AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM (Supplementary Material)

Jiarui Wang[1], Huiyu Duan[1,2], Guangtao Zhai[1,2], Juntong Wang[1], Xiongkuo Min[1,*]
[1]Institute of Image Communication and Network Engineering,
[2] MoE Key Lab of Artificial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China

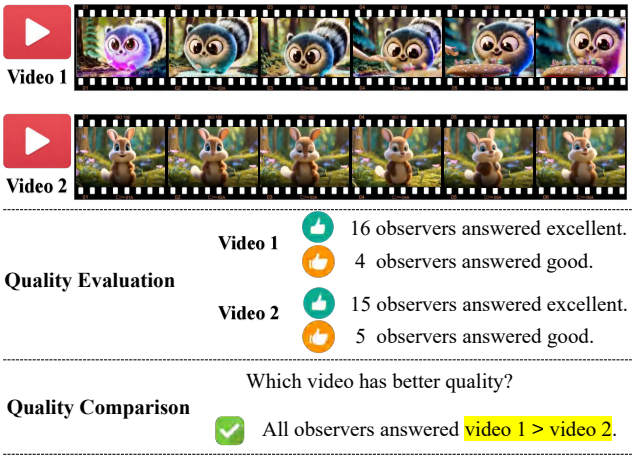


Figure 1. The motivation for visual quality comparison: single stimulus absolute ratings like "excellent" or "good" often involve randomness or inconsistency due to varying personal standards. Double stimuli comparative settings avoid the ambiguity of absolute evaluations for single videos, providing clearer and more consistent judgments.

1. Significance of AIGVQA-DB Construction

Mean opinion scores (MOS) have traditionally served as the primary metric for measuring overall quality. While MOS is effective for providing a general indication of quality, it has notable limitations, especially when it comes to high-quality content. For example, when evaluating closely matched, high-quality images or videos, MOS often results in similar scores across samples, leading to coarse evaluations that fail to capture subtle differences in factors like exposure, motion smoothness, or color fidelity. As illustrated in Figure 1, human assessors applying absolute scoring frequently yield inconsistent ratings due to varying personal standards or subjective preferences. Despite this, when asked to make relative comparisons—such as deciding whether "video1 is better than video2", they exhibit greater consistency and are able to reach a reliable consensus. This highlights a crucial insight: relative comparisons offer more precision and consistency than absolute scoring alone. Pairwise comparisons, which focus on directly comparing two samples, have thus emerged as a valuable complement to MOS. By emphasizing relative differences, pairwise assessments allow for finer granularity, capturing nuanced distinctions that absolute scores may miss. Numerous studies have demonstrated the effectiveness of pairwise comparisons in reducing ambiguity in scoring and providing more detailed evaluations, especially when assessing high-quality content [20, 30, 41, 44, 45].

The development of AI-generated image quality assessment (AIGIQA) datasets is already relatively well-established, incorporating both MOS for absolute quality evaluation and pairwise comparisons for assessing relative quality differences. This dual approach has proven effective in capturing both the overall and relative quality aspects of AI-generated images. However, existing AI-generated video quality assessment (AIGVQA) datasets primarily rely on MOS alone, which significantly limits their ability to capture the fine-grained quality differences inherent in video content. Videos, unlike static images, present unique challenges such as temporal coherence, spatial consistency, and the dynamic nature of objects and motion. These challenges make pairwise comparisons particularly valuable in video quality assessment, as they allow for more accurate evaluations of complex attributes like motion smoothness and frame-to-frame consistency. Unfortunately, most current VQA datasets lack systematic, large-scale pairwise data, limiting their capacity to assess these dynamic and temporal aspects effectively. To address this gap, we propose the AIGVQA-DB dataset, including both MOS and pairwise comparison data. By incorporating both absolute quality judgments and relative assessments, AIGVQA-DB enables more accurate model training and evaluation, allowing models to learn not only the absolute quality standards but also the relational nuances inherent in video sequences.

*Corresponding Author.

Table 1. Prompt categorizations with subcategories, detailed descriptions, and representative keyword examples.

Category	Subcategory	Descriptions	Keyword examples
Spatial major content	People	Prompts that include humans.	person, man, woman, men, women, kid, girl, boy, baby
	Plants	Prompts that include plants.	flower, leaf, tree, grass, forest, wheat, plant, peony
	Animals	Prompts that include animals.	panda, dog, cat, elephant, horse, bird, butterfly, rabbit
	Vehicles	Prompts that include vehicles.	car, van, plane, tank, carriage, rocket, motorcycle
	Artifacts	Prompts that include human-made objects.	robot, doll, toy, microphone, paper, plate, bowl, ball
	Illustrations	Prompts that include geometrical objects and symbols.	abstract, pattern, particle, gradient, loop, graphic, line
	Food and beverage	Prompts that include food and beverage.	water, wine, coffee, apple, butter, egg, chocolate, lime
	Buildings and infrastructure	Prompts that include buildings and infrastructure.	room, building, bridge, court, concert, hotel, factory
	Scenery and natural objects	Prompts that include lifeless natural objects and scenery.	wind, sand, snow, rain, sky, fog, mountain, river, sun
Temporal major content	Actions	Prompts that include the motion of solid objects	sing, dance, laugh, cry, smile, jump, walk, eat, drink
	Kinetic motions	Prompts that include the motion of solid objects.	fly, spin, race, move, rotate, fall, rise, bounce, sway
	Fluid motions	Prompts that include the motions of fluids or like fluids.	waterfall, wave, fountain, smoke, steam, inflate, melt
	Light change	Prompts from which the generated videos may involve light change.	sunset, sunrise, firework, shine, glow, burn, flash, bright
Attribute control	Color	Prompts that include colors.	white, pink, black, red, green, purple, blue, yellow
	Quantity	Prompts that include numbers.	one, two, three, four, five, six, seven, eight, nine, ten
	Camera view	Prompts that include control over the camera view.	view, macro, film, close, capture, aerial, shot, camera
	Speed	Prompts that include control over speed.	fast, slow, rapid, speed, motion, time, quick, swift, lag
	Event order	Prompts that include control over the order of events.	then, before, after, first, second
	Motion direction	Prompts that include control over the motion direction.	forward, backward, from, into, through, out of, left, right
Prompt complexity	Simple	Prompts that involve $0 \sim 8$ non-stop words.	-
	Medium	Prompts that involve $9 \sim 11$ non-stop words.	-
	Complex	Prompts that involve more than 11 non-stop words.	-

2. More Details of Video Generation

2.1. Detailed Information of Prompts

The AIGVQA-DB dataset offers a rich and diverse collection of prompts, carefully constructed from two sources: (1) existing open-domain text-video pair datasets, including InternVid [35], MSRVTT [42], WebVid [8], TGIF [25], FETV [26] and Sora website [7]. These datasets contribute a robust foundation of real-world and generalizable scenarios, providing a solid basis for training and evaluation. (2) manually written prompts designed to push the boundaries of model robustness and generalization. Inspired by unique categories such as "imagination" and "conflicting", these prompts introduce rare or non-realistic scenarios, like "A panda is flying in the sky," that test a model's ability to handle creative and unconventional inputs. As illustrated in Table 1, each prompt in our dataset is categorized based on four key aspects, including "spatial major content", "temporal major content", "attribute control", and "prompt complexity". For each aspect, we include typical elements that frequently occur in daily life. As shown in Figure 5, the spatial major content focuses on objects described in the prompt, including ten subcategories: people, animals, plants, and *etc.* In contrast, temporal major content highlights dynamic actions or changes and is divided into four subcategories, including actions, kinetic motions, fluid motions, and light change, as shown in Figure 6. Similarly, the attribute control covers specific stylistic or compositional controls embedded in prompts, enabling nuanced customization of generated content, including color, quantity, camera view, speed, motion direction, and event order, as shown in Figure 7. Additionally, we classify prompt complexity into three levels including: simple, medium, and complex, based on the number of descriptive elements in the text. By integrating real-world scenarios, imaginative constructs, and a structured categorization system, AIGVQA-DB ensures a comprehensive evaluation framework that challenges text-to-video generation models in both realistic and highly creative contexts.

To ensure a comprehensive and systematic classification of prompts within the AIGVQA-DB dataset, we employed the GPT-4 [28] API for multi-aspect prompt categorization. The GPT-4 [28] was provided with task-specific instructions designed to guide its classification process. These instructions included detailed descriptions of the categorization task, along with illustrative examples to ensure consistent and accurate labeling. Prompts were analyzed based on key aspects. For instance, to classify spatial major content, the GPT-4 [28] was prompted with a detailed instruction template, such as:

> *"Analyze the following prompt and classify it into one or more categories based on the type of object it describes. The categories include People, Buildings and Infrastructure, Animals, Artifacts, Vehicles, Plants, Scenery and Natural Objects, Food and Beverage, and Illustration. Provide only the category names as the output. Example: 'A cat is sitting under a tree.' Spatial major content: Animals, Plants."*

Under this framework, GPT-4 [28] processes the given prompt and assigns appropriate category labels based on its analysis. For example, a prompt like "A dog is driving a

Table 2. Video formats and numbers generated by the 15 text-to-video (T2V) models in the AIGVQA-DB. ✓ in the Pairs and MOS columns indicate which generative models are utilized in each of the two subsets. † Representative variable. *Representative open-source.

Models	Number	Prompts	Frames	FPS	Resolution	MOS	Pairs	URL
*CogVideo [16]	4,000	1,000	32	10	480×480	-	✓	https://github.com/THUDM/CogVideo
*LVDM [13]	4,048	1,048	16	8	256×256	✓	✓	https://github.com/YingqingHe/LVDM
*Tune-A-Video [39]	4,048	1,048	8	8	512×512	✓	✓	https://github.com/showlab/Tune-A-Video
*VideoFusion [27]	4,048	1,048	16	8	256×256	✓	✓	https://github.com/modelscope/modelscope
*Text2Video-Zero [19]	4,048	1,048	8	4	512×512	✓	✓	https://github.com/Picsart-AI-Research/Text2Video-Zero
*Lavie [34]	4,048	1,048	16	8	512×320	✓	✓	https://github.com/Vchitect/LaVie
*VideoCrafter [10]	4,048	1,048	16	10	1024×576	✓	✓	https://github.com/AILab-CVC/VideoCrafter
*Hotshot-XL [1]	4,048	1,048	8	8	672×384	✓	✓	https://github.com/hotshotco/Hotshot-XL
*StableVideoDiffusion [9]	1,000	1,000	14	6	576×1024	-	✓	https://github.com/Stability-AI/generative-models
Floor33 [2]	4,048	1,048	16	8	1024×640	✓	✓	https://discord.com/invite/EuB9KT6H
Genmo [3]	4,048	1,048	60	15	2048×1536†	✓	✓	https://www.genmo.ai
Gen-2 [4]	48	48	96	24	1408×768	✓	-	https://research.runwayml.com/gen2
MoonValley [5]	48	48	200†	50	1184×672	✓	-	https://moonvalley.ai
MorphStudio [6]	4,000	1,000	72	24	1920×1080	-	✓	https://www.morphstudio.com
Sora [7]	48	48	600†	30	1920×1080†	✓	-	https://openai.com/research

car." would be classified under Animals and Vehicles. Similar categorization instructions were devised for temporal major content and attribute control. Additionally, prompt complexity is classified based on the number of non-stop words present in each prompt. This multi-aspect categorization approach ensured that every prompt in the AIGVQA-DB was exhaustively labeled, facilitating fine-grained evaluation of text-to-video models. Examples of prompts and their corresponding categorizations in AIGVQA-DB are shown in Table 5.

2.2. Detailed Information of Text-to-Video Models

To construct AIGVQA-DB, we utilize 15 state-of-the-art text-to-video generative models, encompassing both open-source and closed-source methods, as detailed in Table 2. For open-source models, we rely on official repositories and use default weights to standardize results and maintain consistency across experiments. For closed-source models, we leverage publicly available APIs from open-source platforms. This comprehensive selection ensures that AIGVQA-DB serves as a robust benchmark for evaluating text-to-video generation systems.

CogVideo. CogVideo [16] is built on the text-to-image model CogView2 [12]. It employs a multi-frame-rate hierarchical training strategy to ensure better alignment between text and temporal counterparts in videos, generating keyframes based on textual prompts and recursively interpolating intermediate frames for coherence.

LVDM. LVDM [13] 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.

Tune-A-Video. Tune-A-Video [39] is a one-shot text-to-video generation model that extends text-to-image (T2I) 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 T2I models.

VideoFusion. VideoFusion [27] 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 image-generation models for efficient frame content prediction while maintaining motion dynamics.

Text2Video-Zero. Text2Video-Zero [19] 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. We adopt its official code with default parameters (<motion_field_strength_x&y=12>).

LaVie. LaVie [34] is an integrated video generation framework that operates on cascaded video latent diffusion models. For each prompt, we use the base T2V model and sample 16 frames of size 512×320 at 8 FPS. The number of DDPM [15] sampling steps and guidance scale are set as 50 and 7.5, respectively.

VideoCrafter. VideoCrafter [10] is a video generation and editing toolbox. We sample 16 frames of size 1024×576 at 8 FPS, according to its default settings.

Hotshot-XL. Hotshot-XL [1] is a text-to-gif model trained to work alongside Stable Diffusion XL[1]. We adopt its official code with default parameters and change the output format from GIF to MP4.

Genmo. Genmo [3] is a high-quality video generation platform. We generate 60 frames of size ≤2048×1536 at 15

[1] https://huggingface.co/hotshotco/SDXL-512

FPS for each prompt. The motion parameter is set to 70%.

Gen-2. Gen-2 [4] is a multimodal AI system, introduced by Runway AI, Inc., which can generate novel videos with text, images or video clips. We collect 96 frames of size 1408×768 at 24 FPS for each prompt.

Sora. Sora [7] is particularly known for its ability to handle complex, multi-element prompts, ensuring coherent visual representations of diverse scenarios. Sora [7] currently does not have an open-source API, so the videos we used are downloaded from its official website.

Floor33, MoonValley and MorphStudio. Floor33 [2], MoonValley [5], and MorphStudio [6] are recent popular online video generation application. We use the T2V mode of these applications via commands in Discord[2].

3. More Details of Subjective Experiment

3.1. Annotaion Criteria

The assessment criteria for AIGVQA-DB are systematically structured across four key dimensions: static quality, temporal smoothness, dynamic degree, and text-video correspondence. These dimensions provide a comprehensive framework for video quality assessment, ensuring thorough and reliable assessments through clearly defined scales, detailed annotation criteria, and illustrative reference examples.

- **Static quality** focuses on the video's visual clarity, naturalness, color balance, and detail richness. 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. Conversely, low scores reflect videos with blurriness, unnatural color tones, faded visuals, and 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 8.
- **Temporal smoothness** evaluates the consistency and fluidity of frame-to-frame transitions, and the naturalness of object movements within the video. Videos with high scores exhibit seamless transitions, smooth movements, and no noticeable inconsistencies, creating a natural and immersive viewing experience. Low scores denote irregular or abrupt frame changes and disjointed object movements, which detract from the overall fluidity. For detailed criteria, refer to Figure 9.
- **Dynamic degree** assesses the range and expressiveness of motion within the video. High-scoring videos display diverse, realistic, and natural movements of objects, animals, or humans, contributing to a vivid and engaging experience. Lower scores indicate limited motion or unnatural dynamics. This dimension highlights the importance of motion diversity and realism in engaging content. For detailed criteria, refer to Figure 10.
- **Text-video correspondence** examines the alignment between the video content and its associated text prompt. 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 omissions or mismatches. In contrast, videos with lower scores exhibit inconsistencies, missing elements, or mismatched content. For detailed criteria, refer to Figure 11.

Each of these four dimensions is supported by detailed examples, providing annotators with clear guidelines to perform evaluations. This systematic approach ensures accuracy and consistency in the annotation process, enabling a robust analysis of human preference and video quality.

3.2. Annotation Interface

To ensure a comprehensive and efficient evaluation of video quality, we designed two custom annotation interfaces tailored for different assessment tasks: one for score annotation and the other for pair annotation. The score annotaion interface, shown in Figure 2, is a manual evaluation platform developed using the Python tkinter package, designed to facilitate MOS assessments. 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 512×512 pixels. The duration of the videos remains unaltered, preserving the full content described in the associated text prompts. Meanwhile, the pair annotation interface, illustrated in Figure 3, supports paired comparison assessments, where participants evaluate two videos side-by-side. This interface is designed to explore preference judgments across four key aspects, including: static quality, temporal smoothness, dynamic degree, and text-Video correspondence. In each comparison, participants are shown two videos, labeled "A" and "B", and are required to select their preferred video for each aspect. The interface ensures an unbiased evaluation environment by clearly distinguishing between the two videos while allowing side-by-side playback. Participants can replay either video as needed before making their selection. The evaluation process emphasizes subjective preferences while offering a structured approach to gather comparative insights across multiple dimensions. Navigation options, such as "Replay", "Next", and "Save", streamline the workflow, enabling efficient annotation

3.3. Annotation Management

To ensure ethical compliance and the quality of annotations, we implemented a comprehensive process for the AIGVQA-DB dataset. All participants were fully informed about the experiment's purpose, tasks, and ethical considerations. Each participant signed an informed consent

[2] https://discord.com

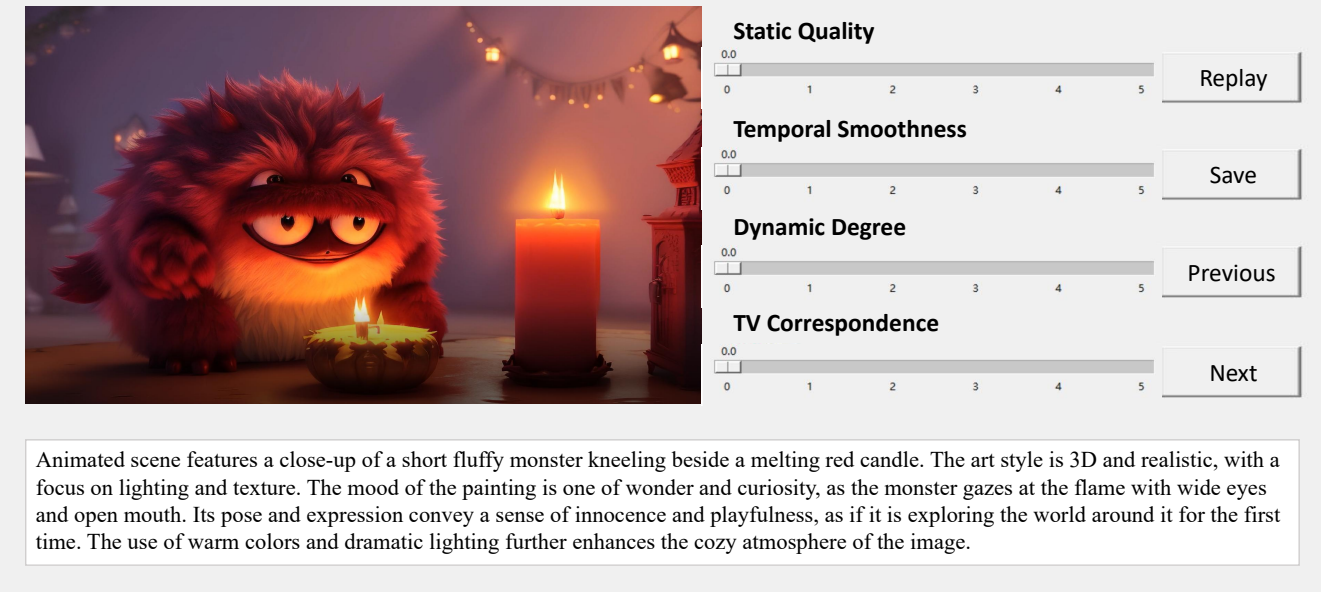


Figure 2. An example of the rating assessment interface for human evaluation. The subjects are instructed to rate four dimensions of AI-generated videos, i.e., static quality, temporal smoothness, dynamic degree, and text-video correspondence, based on the given video and its prompt.

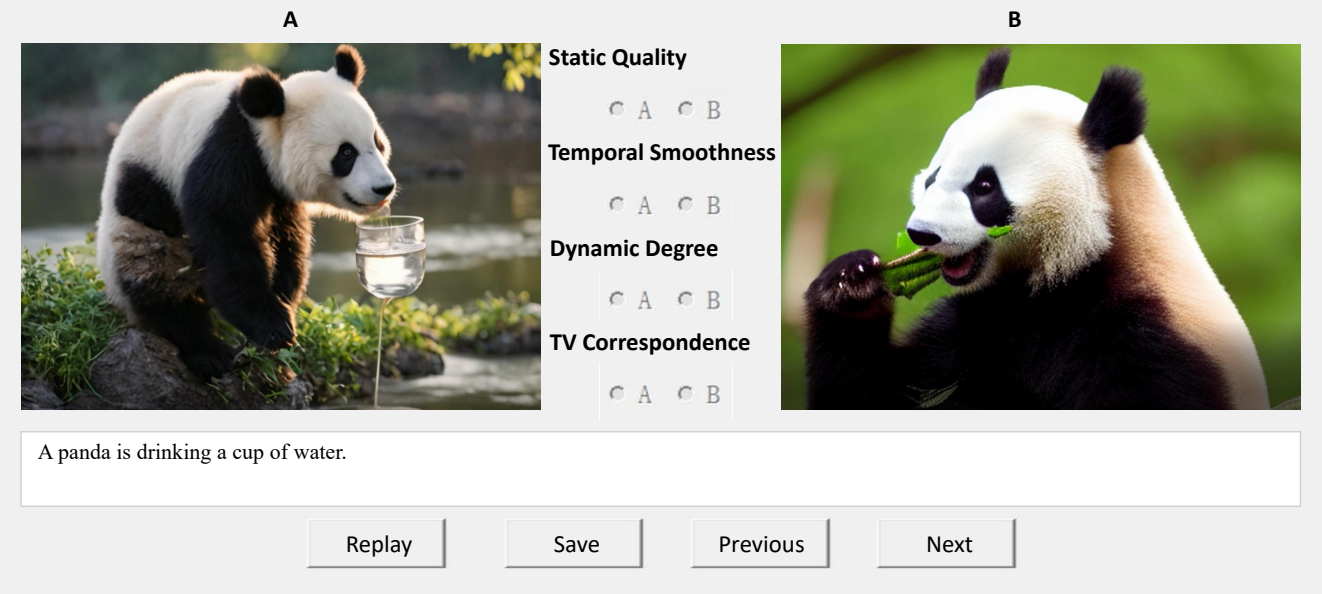


Figure 3. An example of the pair comparison assessment interface for human evaluation. The subjects are instructed to choose which AI-generated video is better among the video pairs, considering four dimensions respectively, i.e., static quality, temporal smoothness, dynamic degree, and text-video correspondence.

agreement, granting permission for their subjective ratings to be used exclusively for non-commercial research purposes. The dataset, consisting of 36,576 AI-generated videos (AIGVs) and their associated prompts, is publicly released under the CC BY 4.0 license. We ensured the exclusion of all inappropriate or NSFW content (textual or visual) through a rigorous manual review during the video generation stage. The annotation was divided into two key components: paired comparison annotation and MOS annotation, each designed to evaluate videos across four dimensions, including: static quality, temporal smoothness, dynamic degree, and text-video correspondence. For the paired comparisons, 30,000 video pairs were evaluated by a total of 100 participants. Each pair was assessed by three participants, and the final result for each pair was determined by majority voting. In cases of discrepancies, the

average opinions of the three participants were calculated to resolve the tie. This approach ensured a balanced and fair evaluation of preferences between video pairs. The MOS annotation task involved 20 participants to rate all videos in the MOS subset individually. Participants scored each video on a 0-5 Likert scale across the four evaluation dimensions. This granular scoring provided a comprehensive dataset for analyzing human preferences and video quality.

Before participating in the annotation tasks, all participants underwent a rigorous training process. They were provided with detailed instructions, multiple standard examples (Figures 8-11), and step-by-step guidance on the annotation criteria. A pre-test was conducted to evaluate participants' understanding of the criteria and their agreement with standard examples. Those who did not meet the required accuracy were excluded from further participation. During the experiment, all evaluations were conducted in a controlled laboratory environment with normal indoor lighting. Participants were seated at a comfortable viewing distance of approximately 60 cm from the screen. To further reduce potential biases, videos from different models were alternately presented in the both MOS and pair comparison tasks. Although individual preferences may vary, the use of detailed explanations and standardized annotation criteria ensured a high level of agreement across participants. This consensus was particularly evident in pair annotations, where majority voting captured group preferences effectively. The documentation of the entire annotation process served as a reference and training standard, ensuring consistency and reliability across all evaluations. This rigorous annotation management strategy makes AIGVQA-DB a robust and ethically sound resource for advancing research in video quality assessment.

4. More Details of AIGVQA-DB

4.1. Detailed Information of the Subsets

Construction of the MOS subset. The MOS subset is specifically designed to evaluate the perceptual quality of videos generated by T2V models, offering a comprehensive benchmark for subjective evaluation. This subset incorporates contributions from 12 generative models in the database, encompassing a broad spectrum of temporal and spatial attributes to ensure diversity. To construct this subset, we initially sourced 48 high-quality videos and their corresponding textual prompts from the Sora platform [7]. These prompts were then used to generate additional videos utilizing 11 other generative models, resulting in a total of 576 videos (48 prompts × 12 generative models). This approach ensured the inclusion of a wide range of visual styles and generative qualities. The dataset spans significant variations in frame count, frame rate (FPS), and resolution, ranging from the compact 256×256 outputs of VideoFusion [27] at 8 FPS to the high-definition 1920×1080 outputs of Sora at 30 FPS. Such diversity in video attributes allows for a robust analysis of generative models under different visual and temporal conditions. Each video in the MOS subset is evaluated by 20 annotators across four dimensions: static quality, temporal smoothness, dynamic degree, and text-video correspondence. This rigorous evaluation process results in 46,080 individual ratings (4 dimensions × 576 videos × 20 annotators). The annotators, equipped with detailed training and examples, provide subjective scores on a 0-5 Likert scale, ensuring consistency and reliability in their assessments. By including videos with diverse visual properties, the MOS subset provides a robust foundation for subjective evaluation tasks, enabling researchers to compare T2V models based on the perceptual quality of AIGVs.

Construction of the Pair comparison subset. To enable detailed comparative analysis, we construct the pair comparison subset. This subset is built based on 1,000 carefully curated textual prompts, including a wide range of scenarios, themes, and levels of complexity. These prompts ensure diversity in content and provide a robust basis for assessing the performance of generative models across various contexts. We use 12 generative models, including 8 open-source models such as Hotshot-XL [1] and Floor33 [2], and 4 closed-source models, such as Gen-2 [4] and MoonValley [5]. For each prompt, open-source models generate four distinct videos, capturing variations in their generative outputs and showcasing intra-model diversity. Closed-source models, due to access constraints, produce one video per prompt. This comprehensive approach results in a dataset of 36,000 videos (1,000 prompts × (8 open-source models × 4 videos + 4 closed-source models × 1 video)). The videos in this subset exhibit a wide range of resolutions and frame counts, from the lower-resolution 256×256 outputs of LVDM [13] to the high-definition 1920×1080 videos from MorphStudio [6]. Each video pair is evaluated by three annotators, who provide individual ratings for all four dimensions. These ratings are aggregated to determine the final result for each dimension, with majority voting or averaged scores used to resolve any disagreements. This process results in 360,000 (4 × 30,000 × 3) ratings, ensuring a rigorous and nuanced analysis of model performance. The pair-comparison subset allows for head-to-head comparisons of generative models, and provides valuable insights into the strengths and weaknesses of different models, offering researchers a robust foundation for comparative studies.

4.2. More Result Analysis

We analyze the subjective pair ratings by calculating the win rates of different generation models across different categories, revealing strengths and weaknesses from four different dimensions. For the evaluation of spatial content categories, as shown in Figure 4(a), models like Genmo [3]

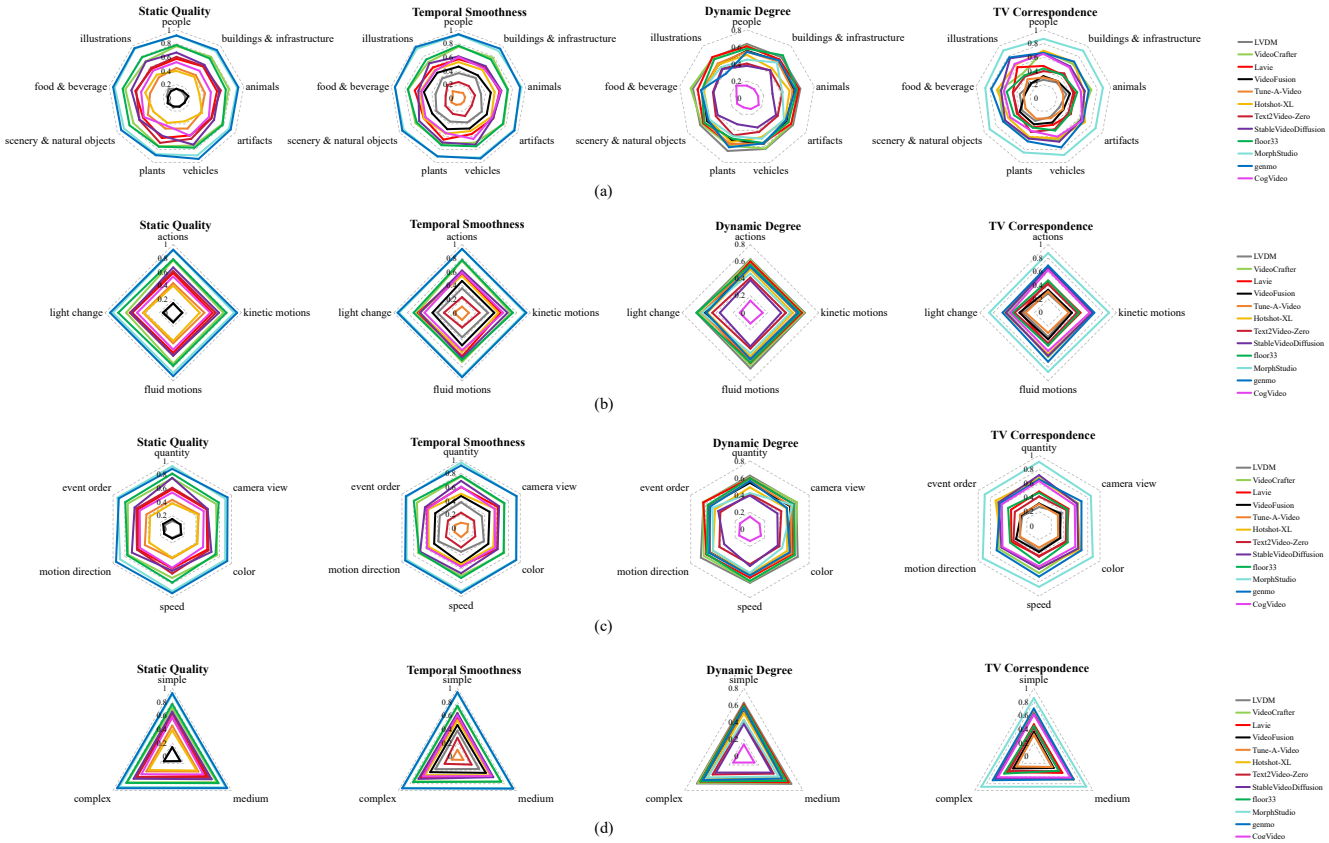


Figure 4. Comparison of averaged win rates of different generation models across different categories. (a) Results across spatial major contents. (b) Results across temporal major contents. (c) Results across dynamic degrees. (d) Results across text-to-video correspondence.

perform exceptionally well in generating realistic representations of people, animals, and vehicles showcasing their strong attention to detail and visual fidelity. MorphStudio [6] consistently leads in producing high-quality outputs for scenery and natural objects, excelling in generating visually appealing and immersive natural environments. Additionally, StableVideoDiffusion [9] demonstrates notable strength in creating illustrations, highlighting its flexibility in handling stylized and artistic content. Conversely, LVDM [13] and VideoFusion [27] lag in these categories, struggling with resolution and detail preservation. For the evaluation of temporal content categories, as shown in Figure 4(b), MorphStudio [6] excels in handling kinetic motions, fluid motions, actions, and scenarios with light changes, making its outputs maintain high temporal smoothness and text-video correspondence. However, models like Text2Video-Zero [19] occasionally produce abrupt transitions, and Tune-A-Video [39] shows limitations in maintaining temporal smoothness under complex motion conditions. For the evaluation of attribute control categories, as shown in Figure 4(c), Genmo [3] performs well in maintaining appropriate quantities of objects. Floor33 [2] and VideoCrafter [10] display superior performance in the logical sequence of events. In contrast, StableVideoDiffusion [9] encounters challenges in event order. Its generative process involves first creating static images and subsequently animating them to produce video sequences. The static-to-dynamic generation pipeline introduces discrepancies in temporal alignment, making it difficult to ensure that actions unfold in a logically consistent manner. For the evaluation of prompt complexity categories, as shown in Figure 4(d), most models demonstrate competence in handling prompts of different complexity, likely due to shared architectures like diffusion-based systems, with common strengths and limitations in handling complex prompts.

5. Details of Loss Function

The training process for AIGV-Assessor is divided into three progressive stages, each utilizing a specific loss function to target distinct objectives: language loss for aligning visual and language features, L1 loss for generating accurate quality scores, and cross-entropy loss for robust pairwise video quality comparisons.

(1) Aligning visual and language features with language loss. In the first stage, spatial and temporal projectors are trained to align visual and language features using the language loss. This involves ensuring that the visual tokens extracted from the vision 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:

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

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) Refining quality scoring 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 AIGVQA-DB, 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:

$$\mathcal{L}_{\text{MOS}} = \frac{1}{N}\sum_{i=1}^{N}|Q_{\text{predict}}(i) - Q_{\text{label}}(i)| \tag{2}$$

where $Q_{predict}(i)$ is the score predicted by the regressor i and $Q_{label}(i)$ 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.

(3) Enhancing pairwise comparisons with cross-entropy Loss. The third stage incorporates the AIGVQA-DB subset into the training pipeline. This dataset contains human annotations for pairwise video comparisons, where two videos are evaluated, and the superior one is selected based on quality. Pairwise training helps the model learn relative quality distinctions, enabling it to compare videos effectively. The objective in this stage is to maximize the probability that the model predicts a higher score for the better video in a pair. The pairwise comparison loss is calculated by comparing the predicted scores for a video pair, which are processed through an LPIPS network to judge which video is better. This predicted logit is then compared with the ground-truth logit labels (0 or 1) using the cross-entropy loss. The label 0 indicates that video2 is better, and 1 indicates that video1 is better. The order of the pair (which video is considered as video1 or video2) is random, but the logit label always corresponds correctly to the better video.

$$\begin{aligned}\mathcal{L}_{\text{Pairs}} = -\frac{1}{N}\sum_{i=1}^{N}\big[&y_{\text{label}}(i)\log y_{\text{pred}}(i) \\ &+(1-y_{\text{label}}(i))\log(1-y_{\text{pred}}(i))\big]\end{aligned} \tag{3}$$

where y_{pred} is the logit predicted by the network for the video pair, y_{label} is the ground-truth label for the video pair (0 for video2 better, 1 for video1 better), and N is the number of videos in the batch. This function encourages the model to predict the better video in a pair, reinforcing its ability to make accurate comparisons. By incorporating the pairwise data into training, the model not only learns to provide accurate quality scores but also becomes proficient in comparing videos and selecting the superior one. This enhances its utility in real-world applications, where users often need to compare the quality of multiple videos directly.

6. Implemention Details

6.1. Detailed Information of Evaluation Criteria

We adopt the widely used metrics in VQA literature [11, 29]: 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:

$$SRCC = 1 - \frac{6\sum_{n=1}^{N}(v_n - p_n)^2}{N(N^2-1)}, \tag{4}$$

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:

$$PLCC = \frac{\sum_{n=1}^{N}(y_n - \bar{y})(\hat{y}_n - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})^2}}, \tag{5}$$

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:

$$(v_i - v_j)(p_i - p_j) > 0, \tag{6}$$

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

$$KRCC = \frac{C - D}{\frac{1}{2}N(N-1)}, \tag{7}$$

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

Table 3. Zero-shot and Cross-dataset performance on LGVQ [47].

Datasets	Metrics	Spatial		Metrics	Temporal		Metrics	Aligment	
		SRCC	PLCC		SRCC	PLCC		SRCC	PLCC
official raw weights →LGVQ [47] (Zero-shot)	MUSIQ [18]	0.389	0.431	VSFA [23]	0.295	0.451	HPS [40]	0.248	0.339
	StairIQA [33]	0.334	0.393	SimpleVQA [32]	0.271	0.419	CLIPScore [14]	0.372	0.405
	LIQE [46]	0.174	0.209	FastVQA [36]	0.374	0.473	BLIPScore [24]	0.379	0.389
	NIQE	0.228	0.293	DOVER [37]	0.254	0.514	ImageReward [43]	0.369	0.371
FETV [26]→LGVQ [47] (Cross-dataset)	MUSIQ [18]	0.406	0.404	VSFA [23]	0.388	0.398	HPS [40]	0.201	0.243
	StairIQA [33]	0.484	0.500	SimpleVQA [32]	0.419	0.407	CLIPScore [14]	0.168	0.205
	CLIP-IQA	0.493	0.501	FastVQA [36]	0.397	0.364	BLIPScore [24]	0.151	0.193
	LIQE [46]	0.461	0.477	DOVER [37]	0.427	0.406	ImageReward [43]	0.193	0.245
	UGVQ [47]	0.521	0.524	UGVQ [47]	0.442	0.432	UGVQ [47]	0.217	0.255
Ours → LGVQ [47]	Ours (MOS)	0.551	0.587	Ours (MOS)	0.479	0.509	Ours (MOS)	0.506	0.535
	Ours (Pairs)	**0.585**	**0.623**	Ours (Pairs)	**0.553**	**0.604**	Ours (Pairs)	**0.513**	**0.541**

6.2. Detailed Information of Evaluation Algorithms

V-Dynamic [17] and **V-Smoothness** [17] are proposed in VBench [17]. We directly used the respective implementation code in VBench [17] without specific changes.

CLIPScore [14] is an image captioning metric, which is widely used to evaluate T2I/T2V models. It passes both the image and the candidate caption through their respective feature extractors, then computing the cosine similarity between the text and image embeddings.

BLIPScore [24] provides more advanced multi-modal feature extraction capabilities. Using the same methodology as CLIPScore [14], 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 builds upon the BLIP model [24] 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 image over another in comparative settings.

AestheticScore is given by an aesthetic predictor introduced by LAION [31]. This metric evaluates the overall aesthetic appeal of an image by leveraging a pre-trained model fine-tuned on datasets annotated with human-judged aesthetic scores.

VSFA [23] 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. We directly used the official code without specific changes.

BVQA [22] leverages the transferred knowledge from IQA databases with authentic distortions and large-scale action recognition with rich motion patterns for better video representation. We used the officially pre-trained model under mixed-database settings and finetuned it on our AIGVQA-DB for evaluation.

SimpleVQA [32] 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. We used the officially pre-trained model and finetuned it on our AIGVQA-DB for evaluation.

FAST-VQA [36] 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. We used the officially released FAST-VQA-B model and finetuned it on our AIGVQA-DB.

DOVER [37] is a disentangled objective video quality evaluator that learns the quality of videos based on technical and aesthetic perspectives. We used the officially pre-trained model and finetuned it on our AIGVQA-DB.

Q-Align [38] is a human-emulating syllabus designed to train large multimodal models for visual scoring tasks. It mimics the process of training human annotators by converting MOS into five text-defined rating levels. We used the officially pre-trained model and finetuned it on our AIGVQA-DB.

Table 4. Zeto-shot alignment performance on T2VQA-DB [21], LGVQ [47] and FETV [26].

Metrics	T2VQA-DB [21]		LGVQ [47]		FETV [26]	
	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
CLIPScore [14]	0.102	0.070	0.372	0.405	0.243	0.177
BLIPScore [24]	0.166	0.111	0.379	0.389	0.309	0.224
ImageReward [43]	0.188	0.127	0.369	0.371	-	-
UMTScore [26]	0.068	0.045	-	-	0.425	0.309
Ours	**0.246**	**0.170**	**0.506**	**0.368**	**0.615**	**0.456**

7. More Results of AIGV-Assessor

To address the zero-shot ability concern, we conduct additional experiments, which manifests that our model outperforms other zero-shot methods, as shown in Tables 3 & 4. It should be noted that AIGV-Assessor achieves much better performance compared to the models trained on FETV in Table 3, especially for alignment score. The main dataset contribution of this paper is creating a large-scale pair comparison subset (similar to ImageReward for AIGI), designed for more granular quality distinction and better generalization. AIGV-Assessor has better zero-shot performance when finetuned on the Pairs compared to trained only on the MOS.

Table 5. Examples of prompts and their corresponding categorizations in AIGVQA-DB.

Prompts	Spatial major content	Temporal major content	Attribute control	Complexity	Source
“A person is running backwards.”	people	actions, kinetic motions	motion direction	simple	FETV [26]
“A plane is flying backwards.”	vehicles	kinetic motions	motion direction	simple	FETV [26]
“A blue horse is running in the field.”	animals	actions, kinetic motions	color	simple	FETV [26]
“A green shark is swimming under the water.”	animals	fluid motions, actions, kinetic motions	color	simple	FETV [26]
“A leave is flying towards the tree from the ground.”	plants	kinetic motions	motion direction	medium	FETV [26]
“The flowers first wilt and then bloom again.”	plants	fluid motions	event order	simple	FETV [26]
“The sun sets on the horizon and then immediately rises again.”	scenery and natural objects	kinetic motions	event order	medium	FETV [26]
“A person pours a cup of coffee from a bottle and then pours the coffee back to the bottle.”	people, food and beverage	actions, fluid motions	event order	complex	FETV [26]
“The three singers are dancing in swim suits.”	people	actions	quantity	simple	InternVid [35]
“a bearded man nods and blows kisses.”	people	actions	event order	simple	InternVid [35]
“A a dog are flipping and riding a skateboard.”	animals	actions	null	simple	InternVid [35]
“A white puppy plays with a slice of lime”	animals, food and beverage	kinetic motions	color	medium	InternVid [35]
“Rain is falling on a black umbrella.”	plants, scenery and natural object	fluid motions	color	simple	InternVid [35]
“Two cars are racing on a track.”	vehicles	kinetic motions	quantity	simple	InternVid [35]
“Two very handsome boys are singing on the stage.”	people	actions	quantity	medium	InternVid [35]
“Two girls are standing in the ocean when they become frightened of something in the water.”	people	actions	quantity, event order	complex	InternVid [35]
“An arial view of animals running”	animals	actions, kinetic motions	camera view	simple	MSRVTT [42]
“Overhead view as pingpong players compete on the table”	people	actions, kinetic motions	camera view	medium	MSRVTT [42]
“There is a orange color fish floating in the water”	animals	fluid motions	color	medium	MSRVTT [42]
“Some blue water in a pool is rippling around”	scenery and natural objects	fluid motions	color	medium	MSRVTT [42]
“A red sport car is driving very fast”	vehicles	kinetic motions	color, speed	medium	MSRVTT [42]
“Four friends are driving in the car”	scenery and natural objects	fluid motions	color	medium	MSRVTT [42]
“Smoke is coming out of a mountain”	scenery and natural objects	fluid motions	motion direction	simple	MSRVTT [42]
“Satellite view of moon we can also see sunlight but surface is not smooth”	scenery and natural objects	light change	camera view	complex	MSRVTT [42]
“Smoke billows from the factory chimney.”	vehicles, buildings and infrastructure	fluid motions	color	simple	Handwritten
“Leaves flutter from the trees in the gusty wind.”	plants	kinetic motions	motion direction	medium	Handwritten
“The crimson hues painted the horizon during the beach sunset.”	scenery and natural objects	light change	color	medium	Handwritten
“The static view of a solar eclipse revealed nature’s cosmic spectacle.”	scenery and natural objects	light change	camera view	medium	Handwritten
“A hiker reaches the summit and then admires the breathtaking view.”	people	actions	event order	medium	Handwritten
“Vinegar drizzling onto a salad, filmed in intricate detail.”	food and beverage	fluid motions	camera view	medium	Handwritten
“An egg cracking open and being whisked vigorously in slow motion.”	food and beverage	kinetic motions	speed	medium	Handwritten
“The coastline transformed into a canvas of fiery colors during the beach sunset.”	scenery and natural objects	light change	color	complex	Handwritten
“Two men playing musical instruments in a city square.”	people, artifacts	kinetic motions	quantity	medium	TGIF [25]
“The bridge of a river being viewed from a cable.”	scenery and natural object	kinetic motions	camera view	medium	TGIF [25]
“A view from inside of a bus showing snow”	vehicles, scenery and natural object	kinetic motions	camera view	medium	TGIF [25]
“Some large metal barrels on a train track.”	vehicles, artifacts	actions	quantity	simple	TGIF [25]
“A person reaching into a dish of beef and vegetables.”	people, food and beverage	actions	quantity	medium	TGIF [25]
“A man wearing a green jacket is fixing a solar panel.”	people	actions	color	medium	TGIF [25]
“A green toy chamelon eating a cookie.”	animals, food and beverage	kinetic motions	color	simple	TGIF [25]
“Two people sit on a table with headphones on”	people	actions	quantity	medium	TGIF [25]
“A screenshot of the dashboard of a korean language software”	illustrations	actions	camera view	medium	TGIF [25]
“A woman’s tight pants with two photos, one showing her wearing the pants and the other showing her with a shirt.”	people, artifacts	actions	quantity	complex	TGIF [25]
“Background - sunset landscape beach.”	scenery and natural object	light change	null	simple	WebVid [8]
“The musician plays the guitar. close up.”	people, artifacts	actions	camera view	simple	WebVid [8]
“Background - sunset landscape beach.”	scenery and natural object	light change	null	simple	WebVid [8]
“Background - sunset landscape beach.”	scenery and natural object	light change	null	simple	WebVid [8]
“Background - sunset landscape beach.”	scenery and natural object	light change	null	simple	WebVid [8]
“Attractive young woman silhouette dancing outdoors on a sunset with sun shining bright behind her on a horizon. slow motion.”	people, scenery and natural object	actions, light change	speed	complex	WebVid [8]
“Night landscape timelapse with colorful milky way. starry sky with tropical palms on the island. milky way timelapse over palms.”	plants, scenery and natural object	light change	speed	complex	WebVid [8]
“A blue wave of fire grows into a large flame and bright sparks on a shiny surface. closeup. slow motion, high speed camera.”	scenery and natural object	fluid motions, light change	camera view, color, speed	complex	WebVid [8]
“Silhouette of happy mom dad and baby at sunset in a field with wheat. farmer and family on the field. a child with parents plays in the wheat. the concept of family relationships.”	people, plants	actions, light change	null	complex	WebVid [8]
“Traditional chinese ink preparation. low angle dolly shot close up focus from brushes on ceramic stand to person hands in background preparing ink for calligraphy.”	artifacts, people	actions	camera view	complex	WebVid [8]
“Beautiful growing network with economic indicators growing abstract seamless. looped 3d animation of moving numbers and lines. cyberspace flashing lights. business concept. 4k ultra hd 3840x2160.”	illustrations	light change	camera view	complex	WebVid [8]

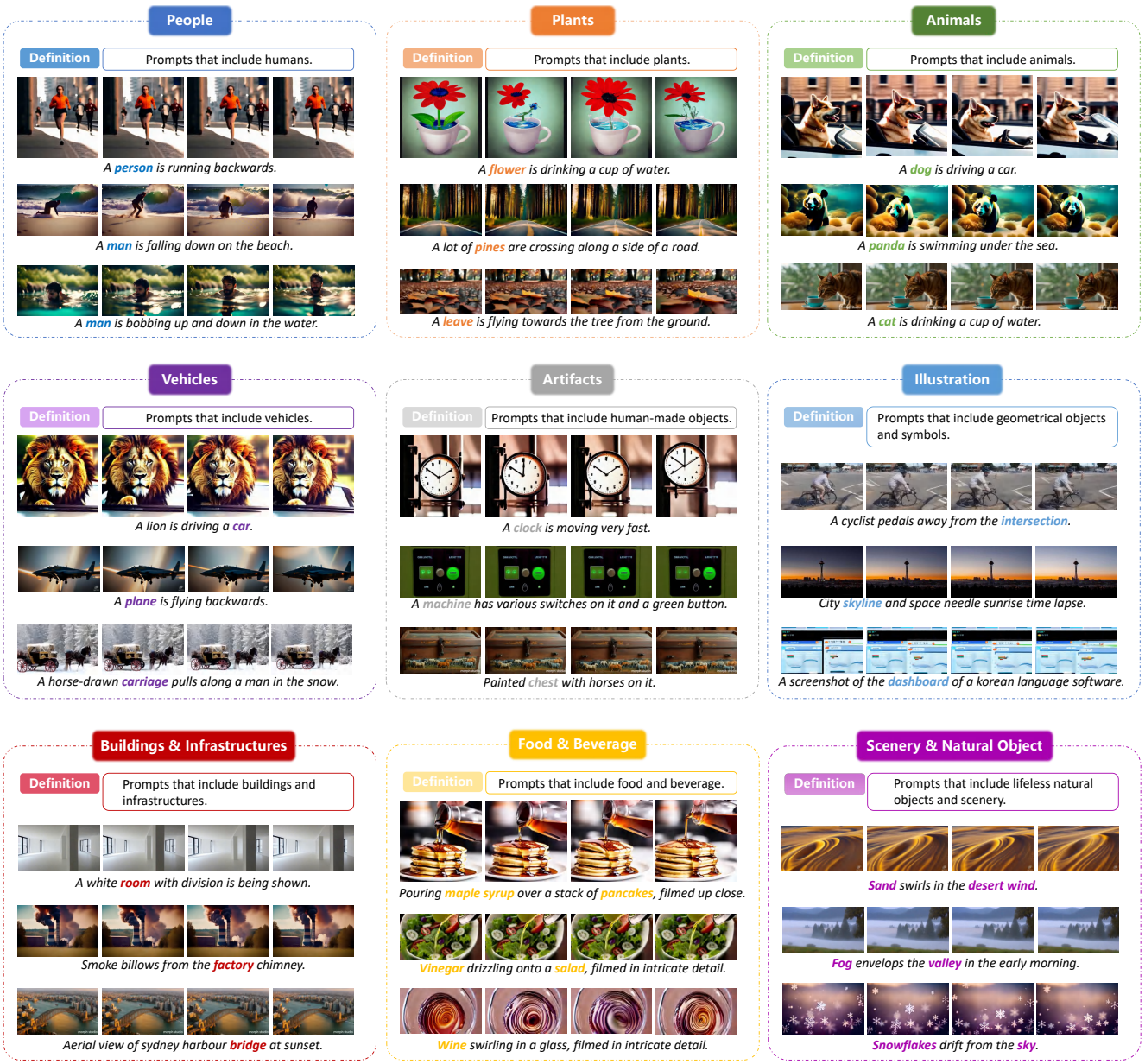


Figure 5. Descriptions and examples of the spatial major contents.

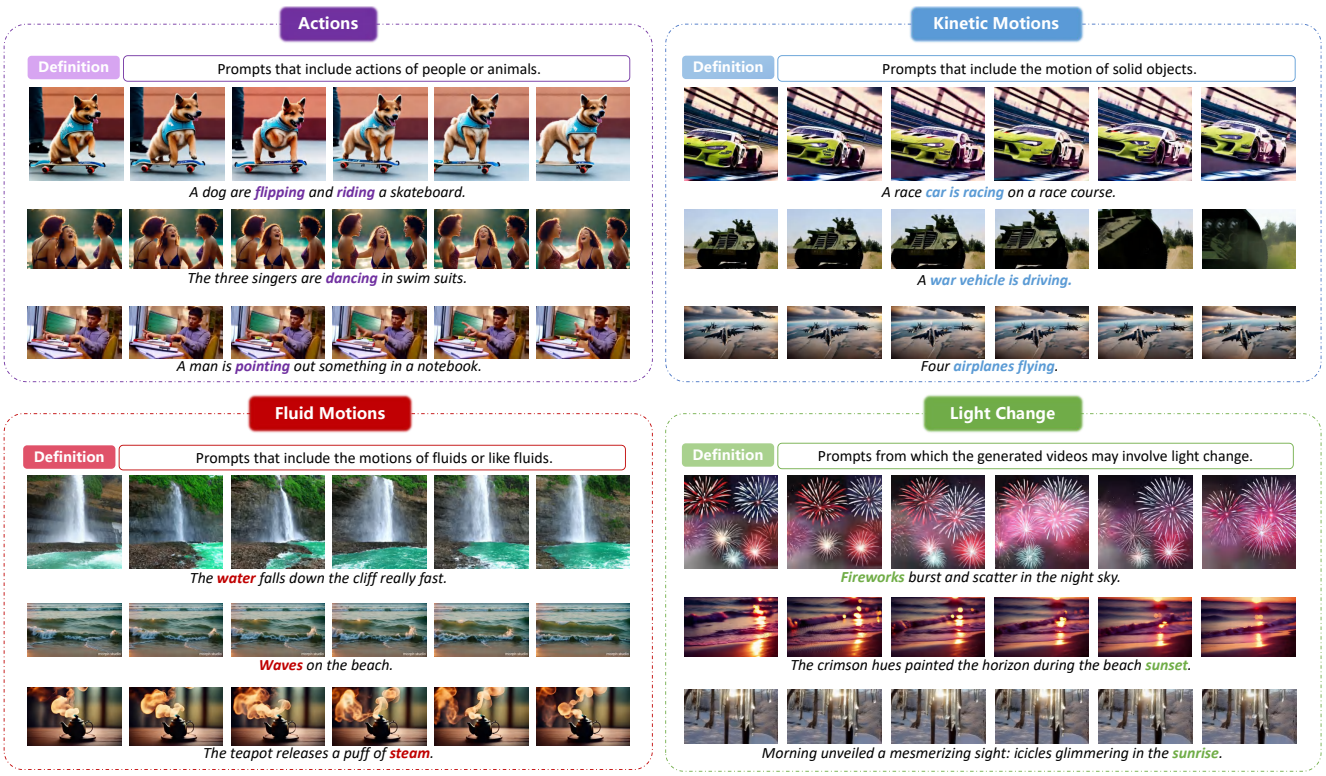


Figure 6. Descriptions and examples of the temporal major contents.

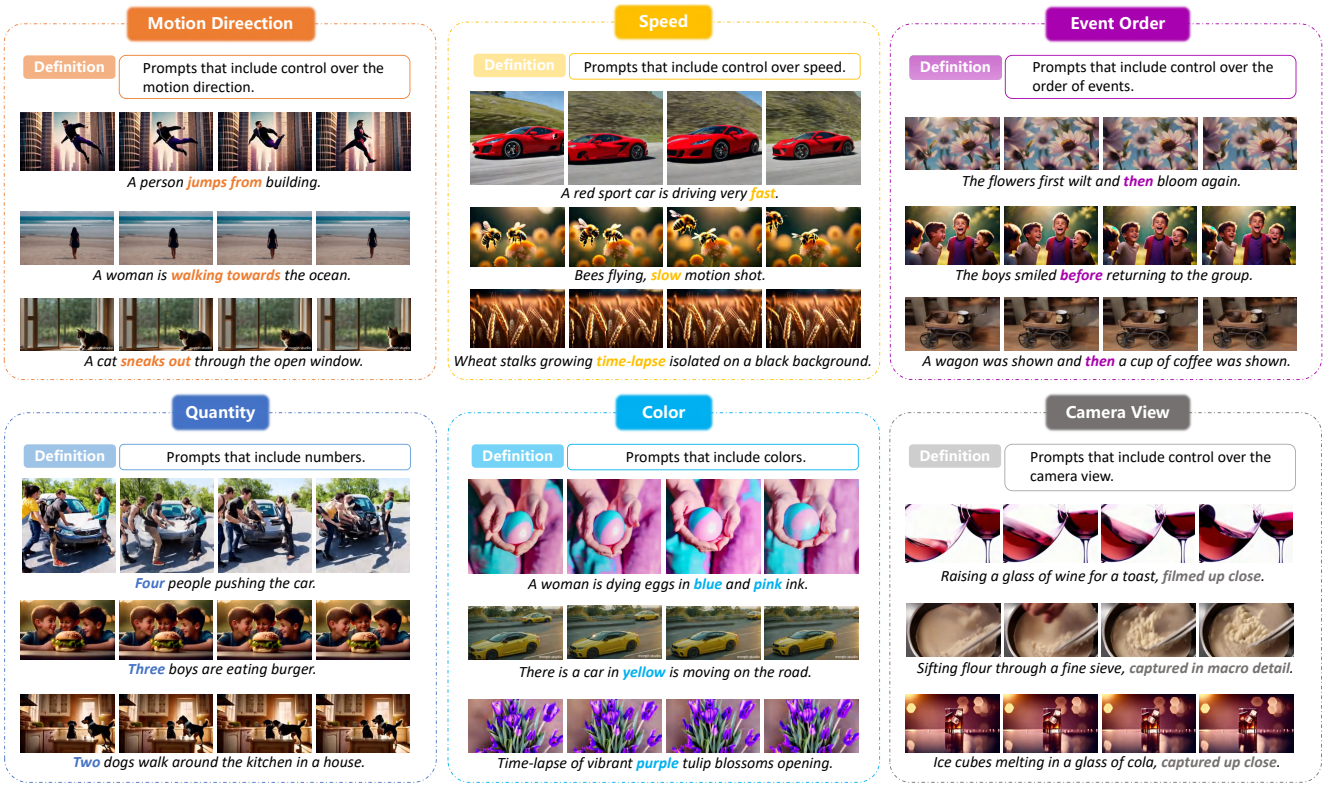
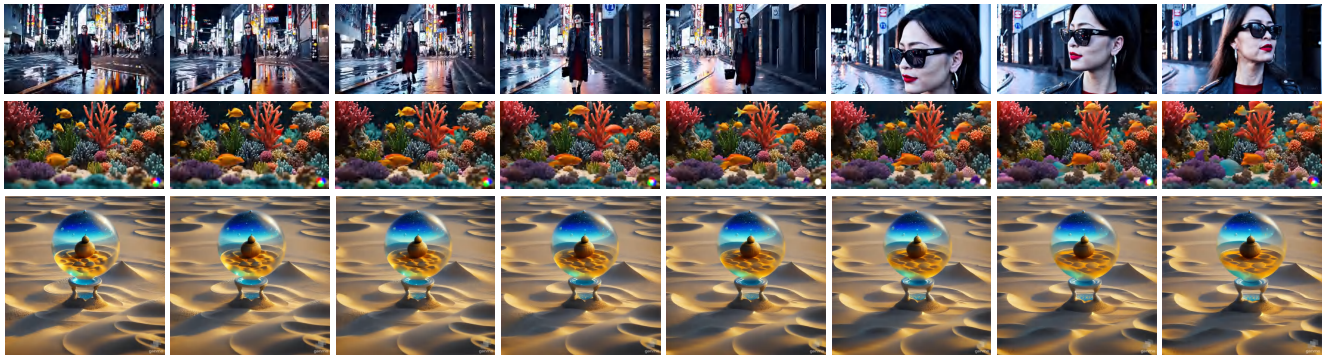


Figure 7. Descriptions and examples of the attribute control.

Static quality **4-5** (**Excellent**): The video content is exceptionally clear, and natural, with vivid and well-balanced colors. All details are flawlessly presented, resulting in a high-quality, immersive, and visually striking experience.



Static quality **3-4** (**Good**): The video content is reasonably clear and natural, with well-preserved details and fairly vibrant colors. The overall quality is satisfactory, offering a pleasant and visually appealing experience.

Static quality **2-3** (**Fair**): The video content shows slight blurriness or appears somewhat unnatural, with colors that are somewhat muted or inconsistent. The quality is acceptable but lacks sharpness, smoothness, or vibrant colors, making it less engaging or realistic.

Static quality **1-2** (**Poor**): The video content is noticeably blurry or unnatural, and the colors appear faded or washed out. While some details may still be identifiable, the lack of clarity and vibrancy results in a poor viewing experience.

Static quality **0-1** (**Bad**): The video content extremely blurry or highly unnatural, with dull or distorted colors, making it difficult to discern details or recognize objects clearly.

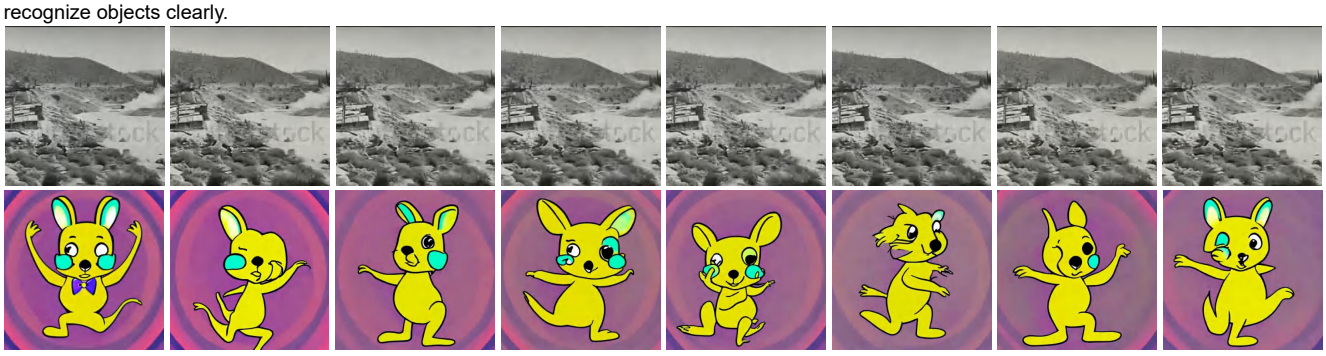
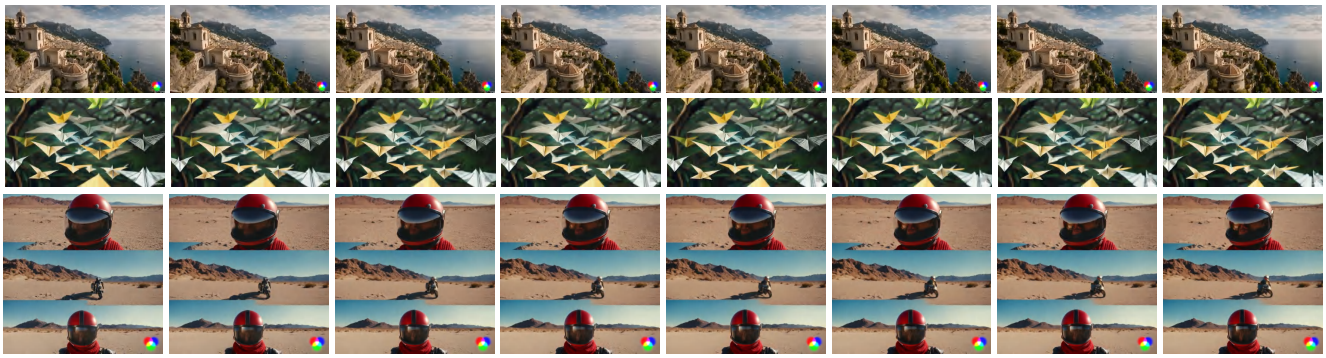


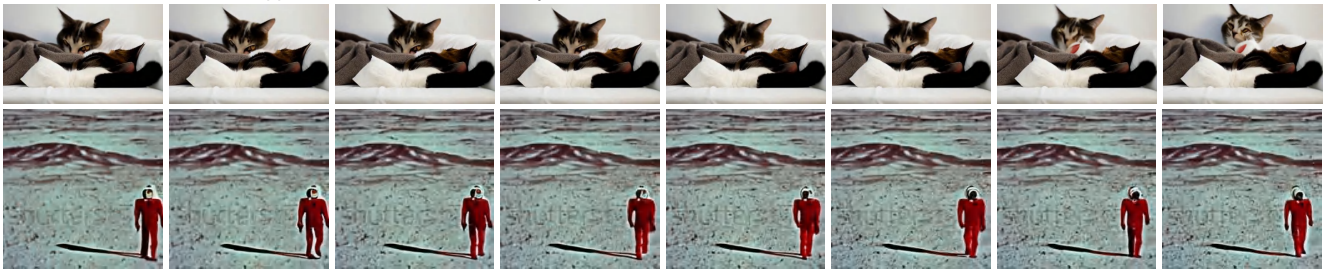
Figure 8. Instructions and examples for manual evaluation of **static quality**.

Temporal smoothness **4-5** (**Excellent**): The video exhibits perfectly smooth frame-to-frame transitions, with natural movements and no noticeable inconsistencies in objects or appearances. Object positions are fluid and realistic, creating an immersive and seamless viewing experience.

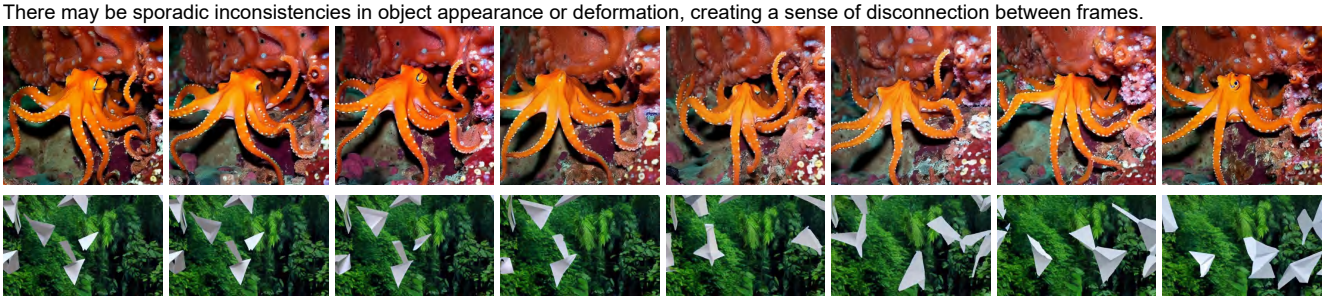


Temporal smoothness **3-4** (**Good**): Frame transitions are mostly smooth, with only rare and minor inconsistencies in object appearance or slight unnatural movements. Object positions and motions are generally well-aligned, and the video feels coherent and natural for the most part.

Temporal smoothness **2-3** (**Fair**): The frame-to-frame transitions are somewhat consistent, but minor unnatural movements, subtle object deformations, or occasional appearance inconsistencies may occur.



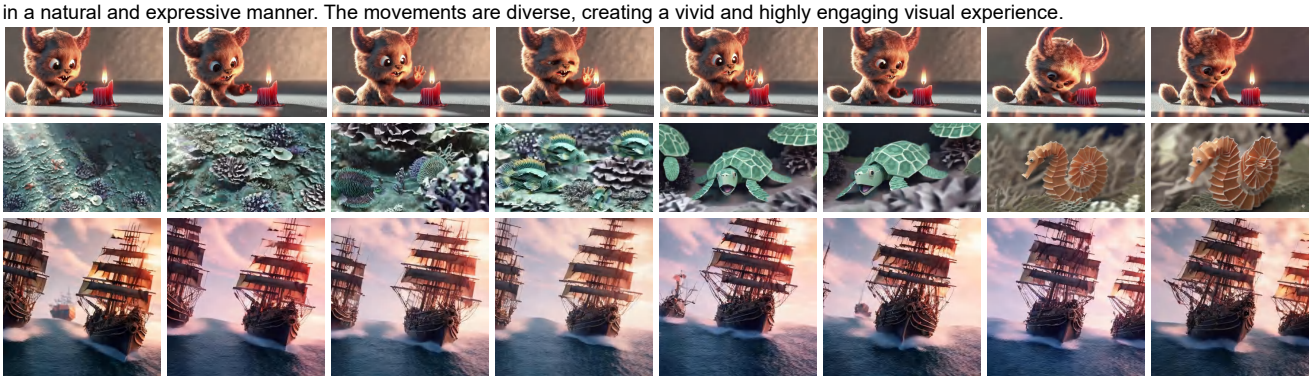
Temporal smoothness **1-2** (**Poor**): Frame transitions are notably rough, with visible unnatural movements or occasional jumps in object positions. There may be sporadic inconsistencies in object appearance or deformation, creating a sense of disconnection between frames.



Temporal smoothness **0-1** (**Bad**): The transitions between frames are highly inconsistent, with noticeable object position jumps, severe unnatural movements, or deformations. Inconsistent objects or appearances are frequent, making the video jarring and disjointed.

Figure 9. Instructions and examples for manual evaluation of **temporal smoothness**.

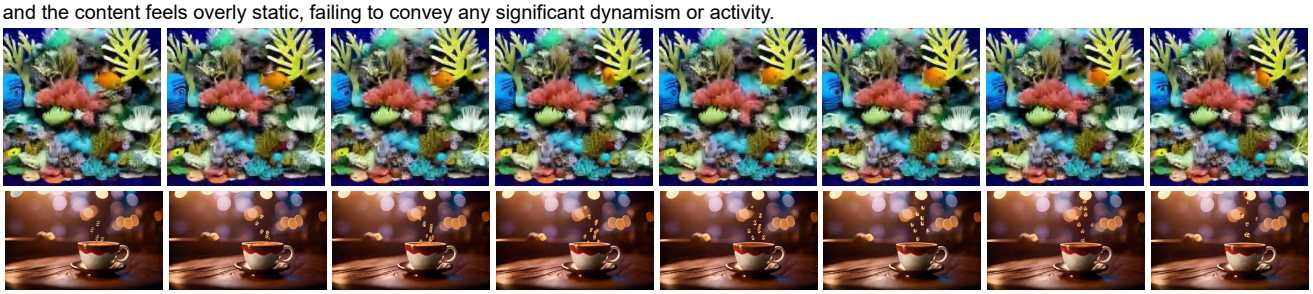
Dynamic degree **4-5** (**Excellent**): The video exhibits a highly dynamic degree of motion, with humans, animals, or objects moving across a wide range in a natural and expressive manner. The movements are diverse, creating a vivid and highly engaging visual experience.



Dynamic degree **3-4** (**Good**): The video contains a broad range of motion, with humans, animals, or objects moving naturally. The movements are engaging and visually dynamic, with clear activity that maintains viewer interest.

Dynamic degree **2-3** (**Fair**): The video displays moderate motion, with humans, animals, or objects moving across a somewhat confined range. Movements may occasionally feel limited, but they are sufficient to convey a basic sense of activity.

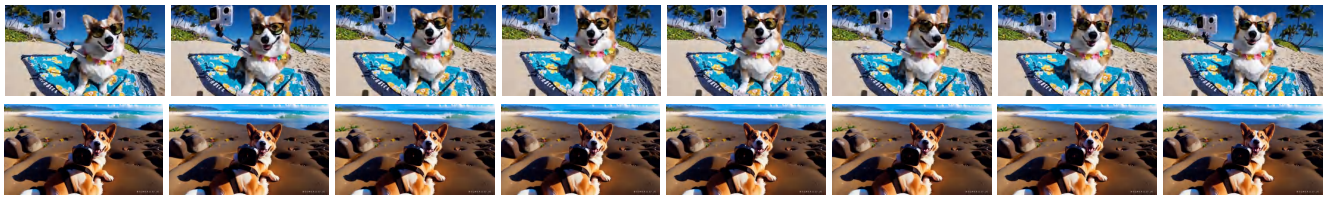
Dynamic degree **1-2** (**Poor**): The video shows limited motion, with only small, repetitive, or localized movements. The range of motion is restricted, and the content feels overly static, failing to convey any significant dynamism or activity.



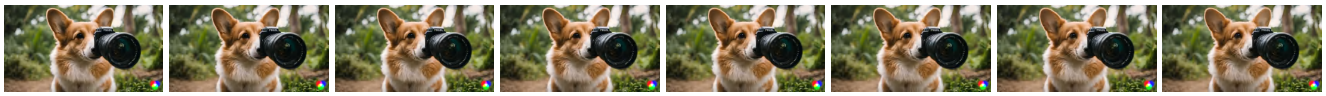
Dynamic degree **0-1** (**Bad**): The motion in the video is almost nonexistent, with minimal or no visible movement of humans, animals, or objects. The content appears static or lifeless, lacking any dynamic visual interest.

Figure 10. Instructions and examples for manual evaluation of **dynamic degree**.

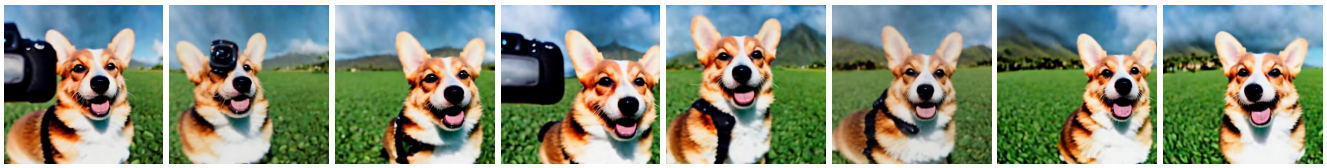
Text-video correspondence **4-5** (**Excellent**): The video content perfectly aligns with the prompt. All described elements are fully realized with high fidelity, leaving no noticeable differences or omissions. The video effectively and naturally captures the exact meaning and intent of the text.



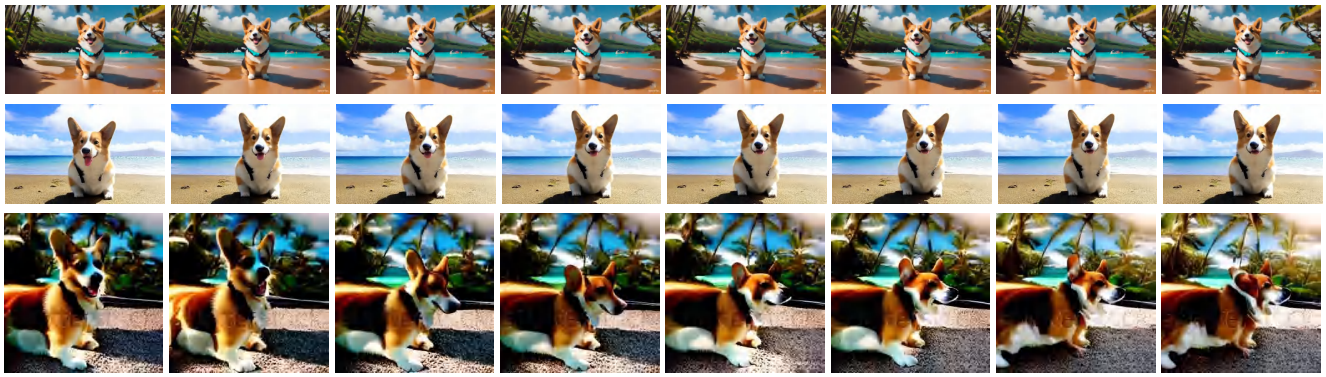
Text-video correspondence **3-4** (**Good**): The video closely matches the prompt, with most key elements represented accurately and in detail. There may be minor omissions or differences, but they do not detract significantly from the overall correspondence between the text and the video.



Text-video correspondence **2-3** (**Fair**): The video content partially aligns with the prompt. Core elements are present, but some details might be missing, incomplete, or slightly different. The overall correspondence is acceptable but lacks precision or completeness.



Text-video correspondence **1-2** (**Poor**): The video shows limited correspondence with the prompt. While some elements might loosely match, there are noticeable discrepancies, missing features, or incorrect interpretations of the text.



Text-video correspondence **0-1** (**Bad**): The video content is entirely inconsistent with the prompt. Key elements described in the prompt are either missing or incorrectly represented, , resulting in a complete lack of correspondence to the text.

Figure 11. Instructions and examples for manual evaluation of **text-video correspondence**. The example videos are of the same prompt "A corgi vlogging itself in tropical Maui."

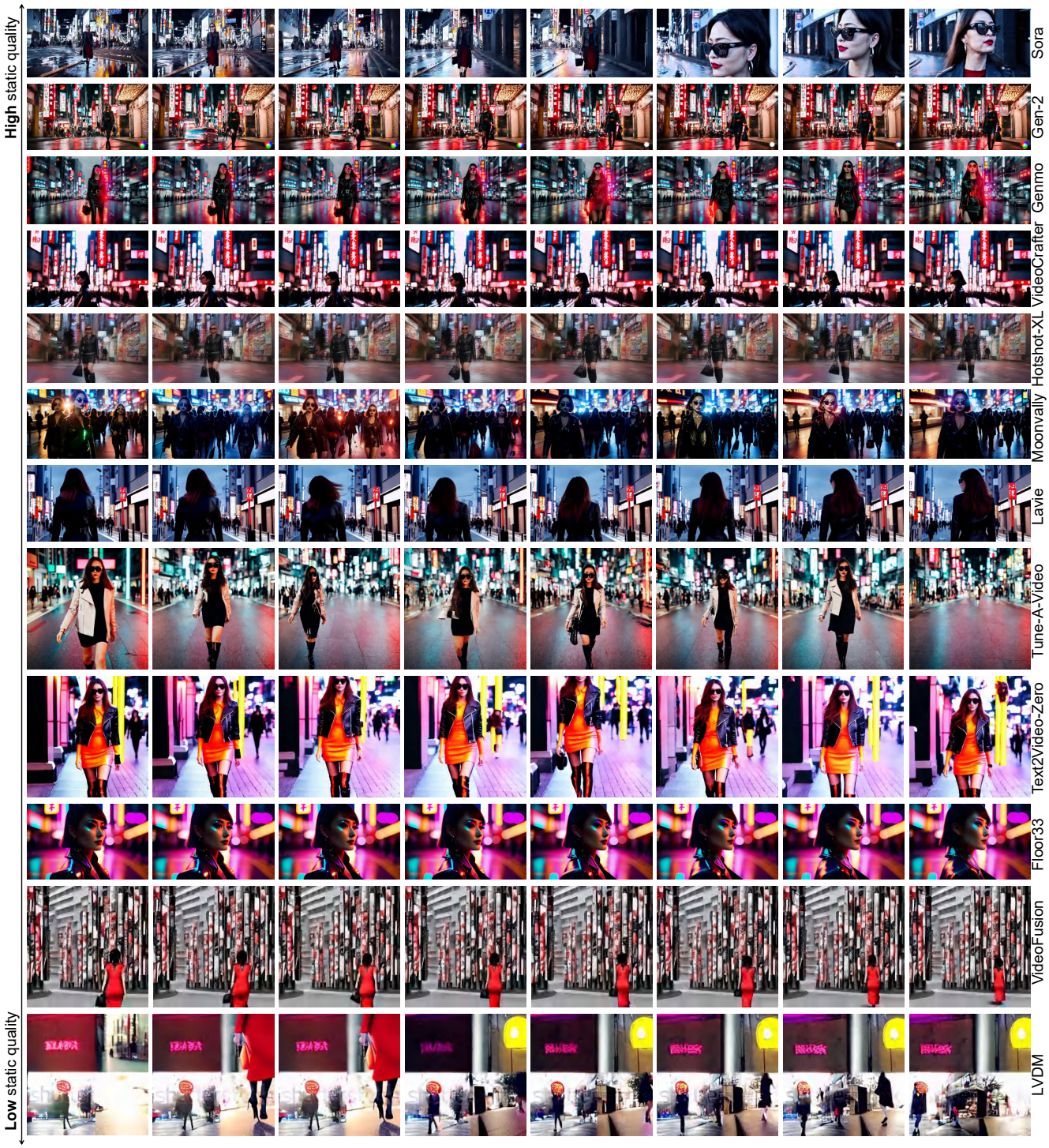


Figure 12. Visualization of generated videos in the MOS subset of AIGVQA-DB: Sort by **static quality** from highest to lowest. The video prompt is "A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about."

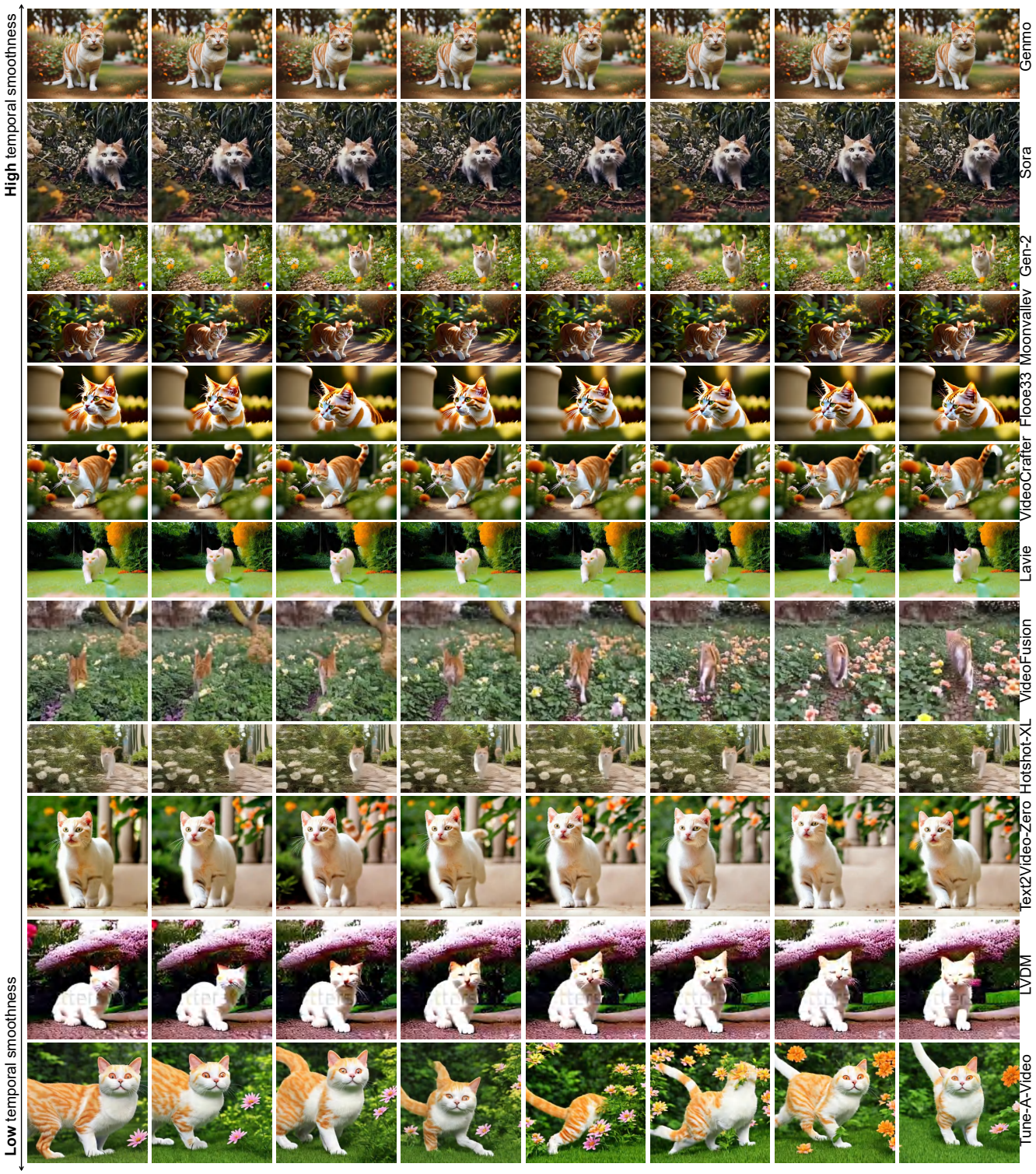


Figure 13. Visualization of generated videos in the MOS subset of AIGVQA-DB: Sort by **temporal smoothness** from highest to lowest. The video prompt is "A white and orange tabby cat is seen happily darting through a dense garden, as if chasing something. Its eyes are wide and happy as it jogs forward, scanning the branches, flowers, and leaves as it walks. The path is narrow as it makes its way between all the plants. the scene is captured from a ground-level angle, following the cat closely, giving a low and intimate perspective. The image is cinematic with warm tones and a grainy texture. The scattered daylight between the leaves and plants above creates a warm contrast, accentuating the cat's orange fur. The shot is clear and sharp, with a shallow depth of field."

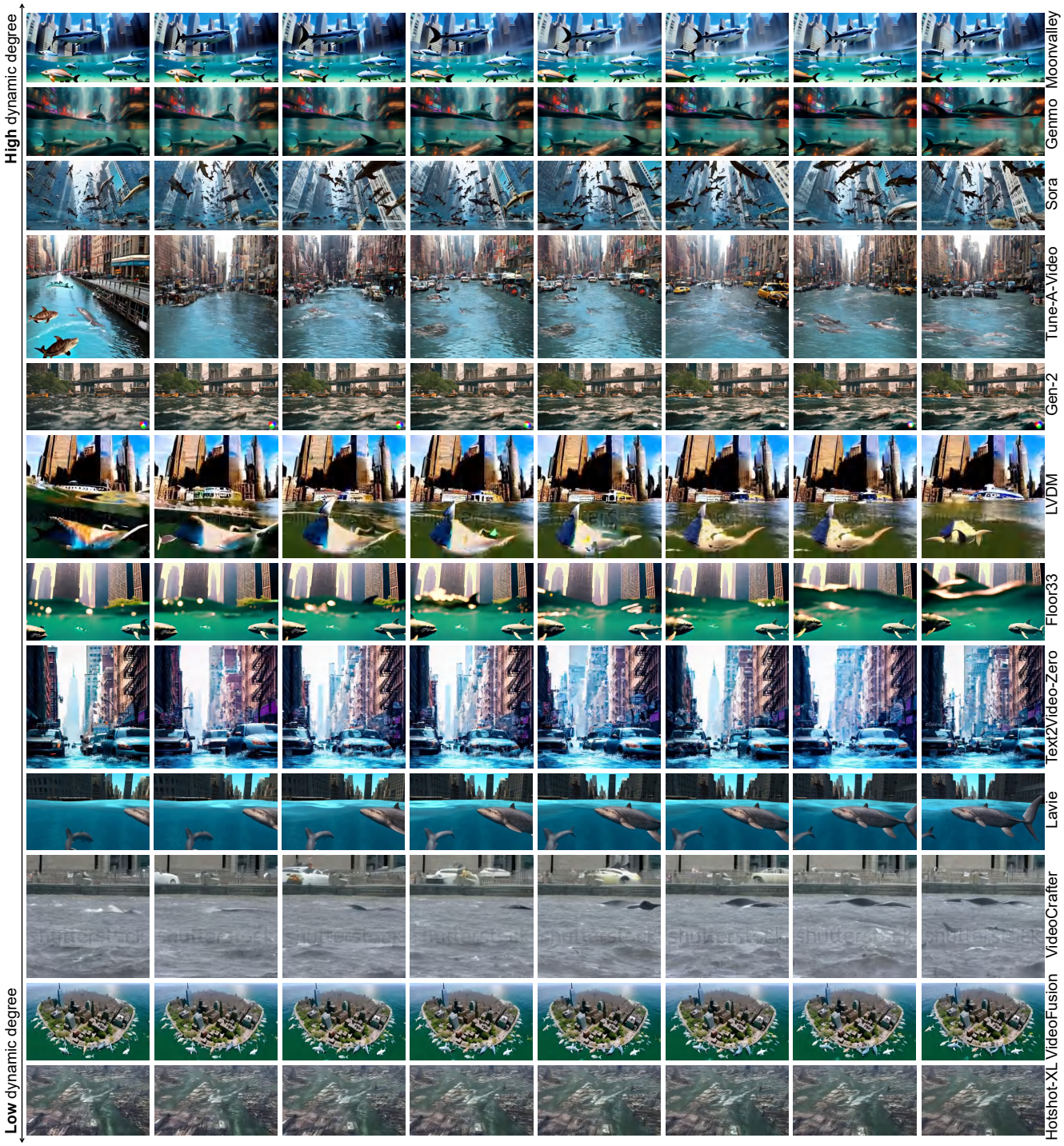


Figure 14. Visualization of generated videos in the MOS subset of AIGVQA-DB: Sort by **dynamic degree** from highest to lowest. The video prompt is "New York City submerged like Atlantis. Fish, whales, sea turtles and sharks swim through the streets of New York."

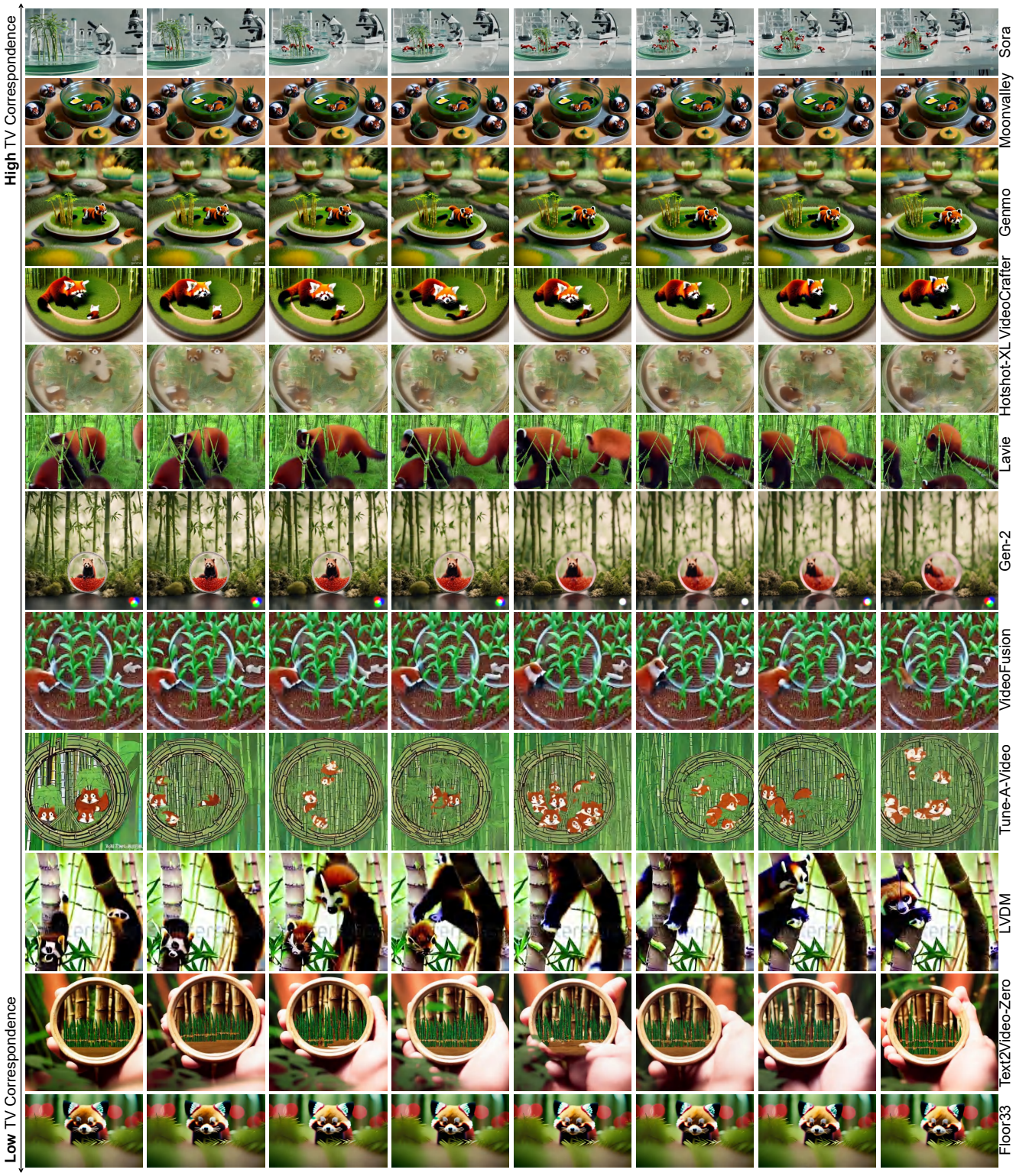


Figure 15. Visualization of generated videos in the MOS subset of AIGVQA-DB: Sort by **TV correspondence** from highest to lowest. The video prompt is "A petri dish with a bamboo forest growing within it that has tiny red pandas running around."

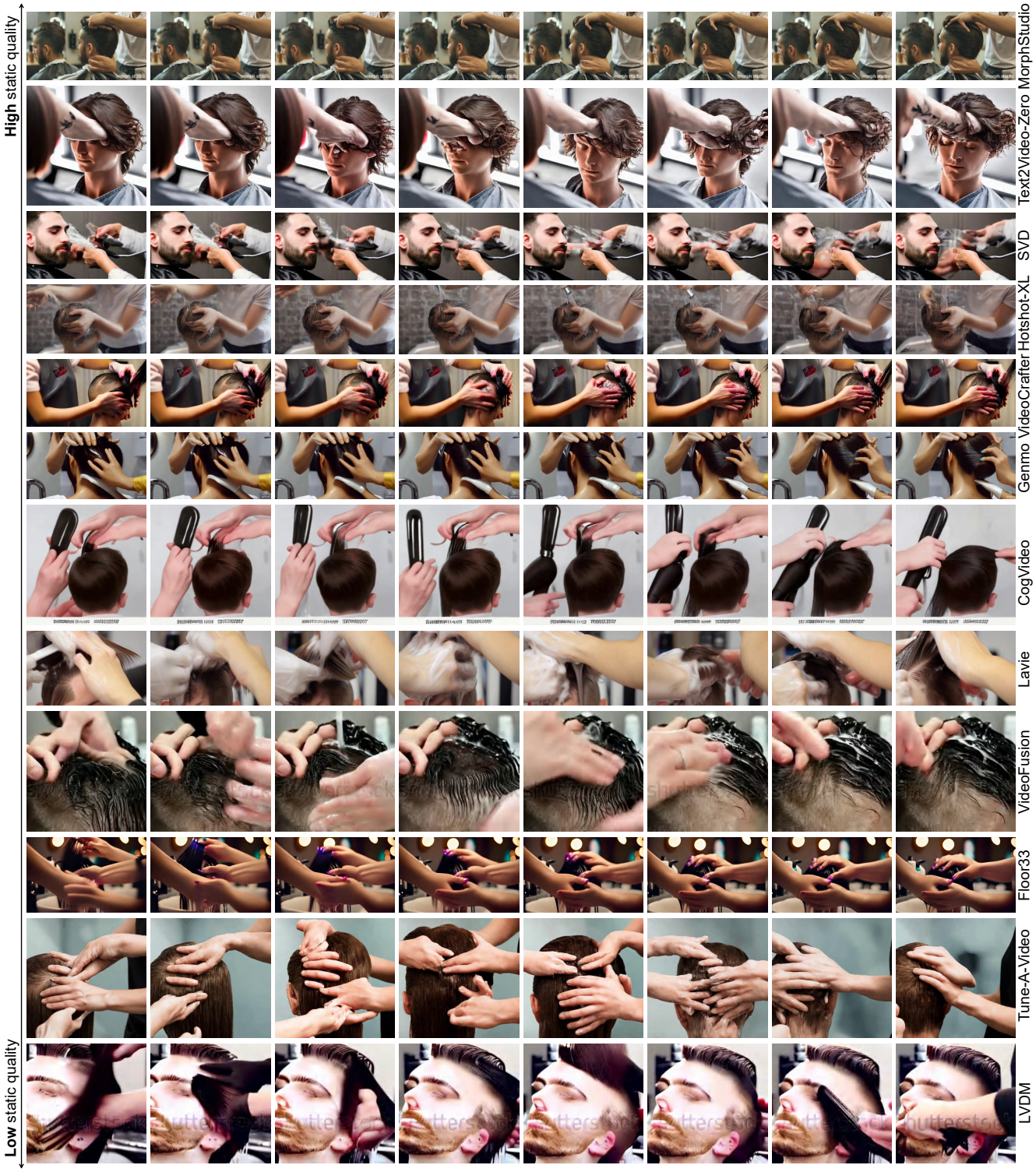


Figure 16. Visualization of generated videos in the pairs subset of AIGVQA-DB: Sort by **static quality** from highest to lowest. The video prompt is "Close up of a hairdresser's hands washing a customers hair before he is getting a haircut."

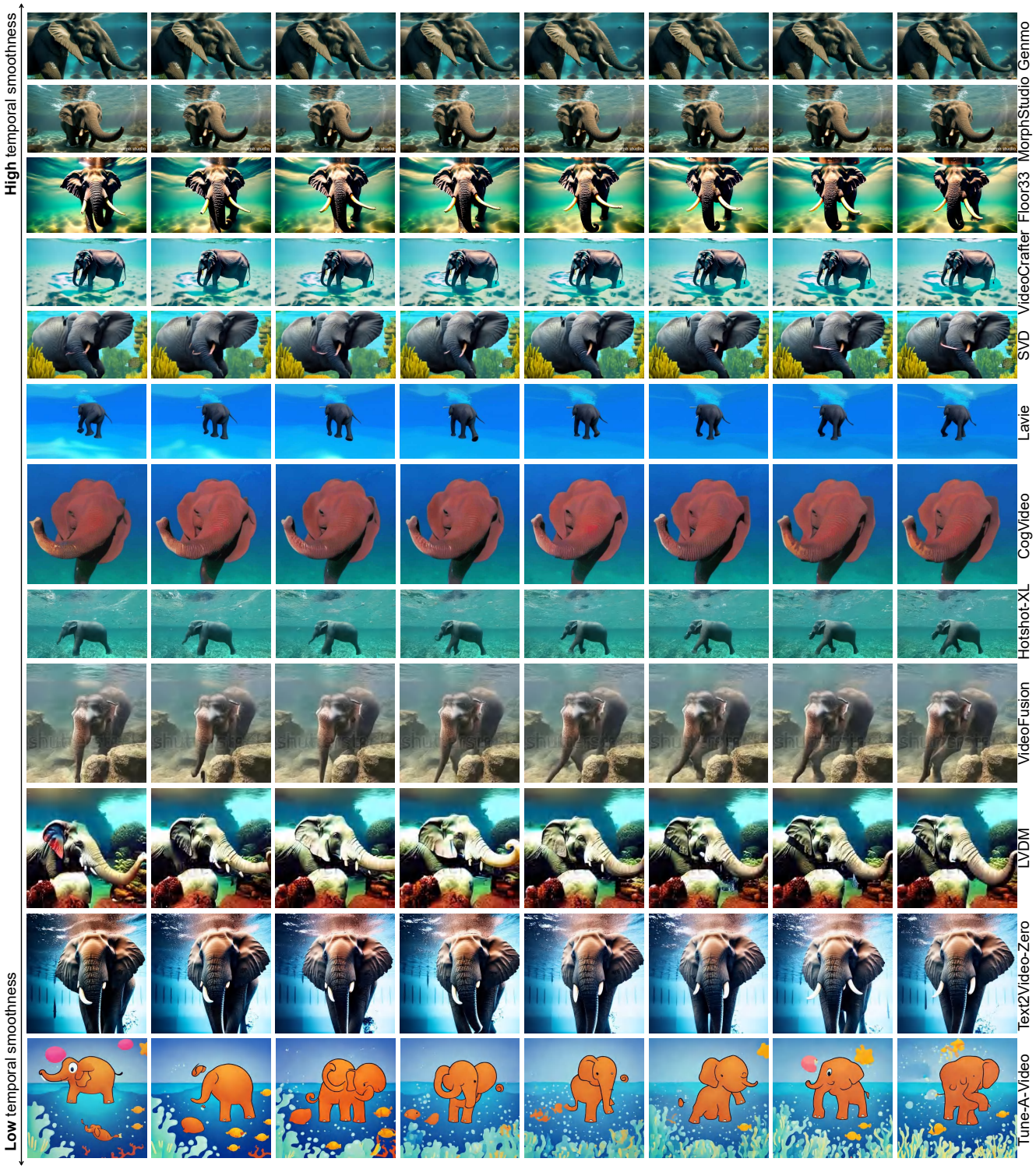


Figure 17. Visualization of generated videos in the pairs subset of AIGVQA-DB: Sort by **static quality** from highest to lowest. The video prompt is "An elephant is swimming under the sea."

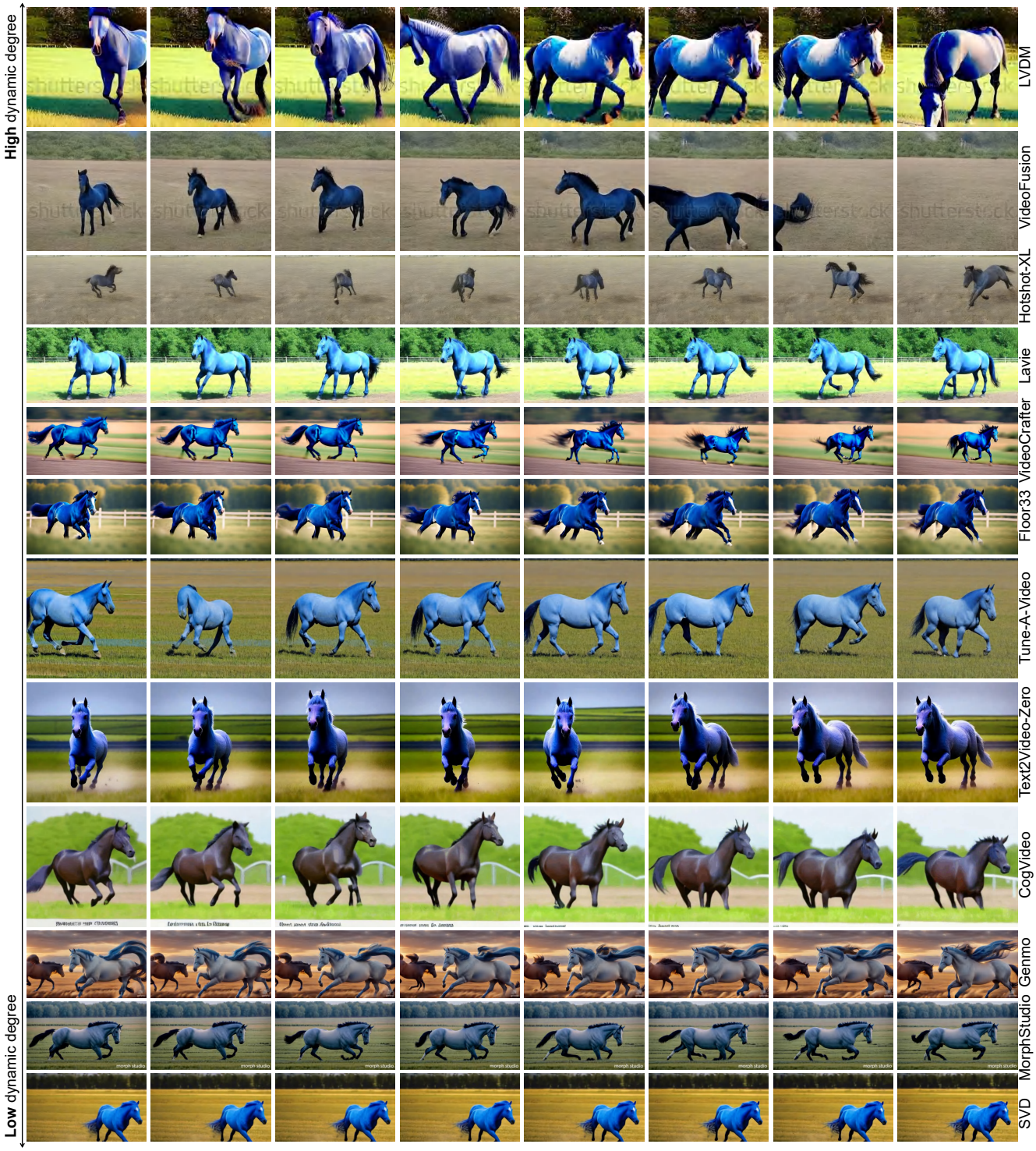


Figure 18. Visualization of generated videos in the pairs subset of AIGVQA-DB: Sort by **dynamic degree** from highest to lowest. The video prompt is "A blue horse is running in the field."

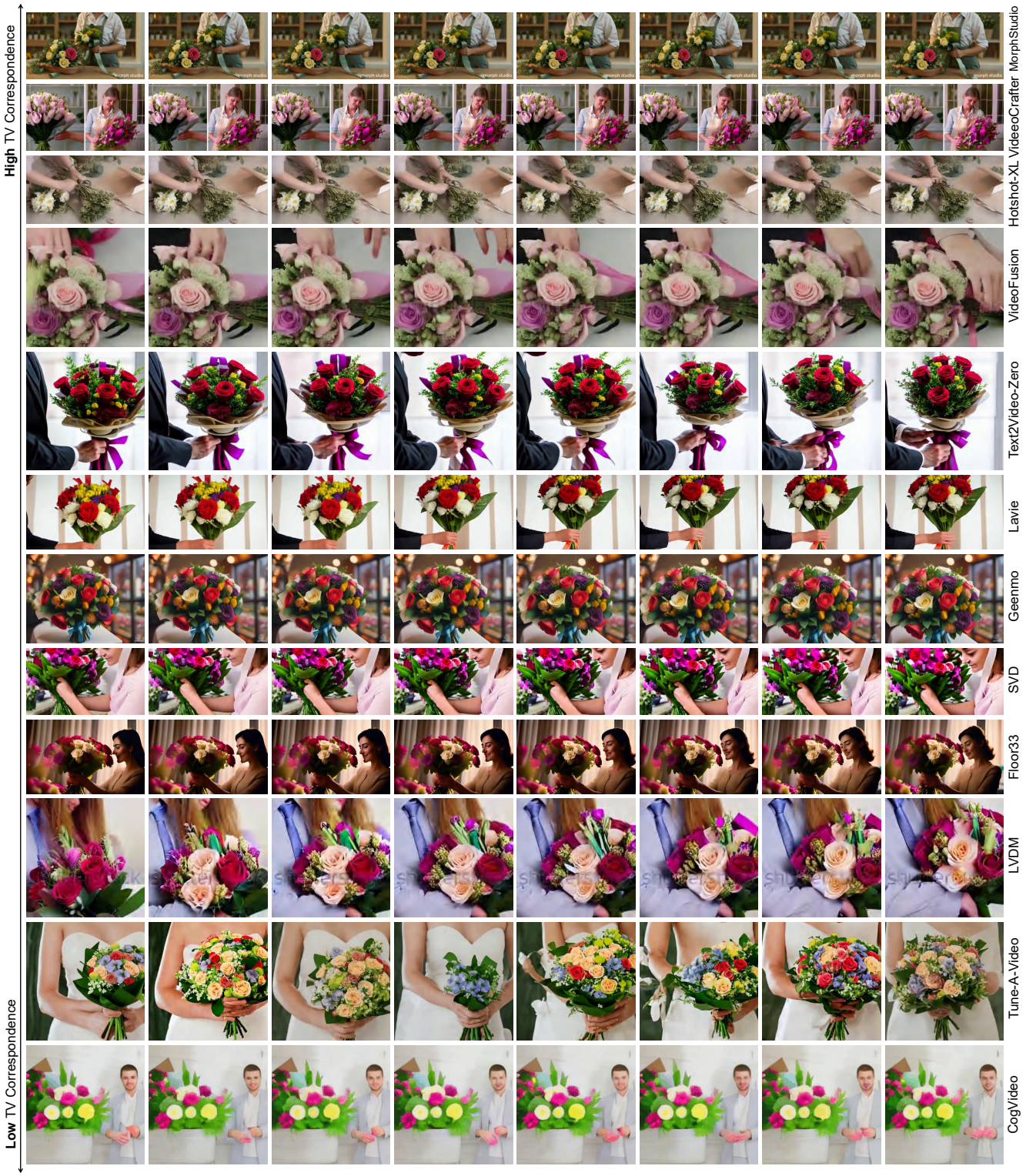


Figure 19. Visualization of generated videos in the pairs subset of AIGVQA-DB: Sort by **TV correspondence** from highest to lowest. The video prompt is "A florist finishes arranging the bouquet and then ties it with a ribbon."

References

[1] Hotshot-XL. https://github.com/hotshotco/hotshot-xl, 2023. 3, 6

[2] Floor33. https://discord.gg/EuB9KT6H, 2023. 3, 4, 6, 7

[3] Gemo. https://www.genmo.ai, 2024. 3, 6, 7

[4] Gen2. https://research.runwayml.com/gen2, 2024. 3, 4, 6

[5] Moonvalley. https://moonvalley.ai, 2024. 3, 4, 6

[6] Morph studio. https://www.morphstudio.com, 2024. 3, 4, 6, 7

[7] Sora. https://openai.com/research/video-generation-models-as-world-simulators, 2024. 2, 3, 4, 6

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew 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)*, pages 1728–1738, 2021. 2, 10

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 7

[10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3, 7

[11] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE transactions on broadcasting (TBC)*, 57(2):165–182, 2011. 8

[12] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 16890–16902, 2022. 3

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3, 6, 7

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528, 2021. 9

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 33:6840–6851, 2020. 3

[16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[17] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 9

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, 2021. 9

[19] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 3, 7

[20] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 36652–36663, 2023. 1

[21] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024. 9

[22] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei 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)*, 32(9):5944–5958, 2022. 9

[23] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*. ACM, 2019. 9

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 9

[25] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo 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)*, pages 4641–4650, 2016. 2, 10

[26] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu 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)*, 2024. 2, 9, 10

[27] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023. 3, 6, 7

[28] OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. 2

[29] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–313, 2019. 8

[30] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1817, 2018. 1

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell 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)*, pages 25278–25294, 2022. 9

[32] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, page 856–865, 2022. 9

[33] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 2023. 9

[34] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3

[35] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui 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. 2, 10

[36] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 538–554. Springer, 2022. 9

[37] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi 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. 9

[38] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 9

[39] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng 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)*, pages 7623–7633, 2023. 3, 7

[40] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 9

[41] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023. 1

[42] Jun Xu, Tao Mei, Ting Yao, and Yong 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. 2, 10

[43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 9

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 1

[45] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3096–3105, 2019. 1

[46] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 9

[47] Zhichao Zhang, Xinyue Li, Wei Sun, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Puyi Wang, Zhongpeng Ji, et al. Benchmarking aigc video quality assessment: A dataset and unified model. *arXiv preprint arXiv:2407.21408*, 2024. 9