reasoning tasks, such as document-level understanding, chart and graph captioning, and visual grounding. These models can reason with videos, such as answering questions based on graphs or maps, and generate captions that describe visual scenes.

Llava-one-vision [61] is an open-source large multimodal model (LMM) designed to enhance vision-and-language tasks in single-image, multi-image, and video scenarios. It utilizes a cost-efficient architecture connecting vision encoders with LLMs, demonstrating strong video understanding through task transfer from images.

CogAgent [59] is designed to facilitate understanding and navigation of graphical user interfaces (GUIs). It utilizes both low and high-resolution video encoders to recognize small text and page elements. CogAgent excels in GUI tasks like navigation and decision-making. CogAgent’s innovative design includes a cross-attention branch to balance high-resolution inputs and computational efficiency.

InternVL2.5 [41] demonstrates strong performance in various benchmarks, including multi-discipline reasoning, document and video understanding, and multimodal hallucination detection. The model features enhanced vision encoders, larger dataset sizes, and improved test-time scaling.

InternVL3 [41] is an advanced multimodal large language model series that surpasses its predecessor, InternVL 2.5, in multimodal perception and reasoning. It extends its capabilities to tool usage, GUI agents, industrial image analysis, 3D vision perception, and more. Built upon Native Multimodal Pre-Training, InternVL3 integrates language and multimodal training in a single stage, enhancing its performance without requiring additional bridging modules.

InternVideo2.5 [60] is an advanced multimodal large language model designed to enhance video understanding by focusing on long and rich context modeling. This approach improves the model’s ability to perceive fine-grained details and capture long-term temporal structures in videos. By incorporating dense vision task annotations and utilizing direct preference optimization (DPO), it creates compact spatiotemporal representations through adaptive hierarchical token compression.

H.3 Question design for LLM-based models

For LMM-based methods, we not only need to input the video to be evaluated, but also the corresponding prompt and instructions to guide the model to output the result we want. Three different questions need to be input for each video to be evaluated. When designing questions from the two dimensions of Perception and T2V correspondence, all videos have a unified template, but to obtain the question-answer pair for each video, different questions need to be designed according to the challenge corresponding to the prompt used to generate the video. We have a total of 20 tasks, so there are 20 question templates for the task-specific challenges. The specific question templates are listed as follows:

- **Perception Quality:** Suppose you are now a volunteer for subjective quality evaluation of videos, and you are now required to rate the perception quality of the given videos on a scale of 0-100. Results are accurate to the nearest digit. Answer only one score.
- **T2V Correspondence:** Please rate the consistency between the video and the text description “<prompt >”. The rating scale is from 0 to 100, with higher scores for descriptions that include important content from the video and lower scores for descriptions that lack important content. Results are accurate to the nearest digit. Answer only a score.
- **Task-specific Questions:**
 - (1) **Object:** Does the video contain <class_name >? Answer yes or no.
 - (2) **Color:** Does the video contain <class_name >in the color of <class_color >? Answer yes or no.
 - (3) **Counting:** Does the video contain <class_count ><class_name >? Answer yes or no.
 - (4) **Texture:** Does the video contain a <class_texture ><class_name >? Answer yes or no.
 - (5) **Position:** Does the video contain both <class1_name >and <class2_name >, and are they positioned as described in “<prompt >”? Answer yes or no.
 - (6) **HOI:** Does the video contain both a person and <object_name >, and is the person’s action <verb_ ing >? Answer yes or no.
 - (7) **Face:** Does the face in the video have <first_feature ><second_feature >and <third_feature >? Answer yes or no.
 - (8) **Emotion:** If there is a person in the video, is their emotion <emotion_class >? If there is no person, does the overall mood of the video convey <emotion_class >? Answer yes or no.
 - (9) **Human:** Do the appearance, hairstyle, accessories, and profession of the person in the video match the description in “<prompt >”? Answer yes or no.
 - (10) **OCR:** Does the video contain the text “<OCR >” with all letters correct? Answer yes or no.
 - (11) **Scene:** Does the video depict a <scene_name >scene? Answer yes or no.
 - (12) **Style:** Is the style of the video <style_name >? Answer yes or no.
 - (13) **Shapes:** Does the video contain a <class_shape ><class_name >? Answer yes or no.
 - (14) **View:** Is the perspective shown in the video <view_class >? Answer yes or no.
 - (15) **World Knowledge:** Does the video contain a famous landmark or celebrity <knowledge_class >? Answer yes or no.
 - (16) **Linguistic Structure:** Does the scene depicted in the video exclude <class_name >? Answer yes or no.
 - (17) **Imagination:** Does the video content show imaginative elements, and does it match the description in “<prompt >”? Answer yes or no.
 - (18) **Motion Direction:** Does the video show the trajectory and path of <class_name >as described in as described in “<prompt >”? Answer yes or no.

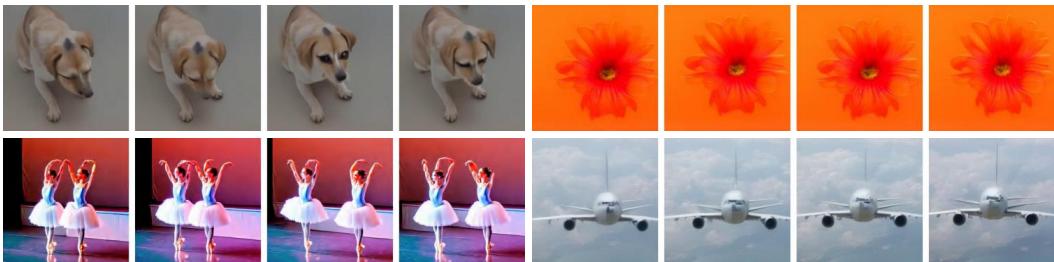
Perception 4-5 (Excellent): The video achieves near-flawless quality with high frame-to-frame coherence, crisp clarity, photorealistic details, and fluid motion, devoid of artifacts and visible distortions.



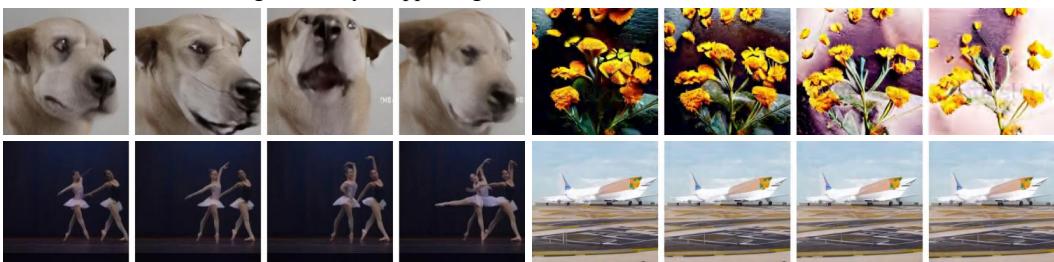
Perception 3-4 (Good): The video is visually appealing with minor flaws, offering clear details, smooth frame transitions and natural colors with only occasional subtle imperfections in motion or clarity.



Perception 2-3 (Fair): The video is acceptable but with vague or contains noticeable imperfections, such as inconsistent frame transitions, artifacts or features that clearly indicate it is AI-generated.



Perception 1-2 (Poor): The video has significant flaws, such as heavy artifacts, poor detail, temporal flickering, or unnatural colors, making it visually unappealing.



Perception 0-1 (Bad): The video is severely distorted, unrealistic, unrecognizable, or motion-incoherent.



Figure 12: Instructions and examples for manual evaluation of the **perceptual quality**.

Correspondence 4-5 (Excellent): The video perfectly matches all details, relations, and nuances in the text.



Correspondence 3-4 (Good): The video closely aligns with the text, accurately representing most described elements with minor errors or omissions.



Perception 2-3 (Fair): The video partially matches the text but has significant inconsistencies, such as missing key objects or incorrect attributes.



Perception 1-2 (Poor): The video shows minimal alignment with the text, containing incorrect representations of the described elements.



Perception 0-1 (Bad): The video completely fails to match the text description.



Yellow phrase "Best Wishes" on a *blue* box.

A green bus and a *purple* microwave.

Figure 13: Instructions and examples for manual evaluation of **T2V correspondence**. Prompt (left): yellow phrase “Best Wishes” on a blue box. Prompt (right): a green bus and a purple microwave.

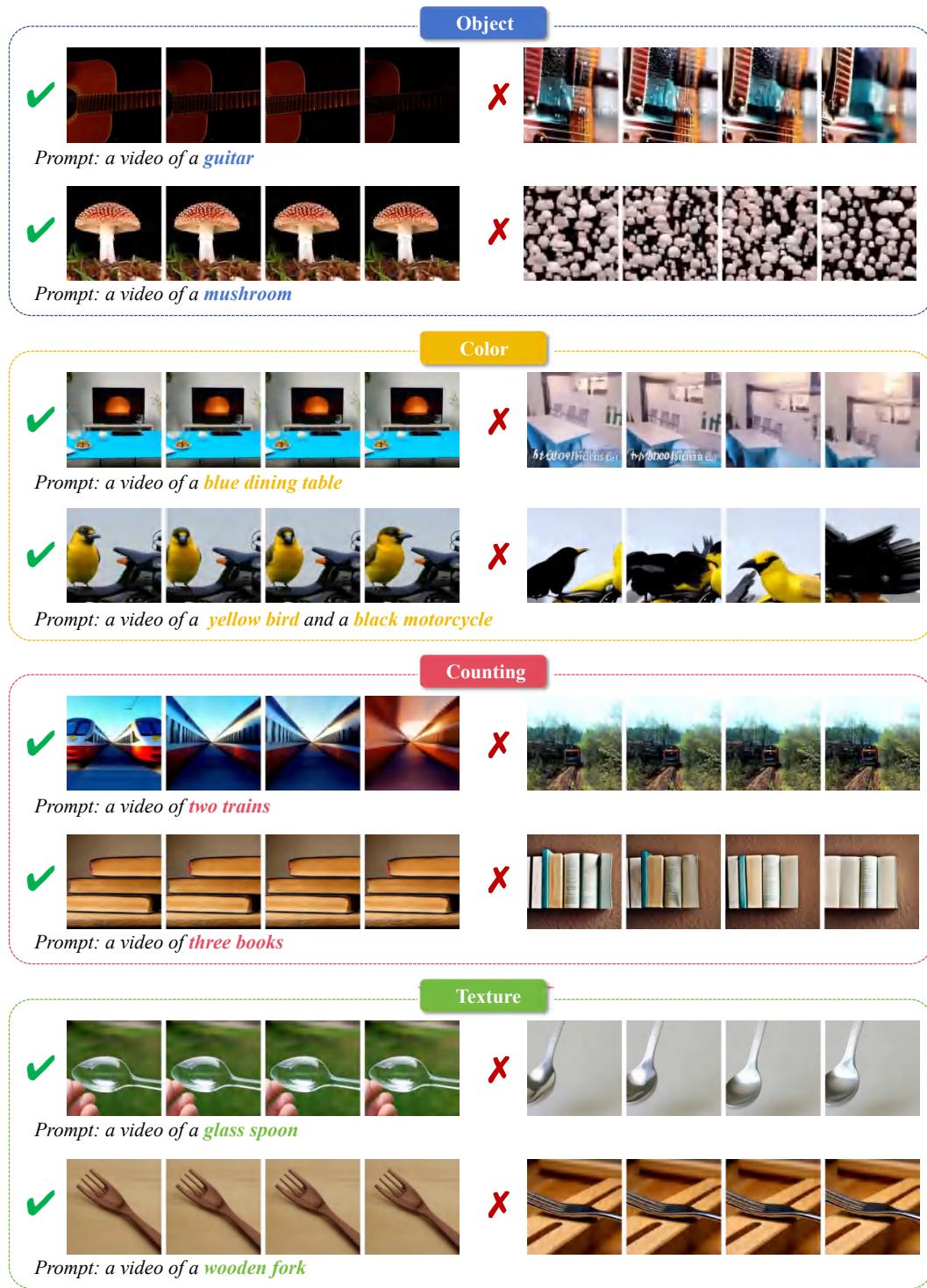


Figure 14: Examples for different task-specific challenges.

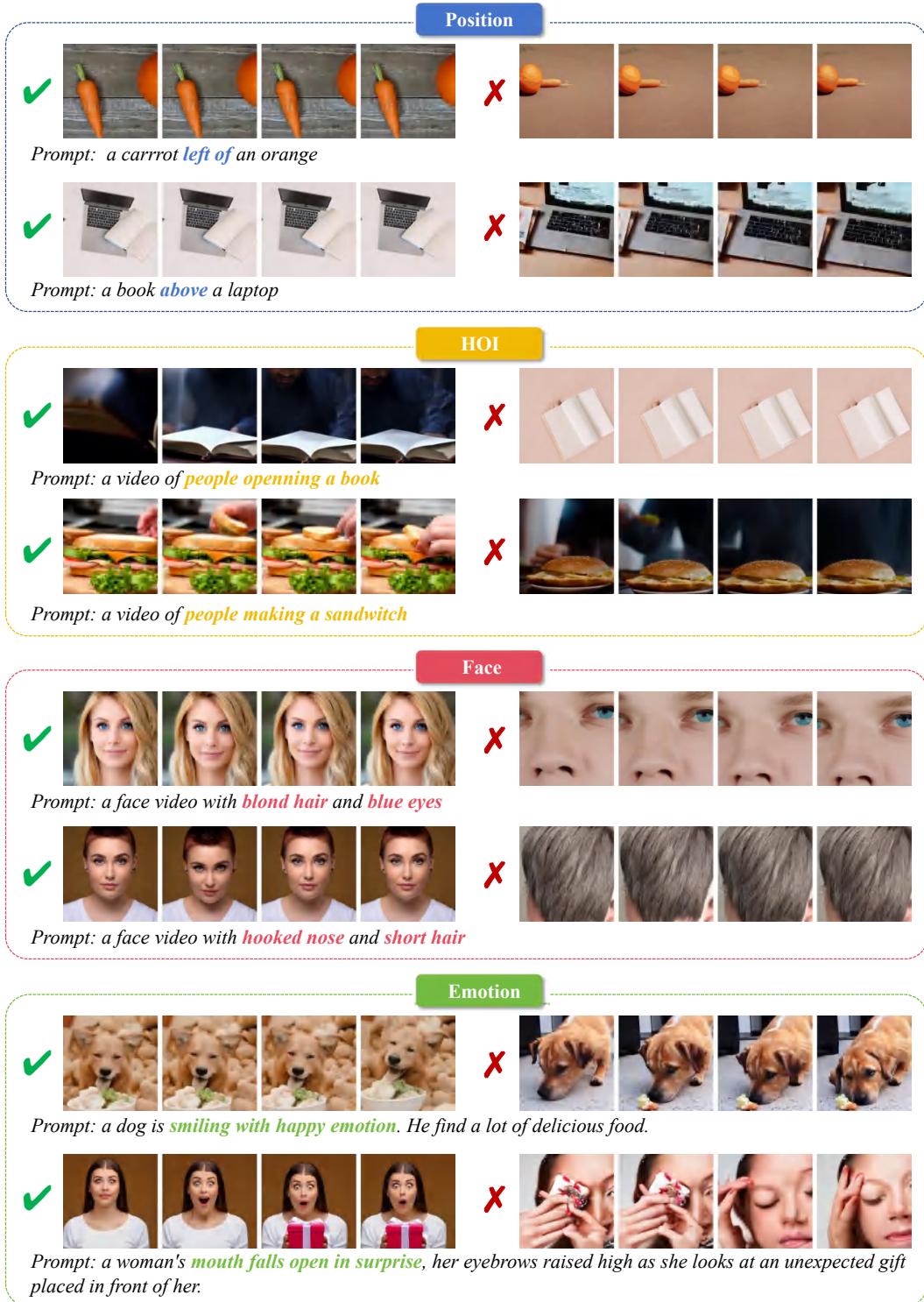


Figure 15: Examples for different task-specific challenges.

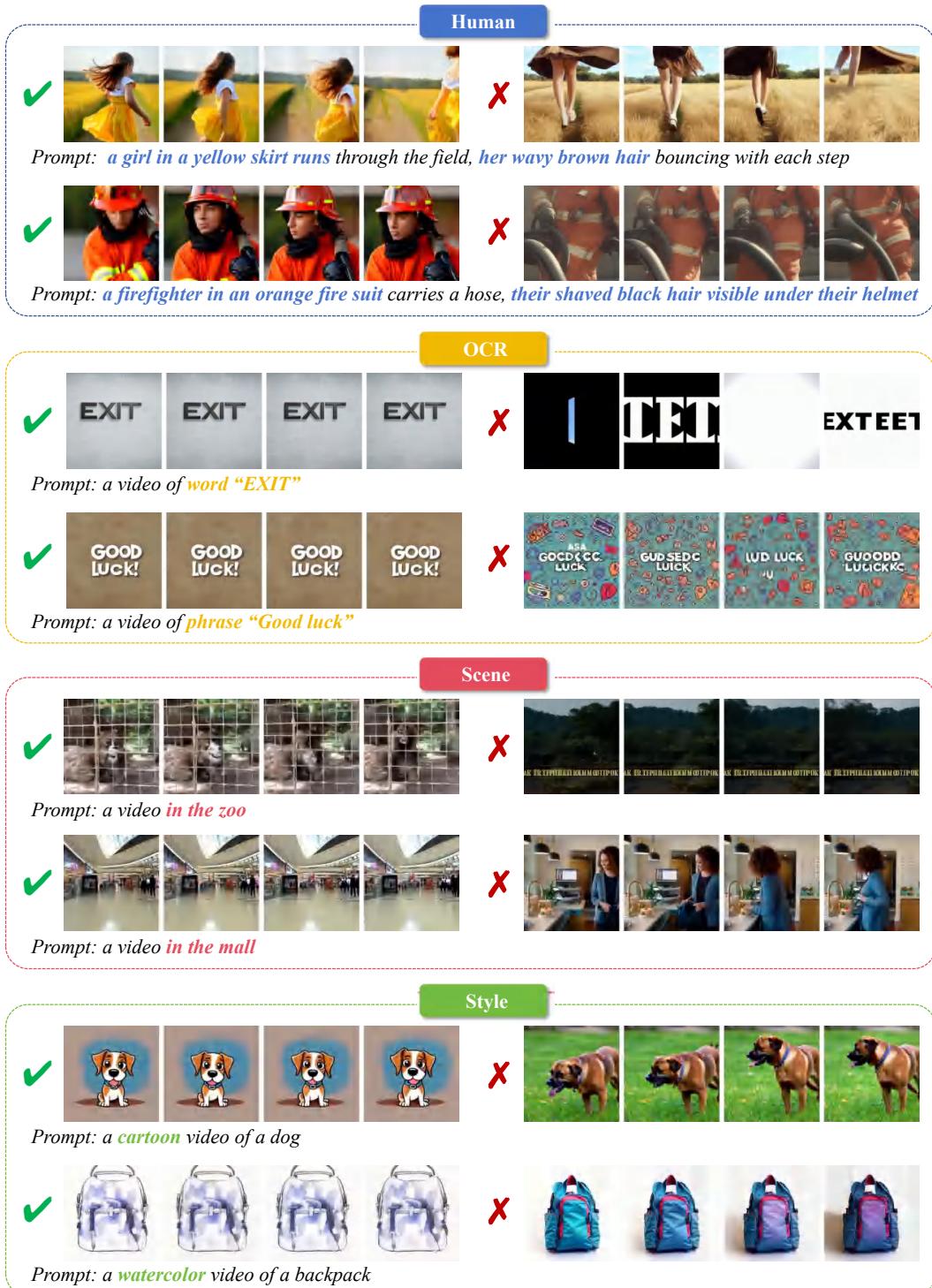


Figure 16: Examples for different task-specific challenges.

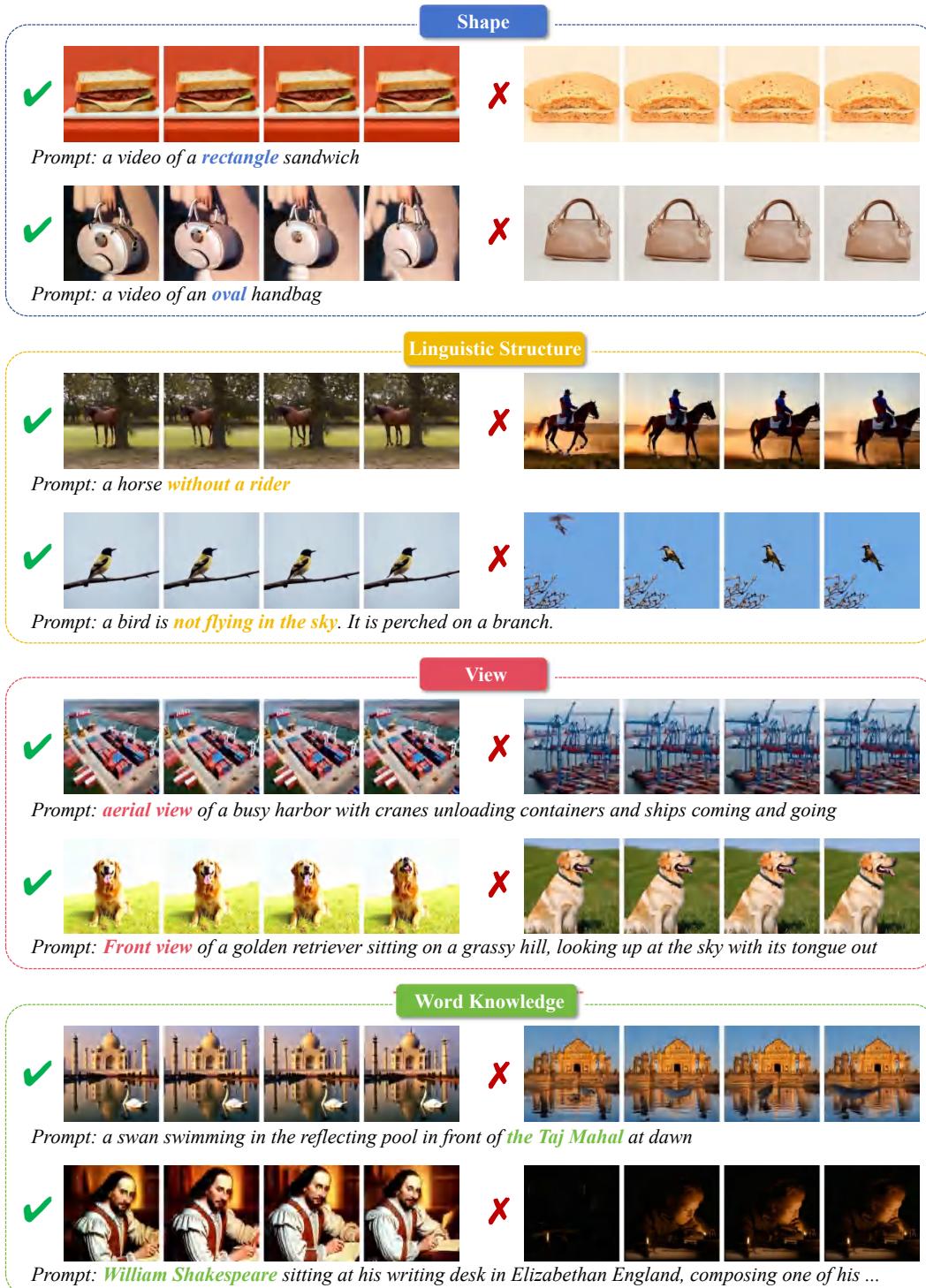


Figure 17: Examples for different task-specific challenges.

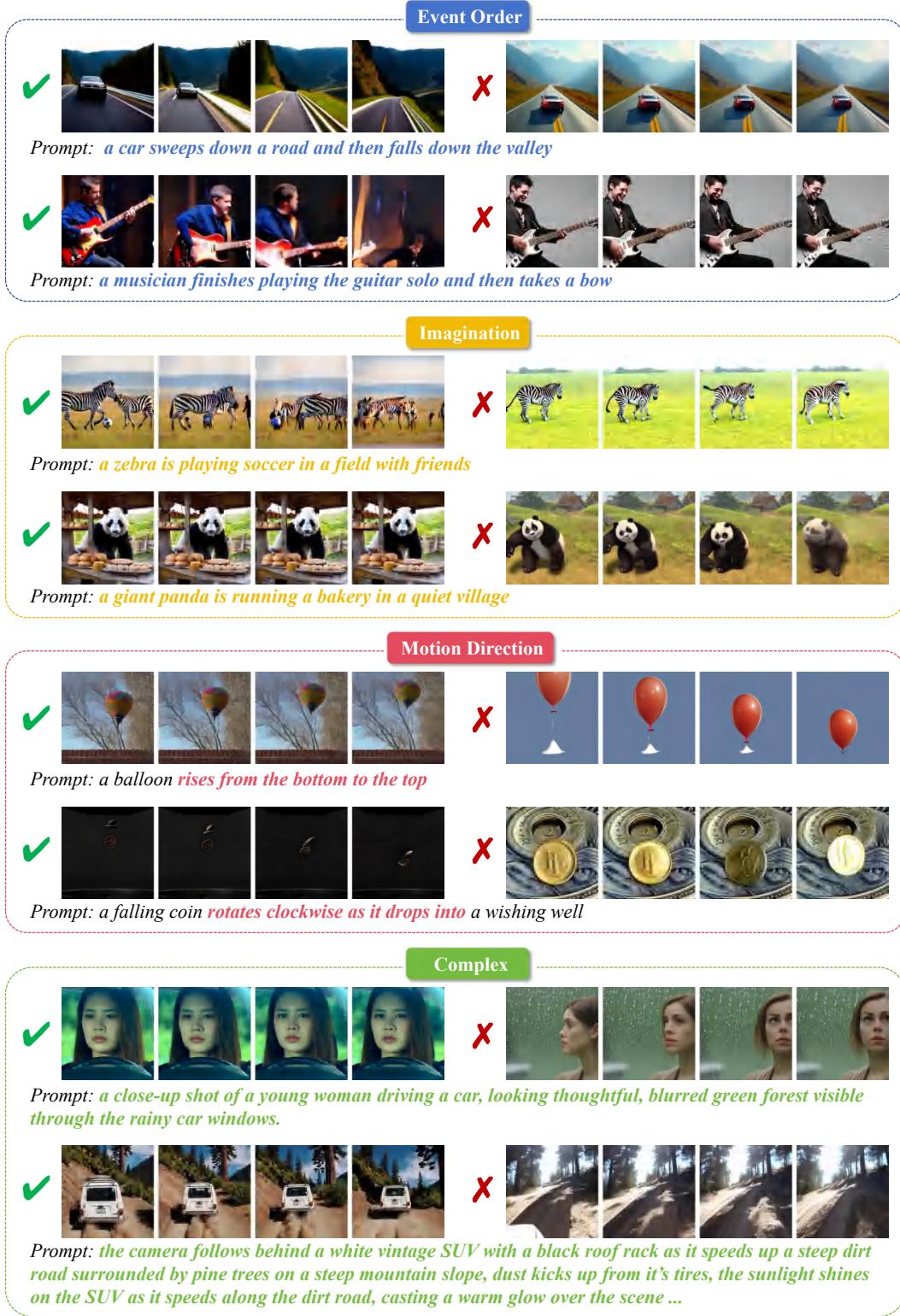


Figure 18: Examples for different task-specific challenges.