Towards A Better Metric for Text-to-Video Generation

Jay Zhangjie Wu[1*] Guian Fang[1*] Haoning Wu[4*] Xintao Wang[3] Yixiao Ge[3]
Xiaodong Cun[3] David Junhao Zhang[1] Jia-Wei Liu[1] Yuchao Gu[1] Rui Zhao[1]
Weisi Lin[4] Wynne Hsu[2] Ying Shan[3] Mike Zheng Shou[1]

[1]Show Lab, [2]National University of Singapore [3]ARC Lab, Tencent PCG
[4]Nanyang Technological University

https://showlab.github.io/T2VScore

Abstract

*Generative models have demonstrated remarkable capability in synthesizing high-quality text, images, and videos. For video generation, contemporary text-to-video models exhibit impressive capabilities, crafting visually stunning videos. Nonetheless, evaluating such videos poses significant challenges. Current research predominantly employs automated metrics such as FVD, IS, and CLIP Score. However, these metrics provide an incomplete analysis, particularly in the temporal assessment of video content, thus rendering them unreliable indicators of true video quality. Furthermore, while user studies have the potential to reflect human perception accurately, they are hampered by their time-intensive and laborious nature, with outcomes that are often tainted by subjective bias. In this paper, we investigate the limitations inherent in existing metrics and introduce a novel evaluation pipeline, the Text-to-Video Score (T2VScore). This metric integrates two pivotal criteria: (1) Text-Video Alignment, which scrutinizes the fidelity of the video in representing the given text description, and (2) Video Quality, which evaluates the video's overall production caliber with a mixture of experts. Moreover, to evaluate the proposed metrics and facilitate future improvements on them, we present the **TVGE** dataset, collecting human judgements of 2,543 text-to-video generated videos on the two criteria. Experiments on the TVGE dataset demonstrate the superiority of the proposed T2VScore on offering a better metric for text-to-video generation. The code and dataset will be open-sourced.*

1. Introduction

Text-to-video generation marks one of the most exciting achievements in generative AI, with awesome video generative models coming out from companies [2, 3, 5, 19, 48] and opensource community [17, 50, 56, 82]. These models, equipped with the ability to learn from vast datasets of text-video pairs, can generate creative video content that can range from simple animations to complex, lifelike scenes.

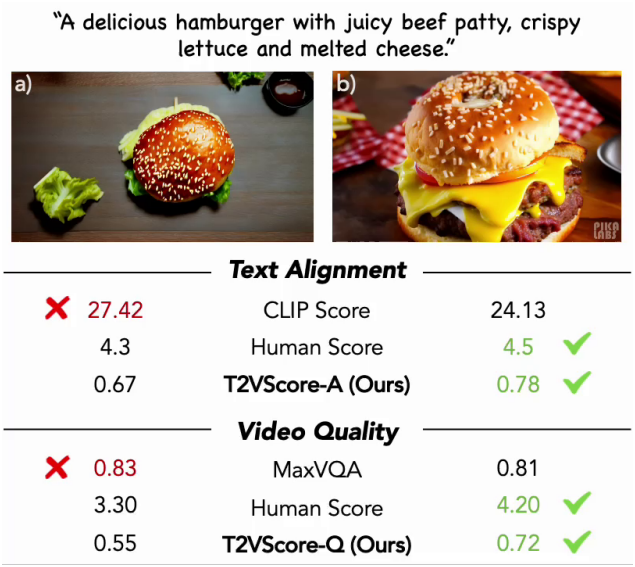


Figure 1. **T2VScore**: We measure text-conditioned generated videos from two essential perspectives: *text alignment* and *video quality*. Our proposed T2VScore achieves the highest correlation with human judgment. *We encourage readers to click and play using Adobe Acrobat.*

To assess text-conditioned generated videos, most existing studies employ objective metrics like Fréchet Video Distance (FVD) [54] and Video Inception Score (IS) [47] for *video quality*, and CLIPScore [45] for *text-video alignment*. However, these metrics have limitations. FVD and Video IS are unsuitable for open-domain video generation due to their Full-Reference nature. Meanwhile, the CLIP Score computes an average of per-frame text-image simi-

*Equal contribution.

larities using image CLIP models, overlooking important temporal motion changes in videos. This leads to a mismatch between these objective metrics and human perception, as evident in recent studies [38, 42]. Current studies also incorporate subjective user evaluations for text-to-video generation. However, conducting large-scale human evaluations is labor-intensive and, therefore, not practical for widespread, open comparisons. To address this, there is a need for fine-grained automatic metrics tailored for evaluating text-guided generated videos.

In this work, we take a significant step forward by introducing `T2VScore`, a novel automatic evaluator specifically designed for text-to-video generation. `T2VScore` assesses two essential aspects of text-guided generated videos: *text-video alignment* (*i.e.*, how well does the video match the text prompt?), and *video quality* (*i.e.*, how good is the quality of the synthesized video?). Two metrics are then introduced: 1) `T2VScore-A` evaluates the correctness of all spatial and temporal elements in the text prompt by querying the video using cutting-edge vision-language models; 2) `T2VScore-Q` is designed to predict a robust and generalizable quality score for text-guided generated videos via a combo of structural and training strategies.

To examine the reliability and robustness of the proposed metrics in the evaluation of text-guided generated videos, we present the **T**ext-to-**V**ideo **G**eneration **E**valuation (TVGE) dataset. This dataset gathers extensive human opinions on two key aspects: *text-video alignment* and *video quality*, as investigated in our `T2VScore`. The TVGE dataset will serve as an open benchmark for assessing the correlation between automatic metrics and human judgments. Moreover, it can help automatic metrics to better adapt to the domain of text-guided generated videos. Extensive experiments on the TGVE dataset demonstrate better alignment of our `T2VScore` with human judgment compared to all baseline metrics.

To summarize, we make the following contributions:

- We introduce `T2VScore` as a novel evaluator dedicated to automatically assessing text-conditioned generated videos, focusing on two key aspects: *text-video alignment* and *video quality*.
- We collect the Text-to-Video Generation Evaluation (TVGE) dataset, which is posited as the first open-source dataset dedicated to benchmarking and enhancing evaluation metrics for text-to-video generation.
- We validate the inconsistency between current objective metrics and human judgment on the TVGE dataset. Our proposed metrics, `T2VScore-A` and `T2VScore-Q`, demonstrate superior performance in correlation analysis with human evaluations, thereby serving as more effective metrics for evaluating text-conditioned generated videos.

2. Related Work

2.1. Text-to-Video Generation

Diffusion-based models have been widely explored to achieve text-to-video generation [5, 13, 16, 19, 20, 32, 56, 59, 60, 73, 80, 82, 86, 87]. VDM [20] pioneered the exploration of the diffusion model in the text-to-video generation, in which a 3D version of U-Net [46] structure is explored to jointly learn the spatial and temporal generation knowledge. Make-A-Video [48] proposed to learn temporal knowledge with only unlabeled videos. Imagen Video [19] built cascaded diffusion models to generate video and then spatially and temporally up-sample it in cascade. PYoCo [13] introduced the progressive noise prior model to preserve the temporal correlation and achieved better performance in fine-tuning the pre-trained text-to-image models to text-to-video generation. The subsequent works, LVDM [16] et al., further explored training a 3D U-Net in latent space to reduce training complexity and computational costs. These works can be classified respectively as pixel-based models and latent-based models. Show-1 [82] marks the first integration of pixel-based and latent-based models for video generation. It leverages pixel-based models for generating low-resolution videos and employs latent-based models to upscale them to high resolution, combining the advantages of high efficiency from latent-based models and superior content quality from pixel-based models. Recently, the text-to-video generation products, such as Gen-2 [2], Pika [3], and Floor33 [1], and the open-sourced foundational text-to-video diffusion models, such as ModelScopeT2V [57], ZeroScope [50], VideoCrafter [17], have democratized the video generation, garnering widespread interest from both the community and academia.

2.2. Evaluation Metrics

Image Metrics. Image-level metrics are widely utilized to evaluate the frame quality of generated videos. These include Peak Signal-to-Noise Ratio (PSNR) [63], and Structural Similarity Index (SSIM) [62], Learned Perceptual Image Patch Similarity (LPIPS) [84], Fréchet Inception Distance (FID) [44], and CLIP Score [45]. Among them, the PSNR [63], SSIM [62], and LPIPS [84] are mainly employed to evaluate the quality of reconstructed video frames by comparing the difference between generated frames and original frames. Specifically, PSNR [63] is the ratio between the peak signal and the Mean Squared Error (MSE) [61]. SSIM [62] evaluates brightness, contrast, and structural features between generated and original images. LPIPS [84] is a perceptual metric that computes the distance of image patches in the latent feature space. FID [44] utlizes the InceptionV3 [51] to extract feature maps from normalized generated and real-world frames, and computes the mean and covariance matrices for

FID [44] scores. CLIP Score [45] measures the similarity of the CLIP features extracted from the images and texts, and it has been widely employed in text-to-video generation or editing tasks [5, 15, 35, 48, 73, 82, 86].

Video Metrics. In contrast to the frame-wise metrics, video metrics focus more on the comprehensive evaluation of the video quality. Fréchet Video Distance (FVD) [54] utilizes the Inflated-3D Convnets (I3D) [7] pre-trained on Kinetics [8] to extract the features from videos, and compute their means and covariance matrices for FVD scores. Differently, Kernel Video Distance (KVD) [53] computes the Maximum Mean Discrepancy (MMD) [14] of the video features extracted using I3D [7] to evaluate the video quality. Video Inception Score (Video IS) [47] computes the inception score of videos with the features extracted from C3D [52]. Frame Consistency CLIP Score [45] calculates the cosine similarity of the CLIP image embeddings for all pairs of video frames to measure the consistency of edited videos [15, 35, 73–75, 85].

2.3. Video Quality Assessment

State-of-the-arts on video quality assessment (VQA) have been predominated by learning-based approaches [24, 27, 67]. Typically, these approaches leverage pre-trained deep neural networks as feature extractors and use human opinions as supervision to regress these features into quality scores. Some most recent works [65, 66, 70] have adopted a new strategy that uses a large VQA database on natural videos [81] to learn better feature representations for VQA, and then transfer to diverse types of videos with only a few labeled videos available. This strategy has been validated as an effective way to improve the prediction accuracy and robustness on relatively small VQA datasets for enhanced videos [36] and computer-generated contents [79]. In our study, we extend this strategy for evaluating text-conditioned generated videos, bringing a more reliable and generalizable quality metric for text-to-video generation.

Despite leveraging from large video quality databases, several recent works [55, 68, 69, 72] have also explored to adopt multi-modality foundation models *e.g.* CLIP [45] for VQA. With the text prompts as natural quality indicators (*e.g. good/bad*), these text-prompted methods prove superior abilities on zero-shot or few-shot VQA settings, and robust generalization among distributions. Inspired by existing studies, the proposed quality metric in T2VScore also ensembles a text-prompted structure, which is proved to better align with human judgments on videos generated by novel generators that are not seen during training.

2.4. QA-based Evaluation

Recent studies have emerged around the idea of using Visual Question Answering (VQA) to test the accuracy of advanced AI models. TIFA [22] utilizes GPT-3 [6] to create questions in various areas, such as color and shape, and checks the answers with VQA systems like mPLUG [26]. VQ^{2}A[9] makes VQA more reliable by synthesizing new data and employing high-quality negative sampling. VPEval[10] improves this process through the use of object detection and Optical Character Recognition (OCR), combining these with ChatGPT for more controlled testing. However, these methods have not yet explored videos, where both spatial and temporal elements should be evaluated. We are adding specific designs to the temporal domain to improve VQA for video understanding. This provides a more comprehensive method for evaluating text-video alignment from both space and time.

3. Proposed Metrics

We introduce two metrics to evaluate text-guided generated videos, focusing on two essential dimensions: Text Alignment (Sec. 3.1) and Video Quality (Sec. 3.2).

3.1. Text Alignment

State-of-the-art multimodal large language models (MLLMs) have demonstrated human-level capabilities in both visual and textual comprehension and generation. Here, we introduce a framework for assessing the text-video alignment using these MLLMs. An overview of our text alignment evaluation process is presented in Fig. 2.

Entity Decomposition in Text Prompt. Consider a text prompt denoted as $\mathcal{P}$. Our initial step involves parsing $\mathcal{P}$ into distinct semantic elements, represented as e_i. We then identify the hierarchical semantic relationships among these elements, forming entity tuples $\{(e_i, e_j)\}$. Here, e_j is semantically dependent on e_i to form a coherent meaning. For instance, the tuple (dog, a) implies that the article "a" is associated with the noun "dog", while $(\text{cat}, \text{playing soccer})$ suggests that the action "playing soccer" is attributed to the "cat". Elements that exert a global influence over the entire prompt, like *style* or *camera motion*, are categorized under a `global` element. This structuring not only clarifies the interconnections within the elements of a text prompt but also implicitly prioritizes them based on their hierarchical significance. For instance, mismatching an element that holds a higher dependency rank would result in a more substantial penalty on the final text alignment score.

Question/Answer Generation with LLMs. Our main goal is to generate diverse questions that cover all elements of the text input evenly. Drawing inspiration from previous studies [22], for a text prompt $\mathcal{P}$, we utilize large language models (LLMs) to generate question-choice-answer tuples $\{Q_i, C_i, A_i\}_{i=1}^{N}$, as depicted on the top of Fig. 2. Different from prior work focusing on text-image alignment, we emphasize the temporal aspects, such as object trajectory and

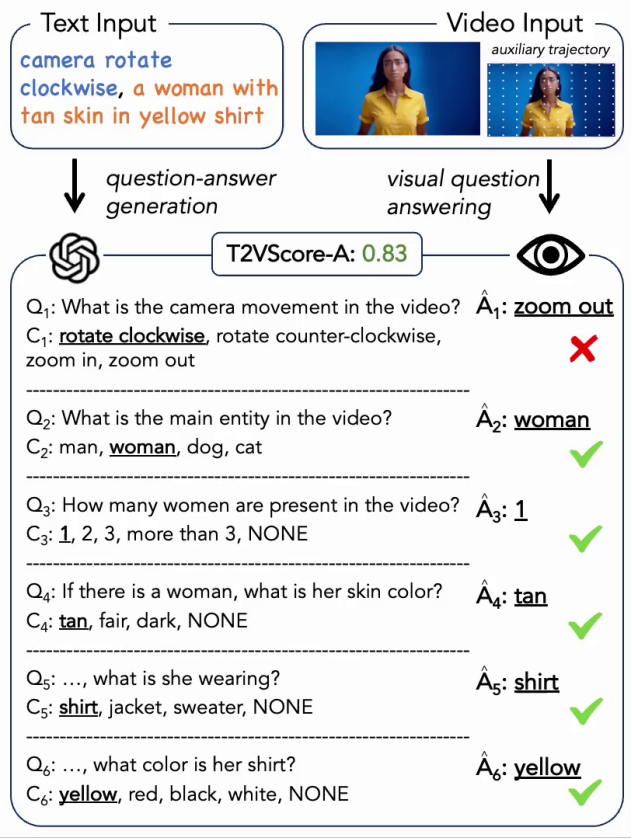


Figure 2. **Pipeline for Calculating `T2VScore-A`:** We input the text pompt into large language models (LLMs) to generate questions and answers. Utilizing CoTracker [23], we extract the auxiliary trajectory, which, along with the input video, is fed into multimodal LLM (MLLMs) for visual question answering (VQA). The final `T2VScore-A` is measured based on the accuracy of VQA. *Please click and play using Adobe Acrobat.*

camera motion, which are unique and essential for evaluating text alignment in dynamic video contexts. We employ a single-pass inference using in-context learning with GPT-3.5 [6, 64] to generate both questions and answers. We manually curate 3 examples and use them as in-context examples for GPT-3.5 to follow. The complete prompt used for generating question and answer pairs can be found in the supplementary.

Video Question Answering with Auxiliary Trajectory. Most open-domain vision-language models are image-centric [11, 30, 33, 34, 77, 89], with only a few focusing on video [31, 39, 78]. These VideoLLMs often struggle with fine-grained temporal comprehension, as evidenced by their performance on benchmarks like SEED-Bench [25]. To address this, we introduce the use of auxiliary trajectories, generated by off-the-shelf point tracking models (e.g., CoTracker [23] and OmniMotion [58]), to enhance the understanding of object and camera movements. We process

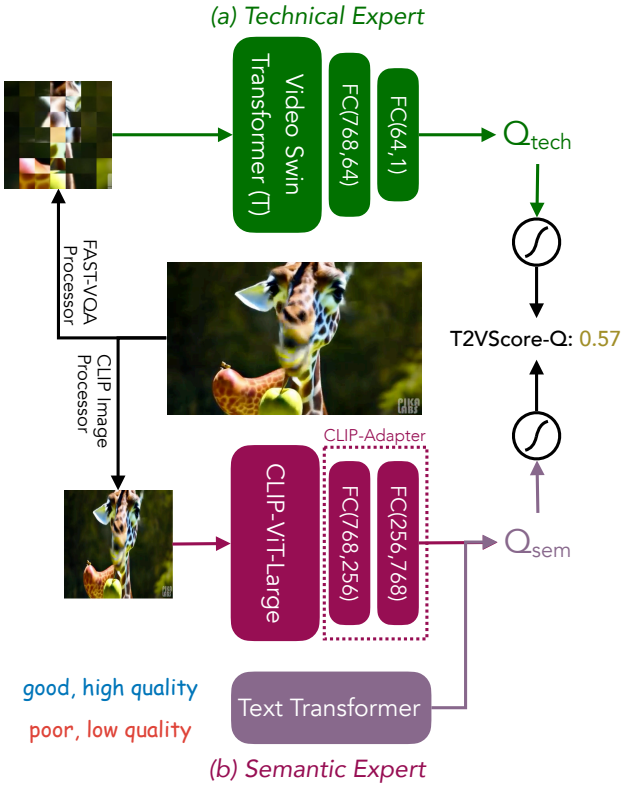


Figure 3. **Pipeline for Calculating `T2VScore-Q`:** a mixture of a ***technical expert*** (a) to capture *spatial* and *temporal* technical distortions, and a text-prompted ***semantic expert*** (b).

a video $\mathcal{V}$, created by T2V models using text prompt $\mathcal{T}$, alongside its tracking trajectory $\mathcal{V}_{\text{track}}$ and question-choice pairs $\{Q_i, C_i\}_{i=1}^{N}$ generated by LLMs. These inputs are then fed into multi-modality LLMs for question answering: $\hat{\text{A}}_\text{i} = \text{VQA}(\mathcal{V}, \mathcal{V}_{\text{track}}, Q_i, C_i)$.

We define the Text-to-Video (T2V) alignment score `T2VScore-A` as the accuracy of the video question answering process:

$$\texttt{T2VScore-A}(\mathcal{T}, \mathcal{V}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{A}i = A_i]. \quad (1)$$

The `T2VScore-A` ranges from 0 to 1, with higher values indicating better alignment between text $\mathcal{T}$ and video $\mathcal{V}$.

3.2. Video Quality

In this section, we discuss the proposed video quality metric in the T2V Score. Our core principle is simple: it should be able to ***keep effective*** to evaluate videos from *unseen* generation models that come up after we propose this score. Under this principle, the proposed metric aims to achieve two important goals: **(G1)** It can more accurately assess the quality of generated videos without seeing any of them (**zero-shot**); **(G2)** while adapted to videos gener-

ated on known models, it can significantly improve generalized performance on **unknown models**. Both aims inspire us to drastically improve the generalization ability of the metric, via a combo of Mix-of-Limited-Expert Structure (Sec. 3.2.1), Progressive Optimization Strategy (Sec. 3.2.2), and List-wise Learning Objectives (Sec. 3.2.3), elaborated as follows.

3.2.1 Mix-of-Limited-Expert Structure

Given the *hard-to-explain* nature of quality assessment [84], current VQA methods that only learn from human opinions in video quality databases will more or less come with their own biases, leading to poor generalization ability [29]. Considering our goals, inspired by existing practices [24, 70, 71], we select two evaluators with different biases as *limited experts*, and fuse their judgments to improve generalization capacity of the final prediction. Primarily, we include a ***technical expert*** (Fig. 3(a)), aiming at capturing distortion-level quality. This branch adopts the structure of FAST-VQA [65], which is pre-trained from the largest VQA database, LSVQ [81], and further fine-tuned on the MaxWell [71] database that contains a wide range of *spatial* and *temporal* distortions. While the technical branch can already cover scenarios related to naturally-captured videos, generated videos are more likely to include *semantic degradations*, *i.e.* failing to generate correct structures or components of an object. Thus, we include an additional text-prompted ***semantic expert*** (Fig. 3(b)). It is based on MetaCLIP [76], and calculated via a confidence score on the binary classification between the positive prompt *good, high quality* and negative prompt *poor, low quality*. We also add an additional adapter [12] to better suit the CLIP-based evaluator in the domain of video quality assessment.

Denote the technical score for the video $\mathcal{V}$ as $\mathtt{Q}_{\mathtt{tech}}(\mathcal{V})$, the text-prompted semantic score as $\mathtt{Q}_{\mathtt{sem}}(\mathcal{V})$, we fuse the two independently-optimized judgements via ITU-standard perceptual-oriented remapping [4], into the `T2VScore-Q`:

$$\mathrm{R}(s) = \frac{1}{1+e^{-\frac{s-\mu(s)}{\sigma(s)}}} \tag{2}$$

$$\mathtt{T2VScore\text{-}Q}(\mathcal{V}) = \frac{\mathrm{R}(\mathtt{Q}_{\mathtt{tech}}) + \mathrm{R}(\mathtt{Q}_{\mathtt{sem}})}{2} \tag{3}$$

The `T2VScore-Q` ranges from 0 to 1, with higher values indicating better visual quality of video $\mathcal{V}$.

3.2.2 Progressive Optimization Strategy

After introducing the structure, we discuss the optimization strategy for the `T2VScore-Q` (Fig. 4). In general, the training is conducted in three stages: *pre-training*, *fine-tuning*, and *adaptation*. The stages come with gradually

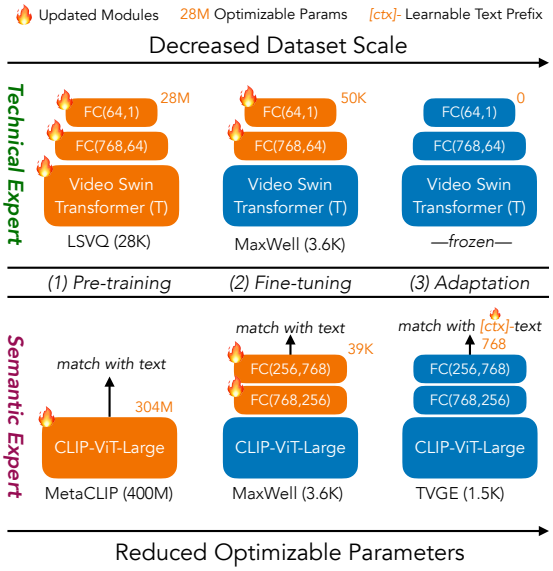


Figure 4. **Optimization Strategy of the Video Quality Metric,** via gradually decreased scales of training datasets, and correspondingly progressively reduced optimizable parameters.

smaller datasets, coped with **progressively reduced optimizable parameters**. For $\mathtt{Q}_{\mathtt{tech}}$, the optimization strategies for each stage are listed as follows: (1) *end-to-end* pre-training with large-scale $\text{LSVQ}_{\text{train}}$ dataset (*28K* videos); (2) *multi-layer* fine-tuning with medium-scale $\text{MaxWell}_{\text{train}}$ dataset (*3.6K* videos); (3) Given that specific distortions on generated videos (See Fig. 5) are usually associated with semantics, to avoid over-fitting, the technical expert is kept *frozen* during the adaptation stage. For $\mathtt{Q}_{\mathtt{sem}}$, (1) we directly adopt official weights from MetaCLIP [76] as pre-training; (2) for the fine-tuning stage, we train a lightweight adapter [12] on $\text{MaxWell}_{\text{train}}$; (3) for adaptation, we train an additional prefix token [55, 69, 88] to robustly adapt it to the domain of generated videos.

3.2.3 List-wise Learning Objectives

Plenty of existing studies [24, 29, 65, 66] have pointed out that compared with independent scores, the rank relations among different scores are more reliable and generalizable, especially for small-scale VQA datasets [49]. Given these insights, we decide to adopt the *list-wise* learning objectives [28] combined by rank loss ($\mathcal{L}_{rank}$) and linear loss ($\mathcal{L}_{linear}$) as our training objective for both limited experts:

$$\mathcal{L}_{rank} = \sum_{i,j} \max((s^i_{pred} - s^j_{pred})\ \mathrm{sgn}\ (s^j_{gt} - s^i_{gt}), 0) \tag{4}$$

$$\mathcal{L}_{lin} = (1 - \frac{< s_{pred} - \overline{s_{pred}}, s_{gt} - \overline{s_{gt}} >}{\|s_{pred} - \overline{s_{pred}}\|_2 \|s_{gt} - \overline{s_{gt}}\|_2})/2 \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{lin} + \lambda \mathcal{L}_{rank} \tag{6}$$

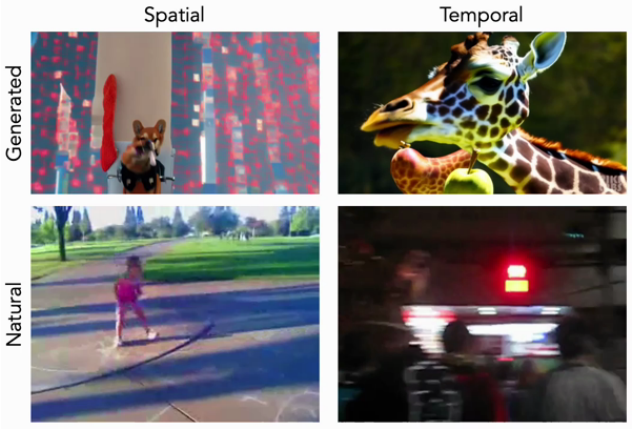


Figure 5. **Domain Gap with Natural Videos.** The common distortions in generated videos (*as in TVGE dataset*) are different from those in natural videos [71], both *spatially* and *temporally*. *We encourage readers to click and play using Adobe Acrobat.*

where s_{pred} and s_{gt} are *lists* of predicted scores and labels in a batch respectively, and sgn denotes the sign function.

4. TVGE Dataset

Motivation. An inalienable part of our study is to evaluate the reliability and robustness of the proposed metrics on text-conditioned generated videos. To this end, we propose the **T**ext-to-**V**ideo **G**eneration **E**valuation (**TVGE**) dataset, collecting rich human opinions on the two perspectives (*alignment & quality*) studied in the T2V Score. On both perspectives, the **TVGE** can be considered as *first-of-its-kind*: First, for the alignment perspective, the dataset will be the first dataset providing text alignment scores rated by a large crowd of human subjects; Second, for the quality perspective, while there are plenty of VQA databases on natural contents [21, 71, 81], they show notably different distortion patterns (both *spatially* and *temporally*, see Fig. 5) from the generated videos, resulting in an non-negligible *domain gap*. The proposed dataset will serve as a validation on the alignment between the proposed T2V Score and human judgments. Furthermore, it can help our quality metric to better adapt to the domain of text-conditioned generated videos. Details of the dataset are as follows.

Collection of Videos. In total, 2543 text-guided generated videos are collected for human rating in the **TVGE** dataset. These videos are generated by *5* popular text-to-video generation models, under a diverse prompt set as defined by EvalCrafter [37] covering a wide range of scenarios.

Subjective Studies. In the TVGE dataset, each video is independently annotated by 10 experienced human subjects from both *text alignment* and *video quality* perspectives. Before the annotation, we trained the human subjects[1] and tested their annotation reliability on a subset of TVGE videos. Each video is rated on a five-point-like scale on either perspective, while examples for each scale are provided in the training materials for subjects.

[1]Training materials are provided in supplementary materials.

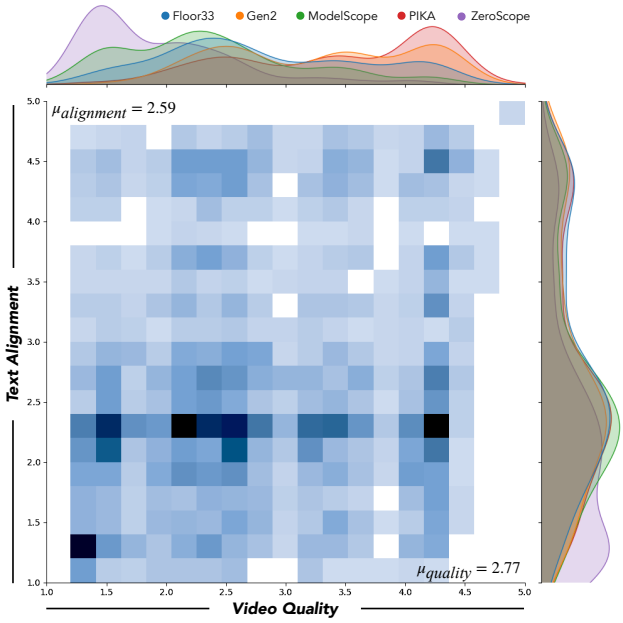


Figure 6. **Score Distributions in TVGE,** suggesting that current text-to-video generation methods generally face challenges in producing video with either good quality or high alignment with text.

Analysis and Conclusion. In Fig. 6, we show the distributions of human annotated *quality* and *alignment* scores in the TVGE dataset. In general, the generated videos receive lower-than-average human ratings ($\mu_{alignment} = 2.59$, $\mu_{quality} = 2.77$) on both perspectives, suggesting the need to continuously improve these methods to eventually produce plausible videos. Nevertheless, specific models also prove decent proficiency on one single dimension, *e.g.* Pika gets an average score of 3.45 on *video quality*. Between the two perspectives, we notice a very low correlation (*0.223* Spearman's ρ, *0.152* Kendall's ϕ), proving that the two dimensions are different and should be considered independently. We show more qualitative examples in the supplementary.

5. Experiments

5.1. Text Alignment

Baselines. We compare our `T2VScore-A` with several standard metrics on text-video alignment, listed as follows:

- *CLIP Score* [18, 45]: Average text-image similarity in the embedding space of image CLIP models [45] over all video frames.
- *X-CLIP Score* [40]: Text-video similarity measured by a video-based CLIP model finetuned on text-video data.

- *BLIP-BLEU* [37]: Text-to-text similarity measured by BLEU [43] score using BLIP-2's image captioning.
- *mPLUG-BLEU*: Same as *BLIP-BLEU* but using a mPLUG-OWL2 for video captioning.

Method	Model	Spearman's ρ	Kendall's τ
Traditional Metric	CLIP Score	0.343	0.236
	X-CLIP Score	0.257	0.175
	BLIP-BLEU	0.152	0.104
	mPLUG-BLEU	0.059	0.040
`T2VScore-A`	Otter†	0.181	0.134
	Video-LLaMA†	0.288	0.206
	mPLUG-OWL2-V†	<u>0.394</u>	<u>0.285</u>
	InstructBLIP*	0.342	0.246
	mPLUG-OWL2-I*	0.358	0.257
	GPT-4V*	**0.486**	**0.360**

† *via* Video QA; * *via* Image QA

Table 1. **Correlation Analysis.** Correlations between objective metrics and human judgment on text-video alignment. Spearman's ρ and Kendall's τ are used for correlation calculation. The best is **bold-faced**, and the second-best is <u>underlined</u>.

Comparison with traditional metrics. We evaluate existing objective metrics using our TVGE dataset and observe a low correlation with human judgment in text-video alignment. This observation aligns with findings in recent research [38, 42] that current objective metrics are incompatible with human perception. In particular, video-based CLIP models exhibit even lower correlations than their image-based counterparts in comprehending videos. This discrepancy may be attributed to the X-CLIP score model, which has been fine-tuned exclusively on the Kinetics datasets, a scope insufficient for broad-domain video understanding. Additionally, while BLEU is a widely employed evaluation metric in NLP research, its effectiveness diminishes in text-video alignment tasks. This is due to the inherent challenge of accurate video captioning. Consequently, video-based models such as mPLUG-Owl-2 prove to be less helpful in this context.

Comparison on MLLMs. Our `T2VScore-A` model is designed to be model-agnostic, which ensures it is compatible with a wide variety of multimodal language learning models (MLLMs). This includes open-source models like Otter [11], Video-LLaMA [83], and mPLUG-OWL2-V [78], as well as proprietary models such as GPT-4V [41]. In our experiments, we tested `T2VScore-A` with both image and video-based LLMs. We found that its performance significantly depends on the capabilities of the underlying MLLMs. Open-source image LLMs like InstructBLIP and mPLUG-OWL2-I show decent results in visual question answering. However, their limited temporal understanding makes them less effective compared to the more advanced open-source video LLMs like mPLUG-OWL2-V in video-based question-answering tasks. Despite this, there is still a notable performance disparity between these open-source MLLMs and GPT-4V, with GPT-4V demonstrating superior performance in video question answering. This is evidenced by its higher correlation with human judgment, outperforming other models by a significant margin.

Effect of auxiliary trajectory. We leverage the point trajectory data generated by CoTracker to enhance fine-grained temporal understanding. This approach effectively captures the subtle motion changes of both the object and the camera, which is instrumental in answering questions related to temporal dynamics. As shown in Fig. 7, models that incorporate trajectory data can accurately identify specific camera movements, such as "panning from right to left" and "rotating counter-clockwise". In contrast, models without trajectory input struggle to perceive these subtle motion changes. The numerical results in Tab. 5 and Tab. 6 further supports our observation.

5.2. Video Quality

Baselines. We compare the `T2VScore-Q` with several state-of-the-art methods on video quality assessment:
- *FAST-VQA* [65]: State-of-the-art technical quality evaluator, with multiple mini-patches ("*fragments*") as inputs.
- *DOVER* [70]: State-of-the-art VQA method, consisting of FAST-VQA and an additional aesthetic branch.
- *MaxVQA* [71]: CLIP-based text-prompted VQA method.

We also validate the performance of multi-modality foundation models in evaluating generated video quality:
- *CLIP* [45, 76]: As CLIP is one of the important bases of the `T2VScore-Q`, it is important to how original zero-shot CLIP variants work on this task. The original CLIPs are evaluated under the same prompts as the proposed semantic expert: *good, high quality*↔*poor, low quality*.

Settings. As discussed in Sec. 3.2, we validate the effectiveness of the `T2VScore-Q` under two settings:
- **(G1):** *zero-shot:* In this setting, no generated videos are seen during model training. Aligning with the settings of off-the-shelf evaluators, it fairly compares between the baseline methods and the proposed `T2VScore-Q`.
- **(G2):** *adapted, cross-model:* In this setting, we further adapt the `T2VScore-Q` to a part of the TVGE dataset with videos generated by one **known** model, and evaluate the accuracy on other 4 **unknown** generation models. It is a rigorous setting to check the reliability of the proposed metric with future generation models coming.

Comparison on the *zero-shot* setting. We show the comparison between the `T2VScore-Q` and baseline methods in Tab. 2, under the *zero-shot* setting without training on any generated videos. Firstly, after our fine-tuning (stage 2, on natural VQA dataset), the two experts that make up the `T2VScore-Q` have notably improved compared with their corresponding baselines; Second, the mixture of the limited

Metric	Spearman's ρ	Kendall's ϕ	Pearson's ρ
FAST-VQA [65]	0.3518	0.2405	0.3460
DOVER [70]	0.3591	0.2447	0.3587
MaxVQA [71]	0.4110	0.2816	0.4002
CLIP-ResNet-50 [45]	0.3164	0.2162	0.3018
CLIP-ViT-Large-14 [76]	0.3259	0.2213	0.3140
the Technical Expert	0.4557	0.3136	0.4426
the Semantic Expert	0.4623	0.3210	0.4353
T2VScore-Q (Ours)	**0.5029**	**0.3498**	**0.4945**
improvements	+22.3%	+24.2%	+23.6%

Table 2. ***Zero-shot* comparison on Video Quality.** Correlations comparison Spearman's ρ, Kendall's ϕ, and Pearson's ρ are included for correlation calculation.

experts also resulted in significant performance gain. Both improvements lead to the final more than **20%** improvements on all correlation coefficients than existing VQA approaches, demonstrating the superiority of the proposed metric. Nevertheless, without training on any T2V-VQA datasets, all *zero-shot* metrics are still not enough accurate to evaluate the quality of generated videos, bringing the necessity to discuss a robust and effective adaptation approach.

Cross-model improvements of adaptation. A common concern on data-driven-generated content quality assessment is that evaluators trained on a specific set of models cannot generalize well on evaluating a novel set of models. Thus, to simulate the real-world application scenario, we abandon the *random* five-fold splits and use rigorous *cross-model* settings during the adaptation stage. As shown in Tab. 3, in each setting, we only adopt the T2VScore-Q on videos generated in **one** among five models in the TVGE dataset and evaluate the changes of accuracy on the rest of videos generated by other 4 models. The table has proven that the proposed prefix-tuning-based adaptation strategy can effectively generalize to **unseen model sets** with an average of **11%** improvements, proving that the T2VScore-Q can be a reliable open-set quality metric for generated videos.

Ablation Studies. We show the ablation experiments in Tab. 4. As is shown in the table, the proposed fine-tuning (stage 2) on both experts improved their single branch accuracy and the overall accuracy of T2VScore-Q, suggesting the effectiveness of the proposed components.

6. Conclusion

In this paper, to address the shortcomings of existing text-to-video generation metrics, we introduced the Text-to-Video Score (T2VScore), a novel evaluation metric that holistically assesses video generation by considering both the alignment of the video with the input text, and the video quality. Moreover, we present the **TVGE** dataset to better evaluate the proposed metrics. The experimental results on the TVGE dataset underscore the effectiveness of T2VScore over existing metrics, providing a more comprehensive and reliable means of assessing text-to-video generation. This proposed metric, along with the dataset, paves the way for further research and development in the field aiming at more accurate evaluation methods for video generation methods.

Strategy	Spearman's ρ	Kendall's ϕ	Pearson's ρ
- Evaluated on other 4 models except *PIKA*			
zero-shot	0.4758	0.3311	0.4643
Trained on *PIKA*, *cross*	**0.5467**	**0.3834**	**0.5377**
- Evaluated on other 4 models except *Floor33*			
zero-shot	0.5467	0.3801	0.5363
Trained on *Floor33*, *cross*	**0.5923**	**0.4192**	**0.5805**
- Evaluated on other 4 models except *ZeroScope*			
zero-shot	0.4148	0.2884	0.4330
Trained on *ZeroScope*, *cross*	**0.4561**	**0.3194**	**0.4623**
- Evaluated on other 4 models except *ModelScope*			
zero-shot	0.4826	0.3340	0.4835
Trained on *ModelScope*, *cross*	**0.5406**	**0.3785**	**0.5368**
- Evaluated on other 4 models except *Gen2*			
zero-shot	0.4964	0.3472	0.4920
Trained on *Gen2*, *cross*	**0.5514**	**0.3895**	**0.5481**
average cross-model gain	+11.2%	+11.2%	+10.6%

Table 3. ***Cross-model* Improvements on Video Quality.** In each setting, we adapt the **T2VScore-Q** with about 500 videos generated with **one** of the models, and test its improvements of accuracy on the rest of the videos generated by the other 4 models.

Components in T2VScore-Q				Spearman's	Kendall's	Pearson's
Q_{sem}	fine-tune	Q_{tech}	fine-tune	ρ	ϕ	ρ
✓				0.3259	0.2213	0.3140
✓	✓			0.4623	0.3210	0.4353
		✓		0.3518	0.2405	0.3460
		✓	✓	0.4557	0.3136	0.4426
✓	✓	✓		0.4458	0.3074	0.4409
✓		✓	✓	0.4629	0.3197	0.4514
✓	✓	✓	✓	**0.5029**	**0.3498**	**0.4945**

Table 4. **Ablation Study.** Spearman's ρ, Kendall's ϕ, and Pearson's ρ are included for correlation calculation.

Limitations and future work. The T2VScore-A heavily relies on multimodal large language models (MLLMs) to perform accurate Visual Question Answering. However, the current capabilities of MLLMs are not yet sufficient to achieve high accuracy, particularly in evaluating temporal dimensions. We anticipate that as MLLMs become more advanced, our T2VScore-A will also become increasingly stable and reliable.

As new open-source text-to-video models continue to emerge, we will keep track of the latest developments and incorporate their results into our TVGE dataset as part of our future efforts.

References

[1] Floor33 pictures. http://floor33.tech/. 2

[2] Gen-2. https://research.runwayml.com/gen2. 1, 2

[3] Pika labs. https://www.pika.art/. 1, 2

[4] Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2000. 5

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 1, 2, 3

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3, 4

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3

[9] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *NAACL*, 2022. 3

[10] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 3

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 4, 7

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 5

[13] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv:2305.10474*, 2023. 2

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J Mach Learn Res*, 2012. 3

[15] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. *arXiv preprint arXiv:2312.02087*, 2023. 3

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv:2211.13221*, 2022. 2

[17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Videocrafter: A toolkit for text-to-video generation and editing. https://github.com/AILab-CVC/VideoCrafter, 2023. 1, 2

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 6

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 1, 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 2

[21] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *QoMEX*, pages 1–6, 2017. 6

[22] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 4, 1

[24] 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 TCSVT*, 2022. 3, 5

[25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 4, 1

[26] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 3

[27] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM MM*, page 2351–2359, 2019. 3

[28] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *ACM MM*, page 789–797, 2020. 5

[29] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4): 1238–1257, 2021. 5

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 4

[31] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao.

Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 4

[32] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4

[35] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. *arXiv preprint arXiv:2310.10624*, 2023. 3

[36] Xiaohong Liu, Xiongkuo Min, Wei Sun, et al. Ntire 2023 quality assessment of video enhancement challenge, 2023. 3

[37] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2023. 6, 7

[38] 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. *arXiv preprint arXiv:2311.01813*, 2023. 2, 7

[39] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 4

[40] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022. 6

[41] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 7

[42] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 7

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 7

[44] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, pages 11410–11420, 2022. 2, 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 6, 7, 8

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[47] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*, 2020. 1, 3

[48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1, 2, 3

[49] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.*, 28(2): 612–627, 2019. 5

[50] Spencer Sterling. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 1, 2

[51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3

[53] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv:1812.01717*, 2018. 3

[54] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR*, 2019. 1, 3

[55] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022. 3, 5

[56] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv:2308.06571*, 2023. 1, 2

[57] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[58] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 4

[59] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv:2305.10874*, 2023. 2

[60] 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:2309.15103*, 2023. 2

[61] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process Mag*, 2009. 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 2

[63] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 2

[64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 4

[65] 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 *ECCV*, 2022. 3, 5, 7, 8

[66] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment, 2023. 3, 5

[67] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4840–4854, 2023. 3

[68] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *International Conference on Multimedia and Expo (ICME)*, 2023. 3

[69] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023. 3, 5

[70] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 3, 5, 7, 8

[71] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable video quality assessment: A database and a language-prompted approach. In *ACM MM*, 2023. 5, 6, 7, 8

[72] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. 2023. 3

[73] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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 *ICCV*, 2023. 2, 3

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

[75] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv:2308.09710*, 2023. 3

[76] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 5, 7, 8

[77] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 4

[78] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 4, 7

[79] Joong Gon Yim, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. Subjective quality assessment for youtube ugc dataset. In *ICIP*, 2020. 3

[80] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv:2303.12346*, 2023. 2

[81] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: 'patching up' the video quality problem. In *CVPR*, 2021. 3, 5, 6

[82] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv:2309.15818*, 2023. 1, 2, 3

[83] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 7

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 5

[85] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv:2305.13077*, 2023. 3

[86] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 2, 3

[87] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv:2211.11018*, 2022. 2

[88] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 5

[89] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 4

Towards A Better Metric for Text-to-Video Generation

Supplementary Material

7. More Details for Subjective Study

Annotation Interface. In Fig. 11, we show the annotation interface for the subjective study in the **TVGE** dataset. The text alignment scores and quality scores are annotated separately to avoid distraction from each other. The input text prompt is shown only for the *alignment* annotation (which is necessary), but not displayed for *quality* annotation, so that the *quality* score can ideally only care about the visual quality of the generated video.

Training Materials. Before annotation, we provide clear criteria with abundant examples of 5-point Likert scale to train the annotators. For **text alignment**, we specifically instruct annotators to evaluate the videos based solely on the presence of each element mentioned in the text description, intentionally ignoring video quality. For **video quality**, we ask the subjects to focus exclusively on technical distortions. We provide five examples for each of the six common distortions: 1) noises; 2) artifacts; 3) blur; 4) unnatural motion; 5) inconsistent structure; and 6) flickering. Samples of the annotated videos can be viewed in Fig. 8.

8. Additional Results

Effect of Auxiliary Trajectory for `T2VScore-A`. As mentioned in Sec. 3.1, we utilize the auxiliary point-level trajectory generated by CoTracker [23] to enhance fine-grained temporal understanding. Fig. 7 presents video samples that exhibit temporal nuances, which state-of-the-art multimodal language models (MLLMs) often fail to detect. Using the trajectory as auxiliary information effectively improves the MLLMs' understanding of subtle temporal changes in camera and object motion. For instance, the *snake* in row 5 appears motionless, though the camera is moving. Upon ablating the auxiliary trajectory, we observe a decrease in visual question answering (VQA) accuracy from 0.58 to 0.48, as shown in Tab. 6. This reduction in VQA accuracy further leads to a diminished alignment with human judgment (see Tab. 5).

`T2VScore-A`	Spearman's ρ	Kendall's ϕ	Pearson's ρ
GPT-4V *w/o tarjectory*	0.4454	0.3289	0.4416
GPT-4V	**0.4859**	**0.3600**	**0.4882**

Table 5. **Effect of Auxiliary Trajectory.** Spearman's ρ, Kendall's ϕ, and Pearson's ρ are included for correlation calculation.

Performance of state-of-the-art MLLMs in VQA. We setup an evaluation set of 500 videos (100 prompts with 5 unique videos per prompt) sampled from our TVGE dataset. Two annotators are tasked with responding to the generated questions, and a third, more experienced annotator is assigned to verify these responses. We compare the accuracy of visual question answering (VQA) across a range of multimodal large language models (MLLMs), focusing on *spatial* and *temporal* QA. As shown in Tab. 6, current MLLMs generally demonstrate weak performance in open-domain VQA tasks, with temporal QA faring even worse. Notably, video-based MLLMs are inferior in temporal QA compared to their image-based counterparts. A similar observation is made in SEED-Bench [25], indicating significant room for further improvement in video-based MLLMs.

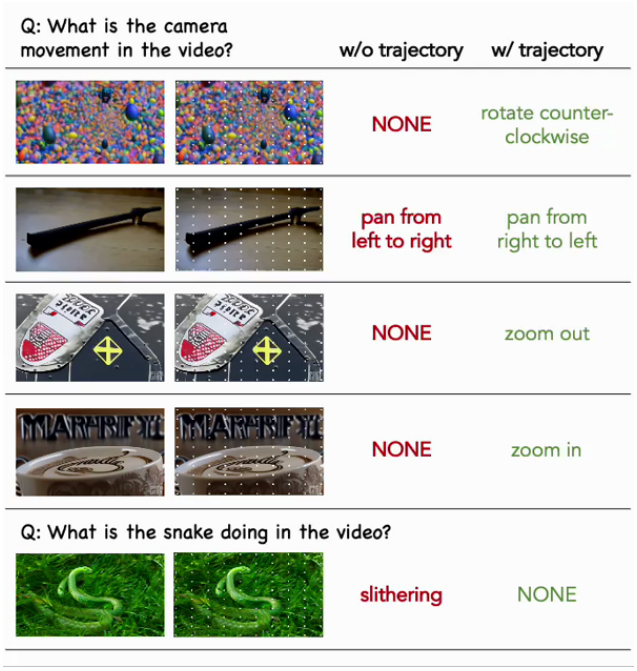


Figure 7. **Quantitative Examples for Auxiliary Trajectory.** Using an auxiliary trajectory effectively enhances multimodal large language models (MLLMs) for fine-grained temporal understanding. *Please click and play using Adobe Acrobat.*

Model	Overall	Temporal QA	Spatial QA
random guess	0.2369	0.2452	0.2327
Otter†	0.1460	0.1059	0.1636
Video-LLaMA†	0.4074	0.3459	0.4435
mPLUG-OWL2-V†	0.5305	0.4280	0.5880
InstructBLIP*	0.5013	0.4762	0.5127
mPLUG-OWL2-I*	0.5107	0.4333	0.5600
GPT-4V* *(w/o trajectory)*	0.4791	0.4411	0.5589
GPT-4V*	**0.5765**	**0.5077**	**0.6308**

† *via* Video QA; * *via* Image QA

Table 6. **Accuracy of Visual Question Answering.**

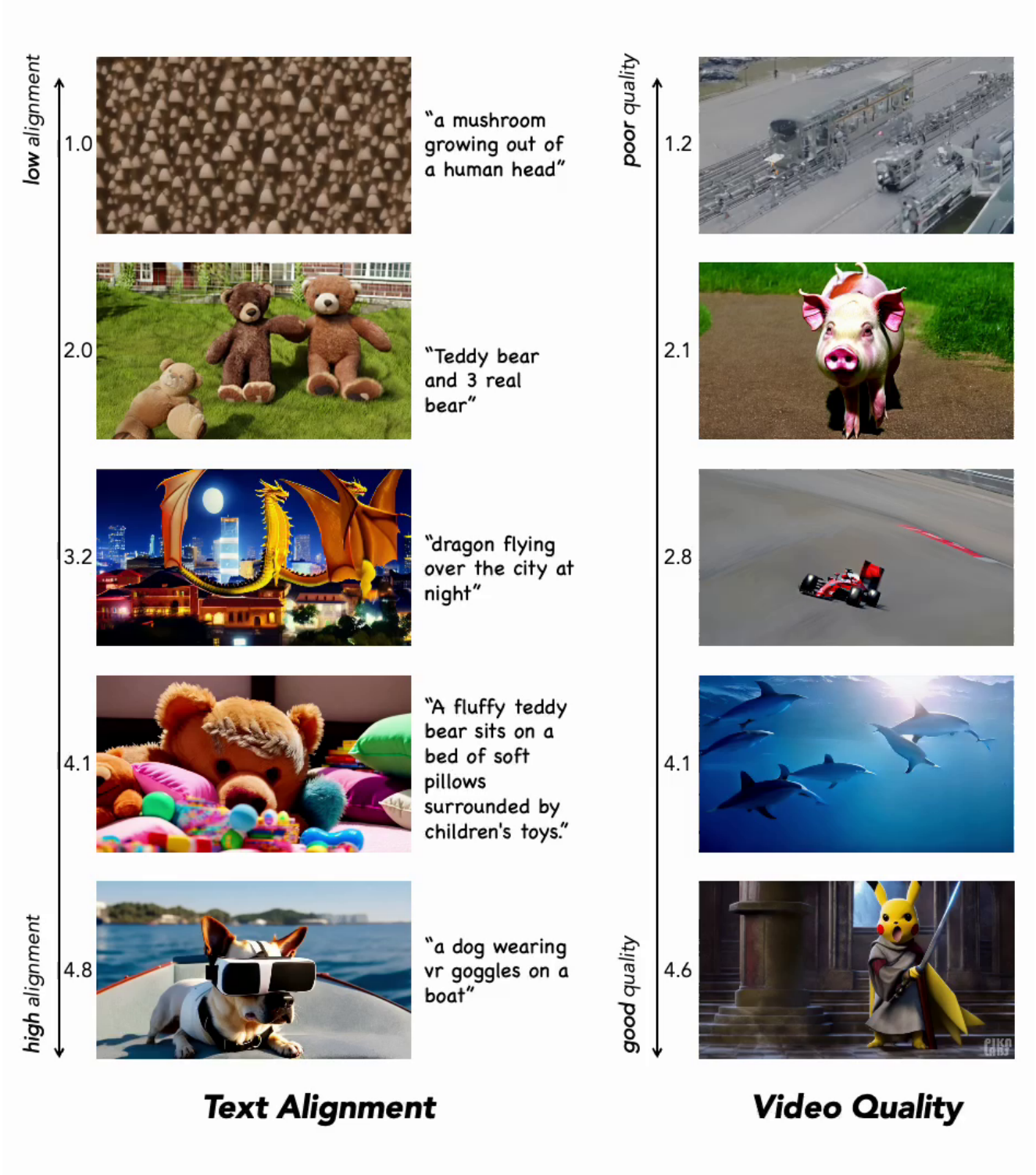


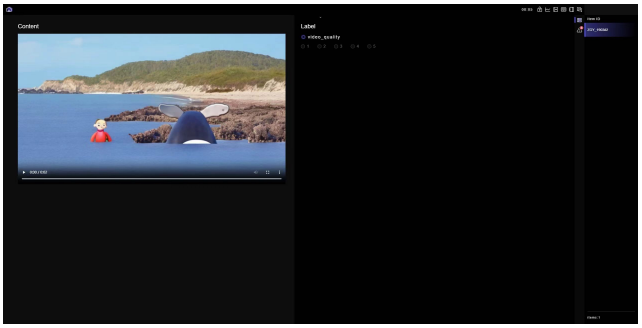
Figure 8. **Human Annotation.** Generated videos and their human ratings of *text alignment* and *video quality*. The scores are the mean of 10 annotators' ratings. *Please click and play using Adobe Acrobat.*

Figure 9. **More Examples for `T2VScore-A`.** We showcase more examples illustrating how `T2VScore-A` is computed. *Please click and play using Adobe Acrobat.*

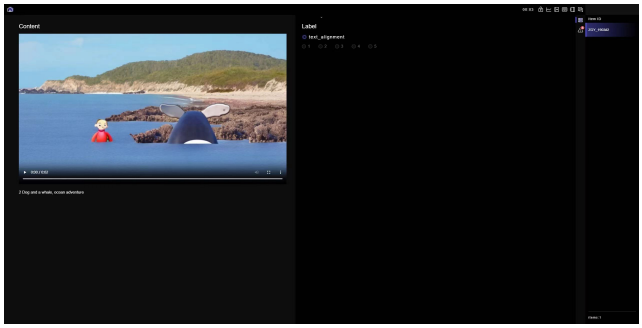
```
# Task Description:
The T2VScore-A is an evaluator for assessing the text alignment of video content generated from textual descriptions. It
scrutinizes the video descriptions and formulates structured questions and answers to ensure the video content aligns precisely
with the provided description.

# Task Steps and Format Specification:
## Input Processing:
On receiving a video description, the T2VScore-A decomposes it into atomic tuples, ensuring each tuple is the smallest unit of
meaning that accurately represents an aspect of the video. Each atomic tuple consists of a correlation: the first element
indicates a global or local Object of the video, and the second element specifies the attribute or detail of that Object,
including but not limited to' activity', 'attribute', 'counting', 'color', 'material', 'spatial', 'location', 'shape', 'OCR',
etc. Questions are formulated based on the atomic tuples' count and order.

## Question Generation:
Generate questions for each atomic tuple, targeting a specific video aspect indicated by the tuple. Questions should reflect the
atomicity principle, avoiding over-fragmentation or excessive aggregation of concepts.

## Answer Formulation:
Provide direct, relevant choices for each question based on the atomic tuple. Include "NONE" as an option where the existence of
an entity or attribute is uncertain. Position the answer immediately following the question for clarity.

## Response Format:
Start with a list of the derived atomic tuples from the video description. Follow with each question and its corresponding
choices. Present the correct answer immediately after each question.

## Consideration for Entity Non-Existence:
Questions related to entities must account for the possibility that the entity may not be present in the video. Hence, the
format "If there is a [entity], ..." should be used where applicable.
```

```
Input: Iron Man is walking towards the camera in the rain at
night, with a lot of fog behind him. Science fiction movie,
close-up
Atomic Tuples:
(entity, Iron Man)
(Iron Man, walking)
(global, towards the camera)
(global, in the rain)
(global, at night)
(global, a lot of fog behind him)
(global, Science fiction movie)
(global, close-up)
Questions and Answers:
Q: What is the name of the character in the video?
Choices: Iron Man, Captain America, Thor, Hulk
A: Iron Man
Q: What is the character doing in the video?
Choices: walking, jumping, flying, NONE
A: walking
Q: What direction is the character moving in the video?
Choices: towards the camera, away from the camera, left to
right, right to left
A: towards the camera
Q: What is the weather condition in the video?
Choices: sunny, rainy, snowy
A: rainy
Q: What time of day is depicted in the video?
Choices: morning, afternoon, evening, night
A: night
Q: What is behind Iron Man in the video?
Choices: a lot of fog, a cityscape, a forest, a desert
A: a lot of fog
Q: What genre does the video belong to?
Choices: comedy, drama, science fiction, horror
A: science fiction
Q: What type of shot is used in the video?
Choices: wide shot, medium shot, close-up, extreme close-up
A: close-up

Input: 2 Dog and a whale, ocean adventure
Atomic Tuples:
(entity, dog)
(dog, 2)
(entity, whale)
(whale, a)
(global, ocean adventure)
Questions and Answers:
Q: Which type of animal appear in the video except for the
whale?
Choices: dog, cat, bird, fish, NONE
A: dog
Q: How many dogs are present in the video?
Choices: 1, 2, 3, more than 3, NONE
A: 2
Q: Which type of animal appears in the video except for the
dog?
Choices: whale, cat, bird, fish, NONE
A: whale
Q: How many whales are present in the video?
Choices: 1, 2, 3, more than 3, NONE
A: 1
Q: What is the theme of the video?
Choices: city adventure, forest adventure, ocean adventure,
mountain adventure
A: ocean adventure

Input: There's a person, likely in their mid-twenties, with
short brown hair and a tattoo on their left forearm. They're
wearing a blue hoodie and seem to be checking their phone.
Atomic Tuples:
(entity, person)
(person, a)
(person, mid-twenties)
(person, short hair)
(person, brown hair)
(person, tattoo)
(tattoo, left forearm)
(person, wearing a hoodie)
(hoodie, blue)
(person, checking their phone)
Questions and Answers:
Q: What is the main entity in the video?
Choices: person, dog, cat, bird
A: person
Q: How many persons are present in the video?
Choices: 1, 2, 3, more than 3, NONE
A: 1
Q: If there is a person, what is their approximate age?
Choices: teenager, mid-twenties, thirties, forties, NONE
A: mid-twenties
Q: If there is a person, what is the length of their hair?
Choices: short, medium, long, NONE
A: short
Q: If there is a person, what is the color of their hair?
Choices: black, brown, blonde, red, NONE
A: brown
Q: If there is a person, do they have a tattoo?
Choices: yes, no
A: yes
Q: If there is a person, where is their tattoo located?
Choices: left forearm, right forearm, back, chest, NONE
A: left forearm
Q: If there is a person, what are they wearing?
Choices: hoodie, t-shirt, sweater, NONE
A: hoodie
Q: If there is a person, what color is the hoodie they are
wearing?
Choices: blue, red, black, white, NONE
A: blue
Q: If there is a person, what are they doing in the video?
Choices: checking their phone, reading a book, eating,
sleeping, NONE
A: checking their phone
```

Figure 10. **Prompt for Question/Answer Generation in `T2VScore-A`.** *Top:* task instruction; *Bottom:* in-context learning examples.



(a) **Annotation Interface** for *Video Quality*. The video is presented to subjects to be rated a quality score among [1,5].



(b) **Annotation Interface** for *Text Alignment*. The video and its text prompt are presented to subjects to be rated an alignment score among [1,5].

Figure 11. **Annotation Interface** for Video Quality (a) and Text Alignment (b).