

---

# AI-Generated Video Evaluation: A Survey

---

Xiao Liu<sup>1\*</sup>, Xinhao Xiang<sup>1\*</sup>, Zizhong Li<sup>1\*</sup>, Yongheng Wang<sup>2</sup>,  
Zhuoheng Li<sup>1</sup>, Zhuosheng Liu<sup>1</sup>, Weidi Zhang<sup>2</sup>, Weiqi Ye<sup>2</sup>, and Jiawei Zhang<sup>1†</sup>

<sup>1</sup>IFM Lab, University of California, Davis

<sup>2</sup>University of California, Davis

{xioliu, xhxiang, zzoli, pipli, zsliu, jiwzhang}@ucdavis.edu  
{yhgwang, vikye, wdizhang}@ucdavis.edu

## Abstract

The growing capabilities of AI in generating video content have brought forward significant challenges in effectively evaluating these videos. Unlike static images or text, video content involves complex spatial and temporal dynamics which may require a more comprehensive and systematic evaluation of its contents in aspects like video presentation quality, semantic information delivery, alignment with human intentions, and the virtual-reality consistency with our physical world. This survey identifies the emerging field of AI-Generated Video Evaluation (AIGVE), highlighting the importance of assessing how well AI-generated videos align with human perception and meet specific instructions. We provide a structured analysis of existing methodologies that could be potentially used to evaluate AI-generated videos. By outlining the strengths and gaps in current approaches, we advocate for the development of more robust and nuanced evaluation frameworks that can handle the complexities of video content, which include not only the conventional metric-based evaluations, but also the current human-involved evaluations, and the future model-centered evaluations. This survey aims to establish a foundational knowledge base for both researchers from academia and practitioners from the industry, facilitating the future advancement of evaluation methods for AI-generated video content.

## 1 Introduction

With the introduction and widespread integration of large generative models like ChatGPT [172], Sora [173], LLaMA [231], and the recent Meta Movie Gen [156], AI-generated content has become increasingly significant in both the production and consumption of contents. In the domain of production, text and video professionals increasingly use generative tools to create and enhance content, from scripts and articles to complex visual sequences, which would traditionally require extensive time and effort to achieve manually, thus streamlining creative workflows and enhancing productivity [103, 303, 7]. On the consumer front, reliance on outputs from generative models has also become commonplace, with applications ranging from information retrieval to the automation of routine tasks. This shift represents a significant transformation from the pre-2023 era, when such tasks were predominantly manual.

As this trend continues to evolve, the methods that can automatically evaluate these AI-generated contents become crucial. These methods help ensure that such content aligns well with human per-

---

\*Equal Contribution

†Corresponding Author

©All rights reserved by authors.

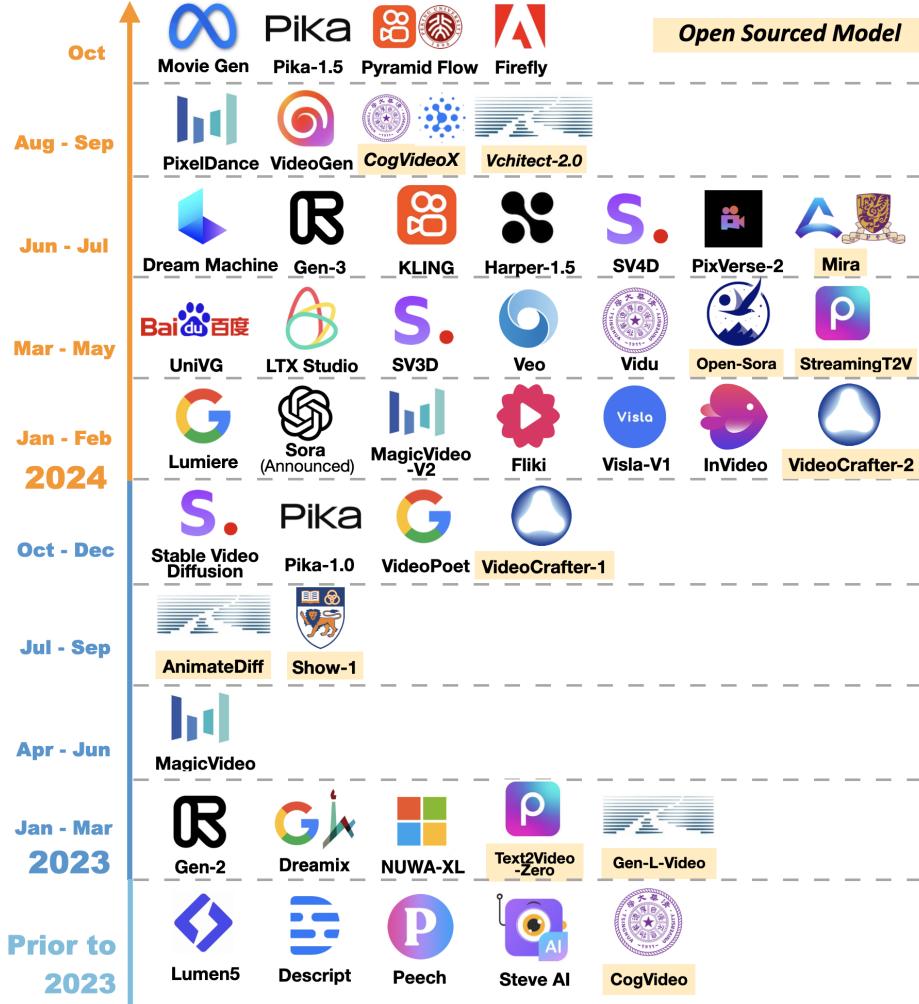


Figure 1: Evolution of Video Generation Models Over Time.

ceptions and instructions. While the evaluation of AI-generated text and images has been thoroughly studied, using techniques including word and pixel matching [128, 73, 204, 178], sophisticated modeling [72, 292, 63], and Large/Vision Language Model evaluation [56, 109, 79, 294], the assessment of AI-generated videos presents unique challenges that are yet to be comprehensively addressed. This discrepancy highlights a significant gap in research, particularly at a time when advanced video generation models are rapidly evolving, as illustrated in Figure 1, and video content is gaining increasing prominence in both professional and personal domains.

Video, by its nature, incorporates both spatial complexity and temporal dynamics, making its evaluation intrinsically more complex than static images or text. Traditional Video Quality Assessment (VQA) metrics focus on technical aspects such as compression effects, transmission quality, [137, 40, 152, 181, 154]. Recent research has shifted focus towards evaluating the perceptual quality of user-generated, in-the-wild videos, considering factors like blur, motion stability, and noise levels [266, 255, 104]. These methods assess whether a video effectively conveys visual information to the viewer. However, in the era of AI-Generated Content (AIGC), an additional critical aspect is whether the generated content aligns with the creator's instruction, which remains less explored within current VQA frameworks.

Besides, current research in the evaluation of AI-generated videos is rapidly emerging but remains unstructured. The continuous introduction of new models and evaluation metrics complicates the process of identifying comprehensive resources, as they are often scattered across various domains. Moreover, those methods focus on various evaluation aspects, which can overlap or be entirely

disjoint. As a result, the absence of a unified framework impedes the progress of the field, resulting in fragmented research efforts. Therefore, there is a critical need for a more cohesive approach in this fast-moving area.

In this survey, we aim to highlight a new area focused on AI-Generated Video Evaluation (AIGVE). To devote our effort, we have collected and integrated existing research related to this field to help academic researchers and industrial practitioners locate the essential foundational knowledge. Our focus is on existing works from related research areas, and we have conducted extensive research on fields such as VQA [159], multimodal text-visual alignment [278], and recent emerging AIGVE evaluation methods [139, 82, 157]. By exploring and categorizing methods related to AIGVE, we aim to build a solid foundation for AI-generated video evaluation and support future research efforts in this rapidly evolving area.

Our contributions through this survey are summarized as:

- **Highlighting an Emerging Field:** We propose and emphasize the need for a new research area on AI-Generated Video Evaluation (AIGVE).
- **Comprehensive Review of Existing Evaluation Methods:** This survey provides a systematic and comprehensive review of current methodologies relevant to AIGVE from multiple research fields. We categorize and analyze these approaches to provide a well-structured outline of the existing landscape.
- **Guidance for Future Research Directions:** We also locate several potential areas that call for more future investigations and development in AIGVE. These areas include integrating evaluation frameworks with vision language models, enhancing the interpretability of evaluation scores, and addressing the ethical and safety considerations of these frameworks. This survey aims to serve as a foundational resource for researchers and industrial practitioners, providing insights that can guide the advancement of more effective and comprehensive evaluation methodologies for AI-generated video content.

## 2 Advancements in Video Generation

Generating videos that are consistent with the offline reality world and conform to the currently known world physical laws has long been a popular and necessary research topic. Most of the previous studies perform video generation tasks based on three different types of generative models: 1) Generative Adversarial Networks (GAN) [246, 237, 6, 18, 91, 283, 55, 230], 2) Autoregressive Transformers [262, 59, 76, 102, 245, 275, 222], and 3) Diffusion Models [75, 212, 74, 20, 287, 19, 129].

**GAN-based Models:** In the early exploration of video generation, GAN-based models used to be the mainstream approach, which also can be seen as the temporal extension of image GAN-based generation [153, 192, 273, 288, 183, 92, 89]. Specifically, the GAN-based generative model consists of a generator network (i.e., aims to generate videos) and a discriminator network (i.e., tries to distinguish which ones are generated “fake” videos). In this training phase of the model, this iteration keeps going until a balance is reached where the generator can generate high-quality videos that confuse the discriminator, while the discriminator can maximize the recognition of “fake” generated videos. Vondrick et al. [246] first leverages GAN for video generation, differentiating the video into moving foreground spatiotemporal convolutions and static background spatial convolutions. Tulyakov et al. [237] further adopts a motion and content decomposed representation for video generation within a recurrent mechanism to generate motion embeddings and a CNN framework to generate videos. Instead of using 2D convolutions or recurrent networks to represent the time dimension, several following works [6, 18] also seek to use 3D convolution networks to harmonize video generation temporally. In addition to modeling the temporal representation, Karras et al. [91] tries to progressively generate high-resolution videos in spatial and temporal directions starting with low-resolution images.

**Autoregressive Transformer-based Models:** Meanwhile, the generalizability of the transformer model [242] and its effectiveness on a wide range of tasks [130, 65] make it an alternative pathway to consider for video generation. Similar to the techniques used with transformers in text, video generation also converts the input modalities, such as text and images, into token sequences using an encoder. Then, the autoregressive transformer is trained to decode each frame of the generated videos. Wu et al. [262] first uses VQ-VAE [241] and a three-dimensional sparse attention mechanism for

open-domain text-to-video generation. Ge et al. [59] proposes to combine a time-agnostic VQGAN for generating images with a time-sensitive transformer to generate long videos. Inheriting the remarkable pretrained models in text-to-image generation, Hong et al. [76] proposes a multi-frame-rate hierarchical training strategy based on [48] to better align text and video clips.

With the emergence and rapid growth of Large Language Models (LLMs), a few works [102, 245, 275, 222] seek to leverage the power of LLMs for video generation. One of the most representative works, Kondratyuk et al. [102], encodes all modalities (i.e., text, image, depth+optical flow, masked video) into the discrete token space and directly uses LLM architectures for video generation.

**Diffusion Models:** In recent years, the significant breakthrough in text-to-image generation [203, 53, 46, 169, 199, 57] has made diffusion models the dominant approach in recent video generation studies. Inspired by the successful utilization of diffusion models in image generation, many studies try to use them in video generation. The diffusion models [216] rely on iterative denoising samples drawn from a noise distribution to generate final results. For text-to-video generation, the noisy input video and the corresponding text embeddings are used to feed into the denoising network. Ho et al. [75] first extends the standard image diffusion architecture to video data, modeling entire videos using a 3D U-Net diffusion model architecture. Singer et al. [212] extend the text-to-image diffusion models with pseudo-3D convolutional and attention layers, each spatial 2D-Conv layer is followed by a temporal 1D-Conv layer, reducing the computational resource compared to computing 3D-Conv layers. Based on the prior video diffusion models, Ho et al. [74] explores generating higher definition videos with spatial super-resolution (SSR) and temporal super-resolution (TSR) models; Blattmann et al. [20] extends the text-to-image Latent Diffusion Models (LDM) Stable Diffusion model to generate long-term videos with high resolution.

Zhang et al. [287] further combines the pixel-based and latent-based text-to-video diffusion models (VDMs) for the generation: it first uses pixel-based VDMs to produce a low-resolution video, then translates it via latent-based VDMs to further upsample the low-resolution video to high resolution. Blattmann et al. [19] demonstrates the necessity of the well-curated pretraining dataset for generating high-quality videos, presenting a systematic data preprocessing workflow including captioning and filtering strategies for data preparation. The recent work Lin et al. [129] further adapts pretrained ControlNets [290] to video generation, which supports video control, sparse-frame video control, and a variety of downstream tasks such as video editing and text-guided motion control.

The aforementioned advancements in foundational techniques have substantially enhanced the performance of recent video generation models. These models have experienced rapid growth in both open-source and commercial arenas, achieving increasingly higher fidelity and extended duration in generated content, as shown in the upper section of Figure 2.

Nevertheless, substantial challenges remain concerning the quality of generated videos. As presented in the lower section of Figure 2, generated videos often exhibit two major issues: (1) misalignment with human perceptual expectations and (2) deviation from the creator’s instructions, resulting in misleading outputs. These issues can be subtle and difficult to detect upon initial review. Addressing these shortcomings necessitates the development of robust evaluation methodologies. However, aside from human evaluation, there remains a significant gap in systematic and comprehensive automated evaluation frameworks to consistently assess the quality of AI-generated videos.

### 3 AI-Generated Video Evaluation

The field of evaluating AI-generated videos is in its early stages. As synthesized video content becomes increasingly prevalent, there is a demand for effective evaluation methods that align with the intentions of creators and the perceptions of viewers. This survey seeks to outline a preliminary framework for AI-Generated Video Evaluation (AIGVE), recognizing that our definitions and understandings will evolve as the field matures.

By extensively reviewing and categorizing existing research [159, 266, 141, 139], we propose that AI-generated videos should ideally satisfy two principal criteria: 1) alignment with human perception and 2) alignment with human instructions.

**Alignment with Human Perception.** This aspect emphasizes evaluating video quality by assessing traditional metrics such as high resolution, clarity, and the absence of noise [159]. Besides, AI-generated content introduces additional complexity, as videos must also maintain consistency with the

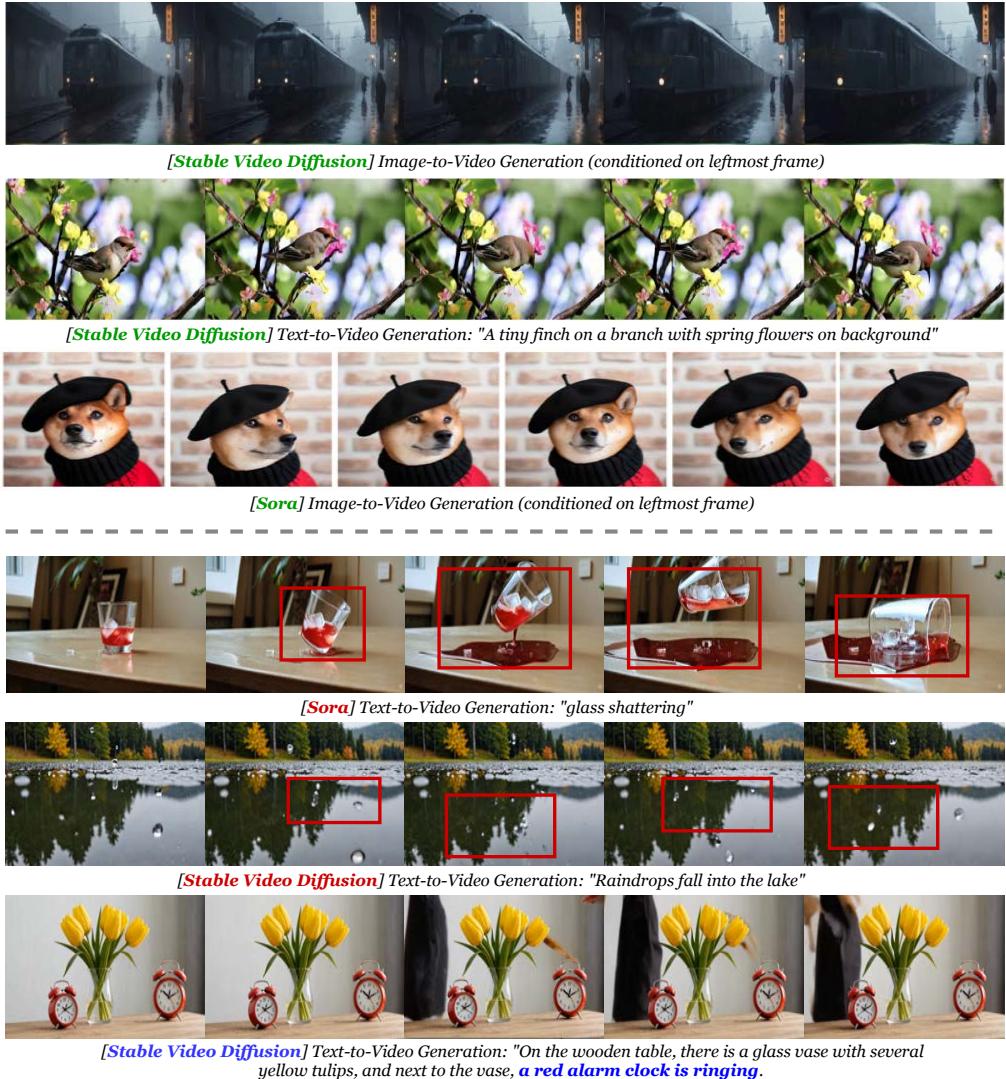


Figure 2: Case Study of AI-Generated Videos. Although current studies can generate high-quality videos (i.e., the green cases), the generated videos still have flaws in certain conditions including physical perception error (i.e., the red cases) and incoherence with the instructions (i.e., the blue cases). Specifically, the areas in red bound boxes indicate the anomaly physical perception contents in the videos, and the blue highlighted fonts indicate the incoherence between human text instructions and the contents in generated videos.

physical world [205, 226]. This involves realistic texture rendering, accurate color representation, and adherence to physical laws, ensuring that the videos are not only of high quality but also believable and immersive.

**Alignment with Human Instructions.** With the introduction of advanced generative AI technologies, a new challenge has emerged in ensuring that videos align precisely with detailed human instructions, which are currently primarily text-based. This involves generating content that accurately mirrors described scenarios, actions, and narratives, thus fulfilling the creative and communicative objectives conveyed by creators. This alignment ensures that the video content not only meets technical standards but also fulfills the creative and communicative intentions of the creator, making it a true reflection of human instruction.

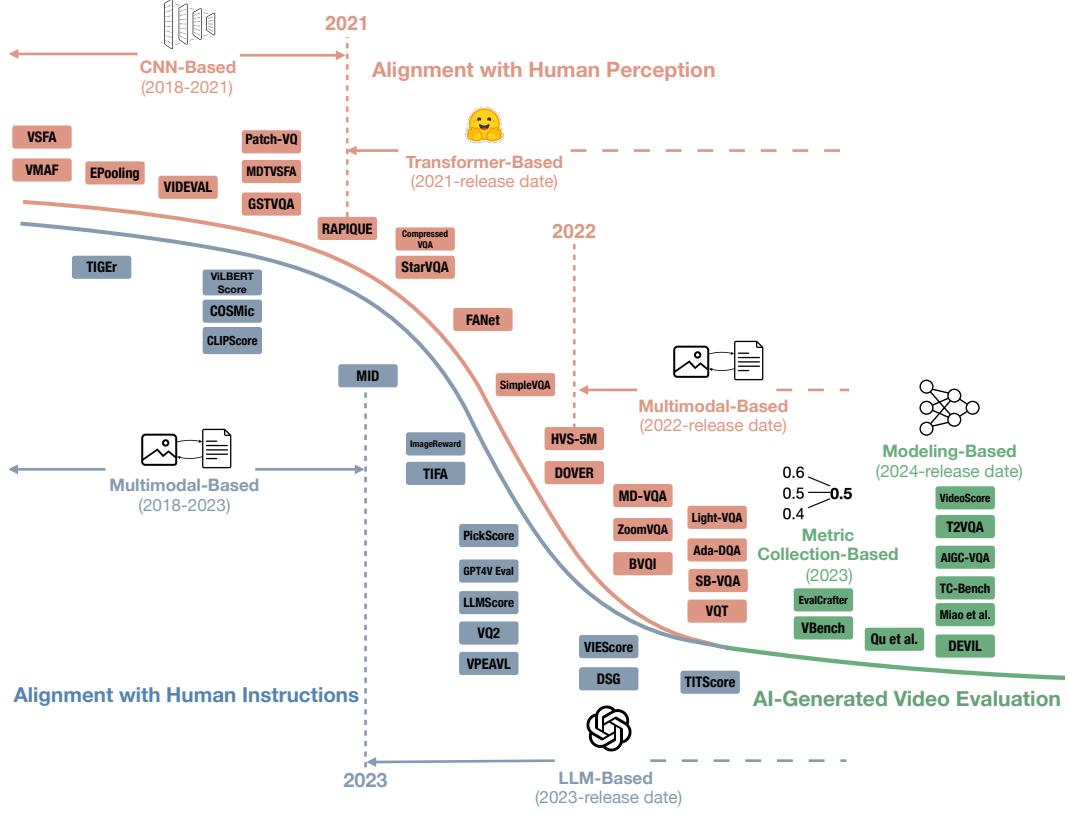


Figure 3: The development and overview of **AI-Generated Video Evaluation (AIGVE)**. AIGVE was built on two initially separate aspects: 1) **Alignment with human perception**, and 2) **Alignment with human instructions**. Note that the timeline scales are different for two aspects. Release date represents the date that this survey is released.

Figure 3 illustrates the development and overview of both aspects and highlights the emergence of AIGVE. Initially, the evaluation of video alignment with human perception and instructions were separate research areas. With the rise of AI-generated videos, both areas need to be considered when evaluating AI-generated videos.

In the remainder of this section, we present the development of AIGVE. It begins by detailing the creation of the benchmark dataset, transforming video-opinion pairs into video-instruction-opinion triplets for more nuanced evaluation. The evaluation methods are then categorized into two main approaches: metric collection evaluation, which utilizes existing metrics to assess various aspects of video quality, and modeling evaluation, where new models are developed and trained on the collected datasets to simulate human judgment.

### 3.1 Benchmark Datasets for AI-Generated Video Evaluation

The construction of benchmark datasets is a crucial preliminary step for both training evaluation models and assessing current video generation models. Recent studies have devoted significant effort to creating large-scale and robust benchmarks.

To align better with human perception and instructions, current research extends the dataset format for video quality assessment from video-opinion pairs  $D = \{V, S\}$  to video-instruction-opinion triplets  $D = \{V, I, S\}$ , where  $V$  represents the video,  $I$  refers to the instruction, and  $S$  denotes the opinion score. We summarize the current standard data collection steps for AIGVE in Figure 4.

In the instruction collection step, the objective is to gather a set of high-quality video generation instructions, denoted as  $I$ , corresponding to various aspects such as dynamics or composition. Note that the instructions could be filtered from the pre-existing text-to-video dataset such as MSR-VTT [272], WebVid [12], or generated by large language models under desired criteria.

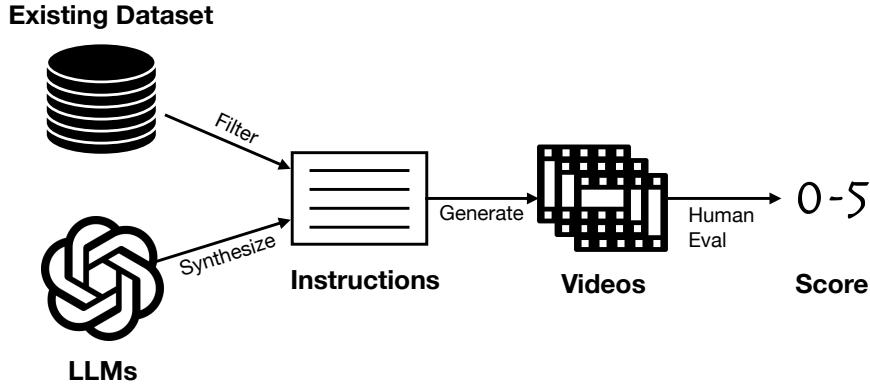


Figure 4: AIGVE Benchmark Dataset Collection Process.

Table 1: Summary of recent AI-Generated Video Evaluation Benchmark datasets: '# Video' is the number of unique videos, '# Frames' the frames per video, '# Prompt' the unique prompts, '# Category' the number of video categories, '# Generators' the number of video generation models, '# Evaluators' the number of human evaluators, and 'Aspect' the evaluation focus.

Dataset	Year	# Video	# Frames	Resolution	# Prompt	# Category	# Generators	# Evaluators	Aspect	Link
EvalCrafter [139]	2023	2,500	8	> 512p	500	4	5	3	general	<a href="#">🔗</a>
VBench [82]	2023	TBD	> 6	> 240p	1746	24	8	-	general	<a href="#">🔗</a>
Chivileva et al. [37]	2023	1,005	> 8	> 128p	201	2	5	24	naturalness	<a href="#">🔗</a>
T2VQA-DB [106]	2024	10,000	16	512p	1,000	7	9	27	general	<a href="#">🔗</a>
T2VBench [86]	2024	5,000	> 16	> 256p	1,600	16	3	3/data	temporal	<a href="#">🔗</a>
VIDEOPHY [15]	2024	9,300	> 25	> 240p	688	3	9	-	physical	<a href="#">🔗</a>
TC-Bench [54]	2024	817	16	> 256p	150	3	5	8	composition	<a href="#">🔗</a>
T2VSafetyBench [157]	2024	17,600	> 25	> 240p	4,400	12	4	60	safety	<a href="#">🔗</a>
VIDEOFEEDBACK [67]	2024	37,600	> 8	> 256p	44,500	5	11	20	general	<a href="#">🔗</a>
AIGC-VQA [145]	2024	10,000	16	512p	1,000	3	10	20	general	<a href="#">🔗</a>
DEVIL [126]	2024	800	> 16	> 240p	4,800	5	6	6	dynamics	<a href="#">🔗</a>
GAIA [36]	2024	9,180	> 4	> 256p	510	3	18	54	action	<a href="#">🔗</a>

These instructions are then fed into  $n$  selected text-to-video models to generate  $n$  videos for each instruction, resulting in the video set  $V$ . Finally, a group of human evaluators is invited and instructed to score each video based on a specified evaluation process, producing the set of scores  $S$ . Table 1 summarizes the current emerging datasets that evaluate both human perception and instruction. Figure 5 presents a detailed analysis of the proportion of videos generated by each model in each benchmark dataset.

**EvalCrafter**[139]: EvalCrafter is one of the first works that focus on comprehensive AIGVE created in 2023. The dataset consists of 2,500 videos, each containing 8 frames with a resolution greater than 512p. It includes 500 prompts generated based on real-world user data and large language model assistance, ensuring a diverse and comprehensive set. The evaluation framework assesses the generated videos across multiple aspects, including visual quality, motion quality, temporal consistency, and text-video alignment. There are 17 objective metrics used, including Inception Score (IS) [204] for video quality, CLIP-Score [71] for text-video consistency, Flow-Score [229] for general motion information, and Warping Error [229] for temporal consistency. Additional metrics include Dover [265], SD-Score [199] for comparing generated video frames, BLIP-BLEU [120, 179] for text alignment evaluation, and several object and attribute consistency scores such as Detection-Score, Count-Score, and Color-Score. Human evaluations are aligned with these objective metrics using a linear regression model to correlate user scores with the evaluation results.

**VBench**[82]: VBench, introduced in 2023, is another pioneer comprehensive benchmark suite designed to evaluate video generative models by breaking down "video generation quality" into 16 distinct, hierarchical dimensions. The dataset includes videos with more than 6 frames each at a resolution greater than 240p, along with 1,746 prompts across 24 sub-categories. The evaluation

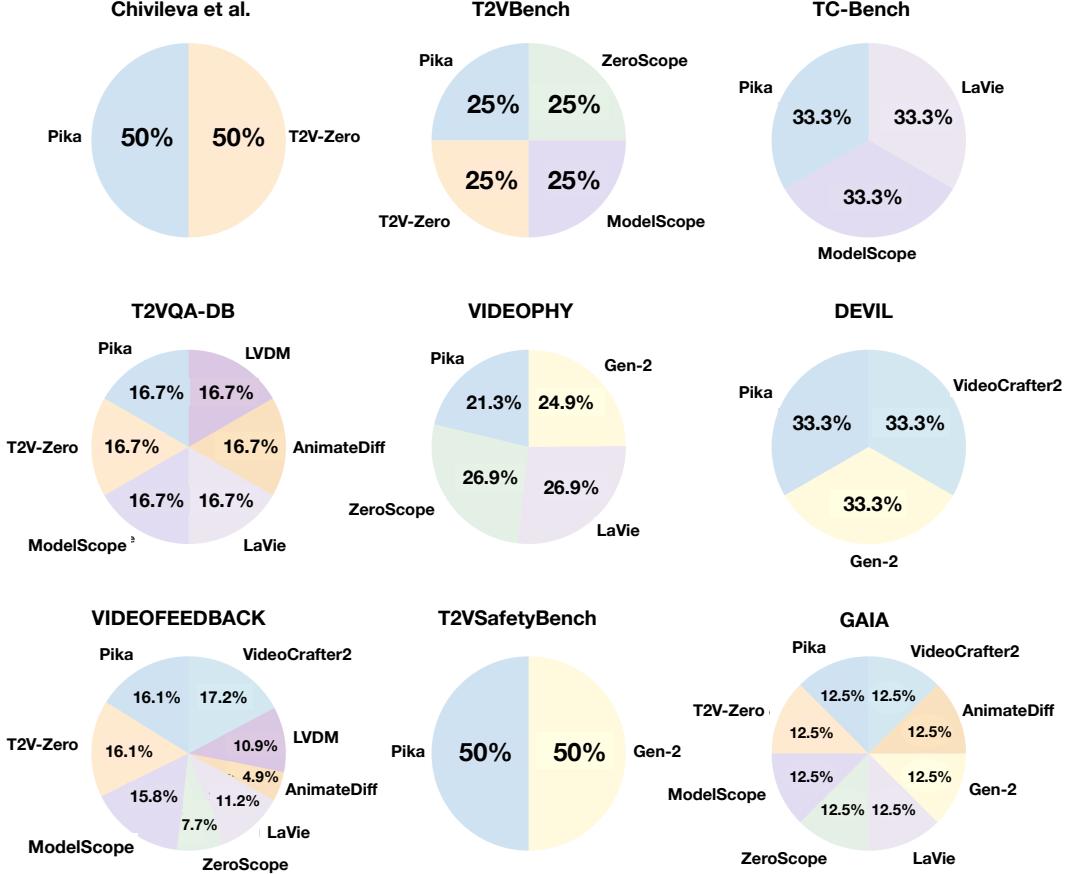


Figure 5: The Proportion of Videos Generated by Each Text-to-video Model.

framework employs a multi-dimensional approach, covering aspects such as video quality, temporal quality, subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, imaging quality, object class, human action, color, spatial relationship, scene, appearance style, temporal style, and overall video-text consistency.

Human preference annotations are used to validate these metrics, ensuring they align with both human perceptions and instructions of video quality. VBench’s multi-dimensional system provides detailed feedback on the strengths and weaknesses of video generation models, offering valuable insights for improving training data, model architecture, and evaluation methods.

**Chivileva et al. [37]:** The dataset consists of 1,005 videos generated from five recent text-to-video (T2V) models, with each video containing more than 8 frames at a resolution greater than 128p. This dataset is evaluated using several commonly used video quality metrics, including Inception Score (IS) [204], Frechet Video Distance (FVD) [239], and CLIPSim [262]. IS [204] measures image quality and diversity using a class probability distribution. FVD [239] compares the feature activations of real and generated videos in a pre-trained video classifier’s feature space, with lower scores indicating better quality. CLIPSim [262] evaluates the semantic alignment between the initial text prompt and the generated video by computing the average frame score using the CLIP model [186]. Additionally, the dataset includes extensive human evaluations to assess video naturalness and alignment with the text prompt. Human evaluators rate the videos on a scale of 1 to 10 for alignment (compatibility with the prompt) and perception (overall perceptual quality), resulting in a comprehensive evaluation of T2V model outputs. This work highlights the limitations of existing metrics and underscores the importance of human assessment in evaluating video naturalness and semantic matching.

**T2VQA-DB [106]:** This dataset is the largest-scale Text-to-Video Quality Assessment Database to date, comprising 10,000 videos generated by nine different T2V models using 1,000 text prompts. The

dataset features videos with 16 frames each at a resolution of 512p. Models used include Text2Video-Zero [96], AnimateDiff [64], Tune-a-Video [267], VidRD [3], VideoFusion [149], ModelScope [228], LVDM [68], Show-1 [287], and LaVie [253]. Each video is evaluated with a Mean Opinion Score (MOS) obtained from 27 subjects, who assess both text-video alignment and video fidelity. The subjects score the videos on a scale of 0 to 100, evaluating how well the generated video content matches the text description and the overall perceptual quality of the video, considering factors like distortion, saturation, motion consistency, and content rationality. The MOS scores are then normalized to account for inter-subject scoring differences, resulting in Z-score MOS (MOSz). The T2VQA-DB dataset is designed to facilitate the development of more accurate and comprehensive metrics for evaluating the quality of text-generated videos, reflecting real user preferences.

**T2VBench** [86]: T2VBench is a comprehensive benchmark specifically designed to evaluate the temporal dynamics of text-to-video generation models. The dataset comprises 5,000 videos generated using three leading T2V models: ModelScope [228], ZeroScope [4], and Pika [2], with each video containing more than 16 frames at resolutions greater than 256p. The benchmark employs 1,680 carefully crafted prompts enriched with temporal dynamics lexicons derived from Wikipedia, covering 16 critical temporal evaluation dimensions. These dimensions include aspects such as explicit and implicit event sequences, scene transitions, event timing, camera perspective transitions, direction of movement, emotional changes, shape changes, weather pattern changes, age changes, acceleration, and lighting and shadows. Each video is evaluated using human ratings collected on a Likert scale, providing a comprehensive assessment of the model’s ability to handle complex temporal dynamics. This benchmark not only highlights the strengths and limitations of current T2V models but also offers valuable insights into improving the temporal consistency and overall quality of video generation in future models.

**VIDEOPHY** [15]: VIDEOPHY is a benchmark designed to evaluate whether videos generated by text-to-video (T2V) models adhere to physical commonsense in real-world activities. The dataset includes 9,300 videos generated using 9 diverse T2V models, such as Pika [2], Lumiere [16], and VideoCrafter2 [31], conditioned on 688 high-quality, human-verified captions that depict interactions between various states of matter, including solid-solid, solid-fluid, and fluid-fluid interactions. Each video is assessed based on semantic adherence (whether the video accurately depicts the actions and entities described in the text) and physical commonsense (whether the depicted actions follow the laws of physics). Human evaluations reveal that the current models struggle significantly with both aspects, with the best-performing model, Pika [2], achieving accurate semantic adherence and physical commonsense in only 19.7% of the cases. This benchmark highlights the gap in current T2V models’ ability to simulate the physical world realistically and provides a crucial resource for developing more accurate and physically plausible video generation models.

**TC-Bench** [54]: TC-Bench is a benchmark specifically designed to assess the temporal compositionality of video generation models, both text-to-video and image-to-video. The benchmark evaluates three types of compositional changes: attribute transitions (e.g., a chameleon changing color), object relation changes (e.g., a person passing an object from one hand to another), and background shifts (e.g., a cityscape transitioning from day to night). TC-Bench includes both text prompts and corresponding ground-truth videos, allowing for evaluation of the model’s ability to generate seamless transitions over time. Two new metrics, Transition Completion Ratio (TCR) and TC-Score, are introduced to measure the extent to which generated videos align with the described transitions and maintain consistency. The benchmark reveals that most current models achieve less than 20% success in accurately completing the compositional changes, highlighting significant challenges and opportunities for improvement in temporal video generation.

**T2VSafetyBench** [157]: T2VSafetyBench is a benchmark specifically designed to evaluate the safety of text-to-video (T2V) generative models. The dataset includes 1,600 videos generated by various T2V models using 400 malicious prompts that were meticulously crafted using large language models and jailbreaking prompt attacks. The prompts are designed to test 12 critical safety aspects, including pornography, violence, gore, discrimination, political sensitivity, and temporal risk, among others. Each generated video is assessed for safety using a combination of human reviews and automated evaluation via GPT-4 [171], which provides high correlation scores with human assessments. The evaluation results reveal that no single model excels across all safety dimensions, with different models showing varying strengths and weaknesses. T2VSafetyBench serves as a critical resource for identifying and mitigating the safety risks inherent in video generation, highlighting the need for ongoing improvements in safety protocols as the capabilities of T2V models continue to advance.

**VIDEOFEEDBACK** [67]: VIDEOFEEDBACK is a large-scale benchmark dataset designed to provide fine-grained human feedback for video generation. The dataset comprises 37,600 videos synthesized by 11 different text-to-video models, including Pika [2], Lavie [253], SVD [19], and Sora [173], among others. These videos were generated based on prompts solicited from the VidProM [250] dataset, which contains diverse and semantically rich text-to-video pairs. VIDEOFEEDBACK is annotated by human raters across five key evaluation dimensions: Visual Quality, Temporal Consistency, Dynamic Degree, Text-to-Video Alignment, and Factual Consistency. Each aspect is scored on a scale from 1 (bad) to 4 (perfect). The annotations are designed to assess not only the visual and technical quality of the videos but also their alignment with the provided text prompts and adherence to factual information.

**AIGC-VQA** [145]: The AIGC-VQA dataset serves as a comprehensive benchmark for assessing the quality of AI-generated content (AIGC) videos. The dataset consists of 10,000 videos generated by various state-of-the-art T2V models such as Pika [2], Lumiere [16], and VideoCrafter2 [31], each video containing 16 frames at a resolution of 512p. AIGC-VQA encompasses three main evaluation aspects: technical quality, aesthetic quality, and video-text alignment. Technical quality measures distortions and temporal consistency using metrics like Frechet Video Distance (FVD) [239] and CLIPSim [262]. Aesthetic quality evaluates visual appeal through factors like composition, colorfulness, and non-toxic content. Video-text alignment assesses the semantic match between the generated video and the text prompt using advanced vision-language models like BLIP [119]. The dataset is annotated with Mean Opinion Scores (MOS) by 20 human evaluators, ensuring a comprehensive assessment that aligns closely with human perception. This benchmark provides a critical resource for developing and refining AIGC models, offering a robust framework for multi-dimensional video quality evaluation.

**DEVIL** [126]: DEVIL is a benchmark designed to evaluate text-to-video (T2V) models from the perspective of dynamics, an essential dimension for measuring the visual vividness and adherence to text prompts. The dataset consists of 800 text prompts categorized into different dynamic grades and evaluated using videos generated by models such as GEN-2 [201], Pika [2], VideoCrafter2 [31], OpenSora [298], StreamingT2V [70], and FreeNoise-Lavie [184]. DEVIL introduces three key metrics: Dynamics Range, Dynamics Controllability, and Dynamics-based Quality. Dynamics Range measures the extent of variations in video content, Dynamics Controllability assesses the model's ability to manipulate video dynamics in response to text prompts, and Dynamics-based Quality evaluates the visual quality of videos with varying dynamics. The evaluation framework incorporates multiple temporal granularities, including inter-frame, inter-segment, and video-level dynamics scores, such as Optical Flow Strength, Structural Dynamics Score, Perceptual Dynamics, Patch-level Aperiodicity, Global Aperiodicity, Temporal Entropy, and Temporal Semantic Diversity. This comprehensive evaluation protocol demonstrates a Pearson correlation exceeding 90% with human ratings, highlighting its effectiveness in advancing the development of T2V generation models.

**GAIA** [36]: GAIA is a comprehensive dataset for evaluating the action quality of AI-generated videos. It includes 9,180 videos from 18 different text-to-video models, both from lab studies and commercial platforms. Each video covers a variety of whole-body, hand, and facial actions. The dataset was created with the involvement of 54 participants who conducted large-scale human evaluations to assess action quality from three perspectives: subject quality, action completeness, and action-scene interaction. GAIA provides quantifiable action state estimations based on human reasoning behavior. The evaluation process resulted in 971,244 human ratings, normalized using Z-score normalization. The main metrics used for evaluation include Spearman's Rank-order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC). GAIA demonstrates the value of multi-dimensional methods and highlights the poor correlation between traditional Action Quality Assessment (AQA) methods and human evaluations. It serves as a benchmark to reveal the strengths and weaknesses of various text-to-video models and aims to facilitate the development of accurate AQA methods for AI-generated videos.

### 3.2 Evaluation Methods for AI-Generated Video Evaluation

With the availability of diverse benchmark datasets that align with various evaluation aspects, research in AIGVE has seen rapid growth. This section categorizes and summarizes the advancements in AIGVE research.

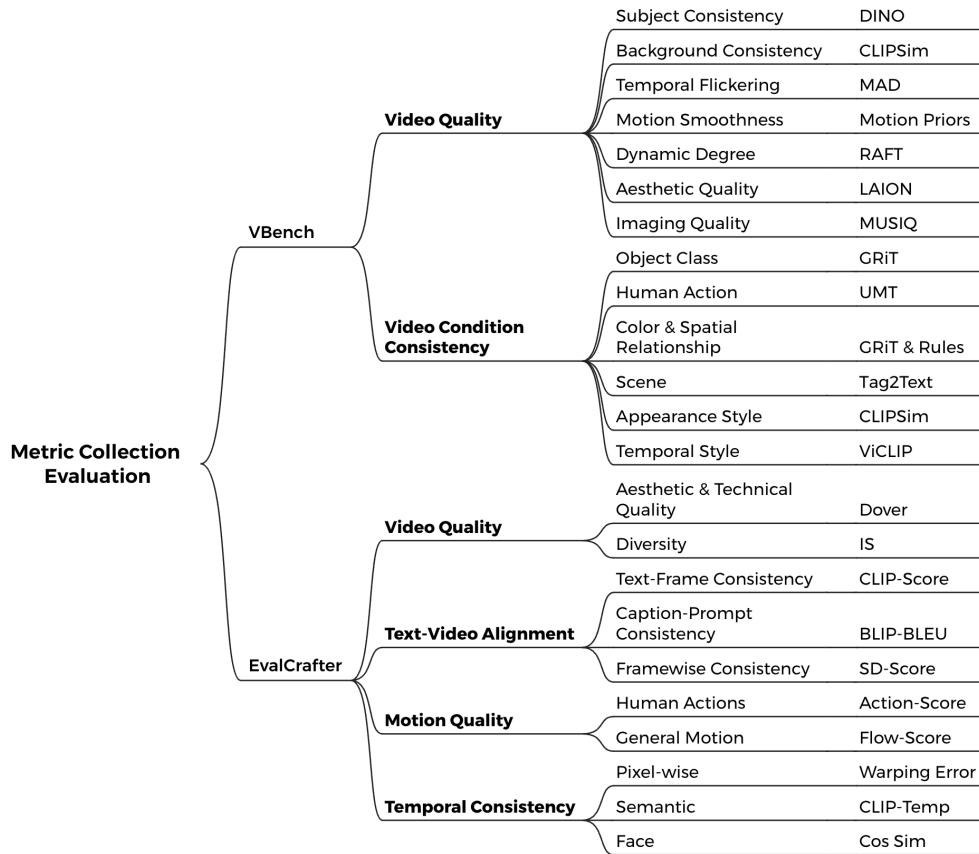


Figure 6: Summary of Metric Collection Evaluation.

### 3.2.1 Metric Collection Evaluation

In the early stages of AI-generated video research, as the demand for robust evaluation methods grew, researchers recognized the complexity of providing a comprehensive assessment. Instead of assigning a single score to summarize the quality of the generated video, they began to decompose the evaluation process into multiple aspects. Each aspect was assessed using established metrics, allowing for a more granular and accurate evaluation of AI-generated videos.

**VBench** [83]: VBench is a comprehensive benchmark suite that dissects video generation quality into 16 distinct, hierarchical dimensions. These dimensions are divided into two primary categories: Video Quality and Video Condition Consistency. The Video Quality category includes aspects such as Subject Consistency, measured by DINO [23] feature similarity across frames; Background Consistency, assessed by calculating CLIP [186] feature similarity; Temporal Flickering, evaluated using mean absolute difference across frames; Motion Smoothness, determined by motion priors in video frame interpolation models; Dynamic Degree, estimated by RAFT [229] to measure the extent of dynamics in synthesized videos; Aesthetic Quality, evaluated using the LAION [207] aesthetic predictor; and Imaging Quality, measured by the MUSIQ [94] image quality predictor.

The Video Condition Consistency category is decomposed into dimensions such as Object Class and Multiple Objects, where the success of generating specific objects and multiple objects in a frame is detected by GRiT [268]; Human Action, assessed by UMT [122] to determine if human actions match those described in the prompts; Color and Spatial Relationship, evaluated by GRiT [268] and rule-based approaches, respectively; Scene, where consistency with the scene described by the text prompt is checked using Tag2Text [81]; and Appearance Style and Temporal Style, measured by CLIP [186] feature similarity and ViCLIP [254], respectively. This multi-dimensional evaluation framework not only aligns well with human perceptions, as validated by human preference

annotations but also provides valuable insights into the strengths and weaknesses of different video generation models.

**EvalCrafter** [139]: EvalCrafter evaluates text-to-video models across four key aspects: visual quality, text-video alignment, motion quality, and temporal consistency. For visual quality, EvalCrafter employs metrics such as video quality assessment aesthetic and technical ratings, which measure common distortions like noise and artifacts, as well as the inception score (IS) [204] to evaluate the diversity of the generated content. For text-video alignment, the framework introduces a CLIP-Score [72] for consistency between text prompts and generated video frames, a BLIP-BLEU [120, 178] score to assess the alignment between generated captions and input prompts, and a novel SD-Score [200] that compares the generated quality with frame-wise results from stable diffusion models. Motion quality is assessed using Action-Score for human actions and Flow-Score for general motion, while temporal consistency is measured through metrics like warping error and semantic consistency across frames.

The aspects and metrics of these works are summarized in Figure 6. Both works establish a foundational framework for evaluating AI-generated video models, demonstrating strong correlations with human judgment. However, the need for multiple individual metrics in these frameworks presents challenges in integrating them into a unified pipeline. Additionally, evaluating a single video across a sequence of metrics is time-consuming. Thus, there remains a need for a streamlined, unified evaluation method.

### 3.2.2 Modeling Evaluation

Supported by the development of large-scale datasets, recent research trends have advanced toward leveraging modeling methods to comprehensively evaluate AI-generated videos in a manner that mimics human judgment. We observe two distinct branches emerging in parallel.

One branch focuses on the general evaluation of AI-generated videos which gives an overall evaluation of the video.

**VIDEOSCORE**[67]: VIDEOSCORE is a significant advancement in the high-level evaluation of AI-generated videos, leveraging the VIDEOFEEDBACK dataset. The model is built on the Mantis-Idefics2-8B [87] backbone and trained using regression scoring with a linear layer. It achieves a Spearman correlation of 77.1 on VideoFeedback-test.

**T2VQA** [105]: T2VQA is built on the T2VQA-DB. T2VQA employs a novel transformer-based architecture that integrates features from both text-video alignment and video fidelity. The model uses BLIP [120] for frame encoding and Swin-Transformer [142] for capturing video fidelity, with a large language model handling quality regression.

**AIGC-VQA** [145]: AIGC-VQA introduces a general perception metric for assessing the quality of AI-generated videos, targeting three key aspects: technical quality, aesthetic quality, and video-text alignment. The model employs a multi-branch architecture, with a 3D-Swin Transformer [277] handling technical quality, ConvNext [143] managing aesthetic evaluation, and a BLIP-based [119] branch, enhanced with a spatial-temporal adapter, assessing video-text alignment. The AIGC-VQA model is trained using a divide-and-conquer strategy, progressively optimizing each branch to ensure comprehensive video quality assessment.

**Qu et al.**[185] addresses the quality assessment of AI-generated videos by categorizing evaluation into three key dimensions: visual harmony, video-text consistency, and domain distribution gap. The method incorporates a multi-modal approach, utilizing explicit prompt injection and implicit text guidance to enhance video-text alignment. Additionally, the framework employs an auxiliary inter-domain classification task to predict the source generative model, which significantly improves the discriminative features and overall quality assessment performance.

Another branch focuses on evaluating specific aspects of the generated video while still considering alignment with both human perception and instruction.

**TC-Bench** [54]: TC-Bench evaluates the temporal compositionality of AI-generated videos. TC-Bench introduces two novel metrics, TCR and TC-Score, designed to measure the completion of compositional transitions and overall text-video alignment.

**Miao et al.** [157] focuses on evaluating the safety aspects of AI-generated videos, addressing concerns around potentially harmful content. Leveraging proposed T2VSafetyBench, it evaluates 12 critical safety dimensions, including pornography, violence, discrimination, misinformation, and temporal risks. The framework utilizes both automated assessments, primarily leveraging GPT-4, and manual evaluations to ensure a comprehensive analysis of these safety aspects. T2VSafetyBench revealed that no single model excels across all dimensions, with different models showing strengths in various areas, such as Stable Video Diffusion [19] performing well against sexual content, while Gen2 [201] excelled in managing gore and disturbing content. This benchmark underscores the need for a balanced trade-off between usability and safety in text-to-video generative models, emphasizing the importance of focusing on video safety as these technologies continue to advance.

**DEVIL** [126]: DEVIL framework emphasizes the evaluation of dynamics in AI-generated videos, a crucial aspect often overlooked by traditional metrics. The framework introduces a comprehensive evaluation protocol focusing on three key metrics: dynamics range, dynamics controllability, and dynamics-based quality. The framework establishes a benchmark with text prompts reflecting various dynamics grades, allowing for a detailed assessment of a model's ability to generate and control dynamic content.

### 3.3 Foundations of AI-Generated Video Evaluation

Although still emerging, AIGVE builds on two key foundations: alignment with human perception and alignment with human instructions. A solid understanding of these two foundational aspects is crucial for comprehensively developing AIGVE.

In the subsequent sections of this survey, we introduce the research works related to these two important foundation aspects and give further prospects based on the ground knowledge of these two aspects.

Section 4 discusses alignment with human perception. This section investigates a detailed survey on benchmark datasets as well as existing quality evaluation methods, offering a clear picture of how the field of evaluating video perception quality evolved over the last decade.

Section 5 covers alignment with human instructions, considering both related benchmark datasets and evaluation frameworks. This section summarizes the prior representative work in terms of the alignment between video content and human instructions to introduce the previous development trajectory of this field and how it has been influenced by the emergence of large language models.

Finally, Section 6 summarizes future perspectives in the field, identifying key challenges and potential research opportunities. We discuss the integration of vision language models for video evaluation, improvements in score interpretability, and the embedding of ethical and safety considerations within AIGVE frameworks.

## 4 Alignment with Human Perception

A key objective of AI-generated video is to produce content that closely aligns with human perception. Over the past decade, numerous benchmark datasets and evaluation techniques have been developed to assess AI-generated videos based on criteria related to human perception. In this section, we introduce several representative benchmark datasets, evaluation methods, and associated metrics.

### 4.1 Benchmark Datasets

A number of datasets have been curated for the development and validation of video quality assessment models. This section details various collections of video sequences with different characteristics and their corresponding quality scores. These videos vary in resolution, duration, distortions, and the environments in which they were assessed, as well as in the methodologies used for assessing video quality, including crowdsourcing platforms, controlled lab environments, and various subjective quality metrics. The evaluation metrics employed across these datasets include Mean Opinion Score (MOS) [220], Differential Mean Opinion Score (DMOS) [211], Peak Signal-to-Noise Ratio (PSNR) [209], Structural Similarity Index (SSIM) [258], and no-reference metrics such as BRISQUE [164] and NIQE [166], providing a comprehensive assessment of video quality from multiple perspectives.

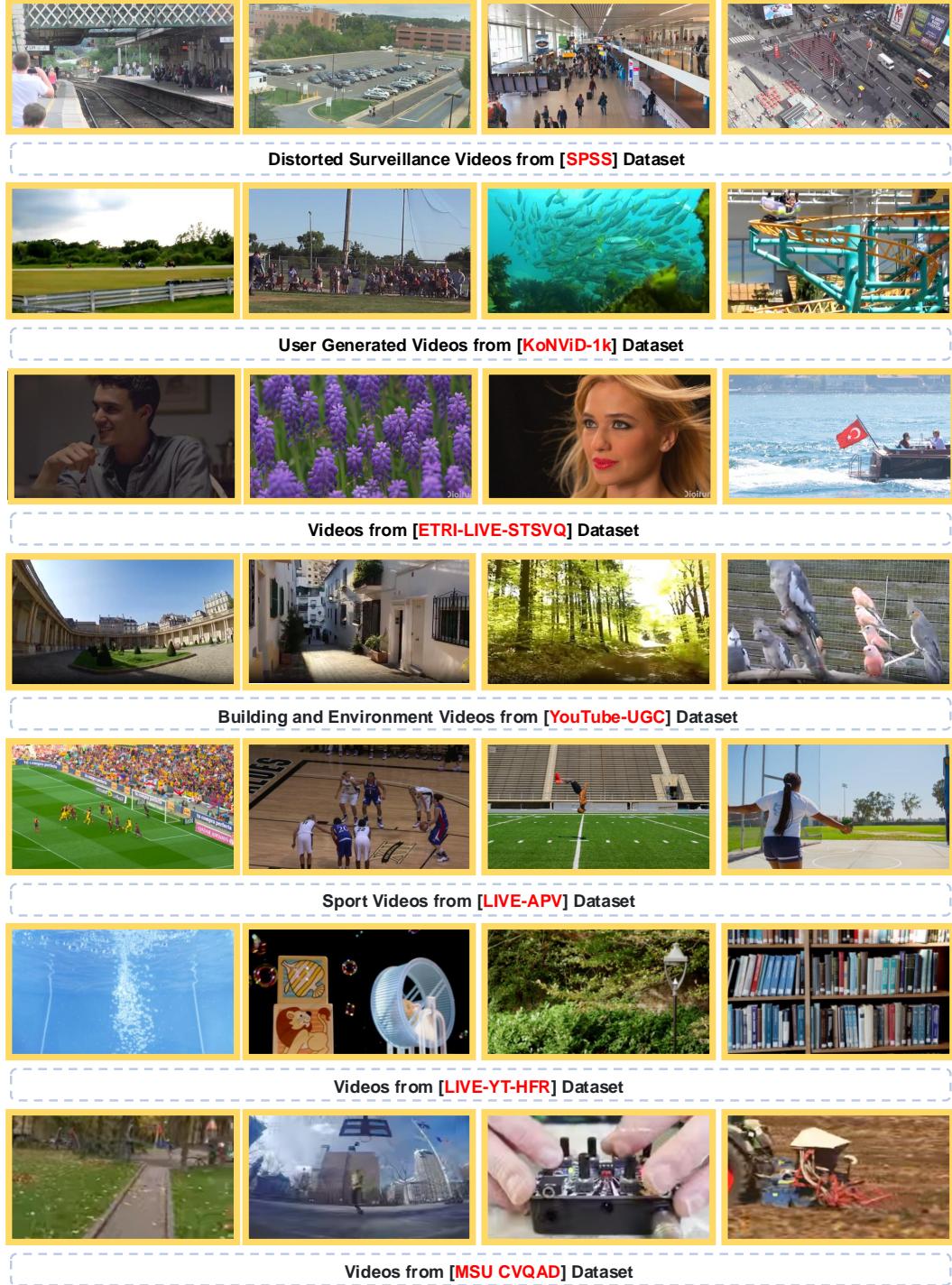


Figure 7: Exemplar Cases from Alignment with Human Perception Benchmark Datasets. Here we chose four representative frames from different video clips as a representation of the video within each dataset.

Table 2: Summary of Databases for Video Quality Assessment. 'Cont.' denotes the number of unique video contents. '#Total' represents the total number of videos. 'Dur.' indicates the video duration in seconds. '#Subj.' refers to the number of subjects involved. 'Env.' specifies the environment for the subjects. 'Crowd' stands for crowdsourcing.

Database	Year	#Cont.	#Total	Resolution	Dur.	Distortions	#Subj.	Env.	Link
CVD2014 [170]	2014	5	234	480p, 720p	10-25	In-capture	210	In-Lab	
KoNViD-1k [78]	2017	1200	1200	540p	8	In-the-wild	642	Crowd	
LIVE-VQC [214]	2018	585	585	1080p, 240p	10	In-the-wild	4776	Crowd	
YouTube-UGC [255]	2019	1500	1500	4K, 360p	20	In-the-wild	>8000	Crowd	
SPSS [17]	2020	14	224	1080p	N/A	In-the-wild	19	Crowd	
UGC-VIDEO [125]	2020	50	550	720p	10	UGC+compression	30	In-lab	
ETRI-LIVE-STSVQ [113]	2021	15	437	4K	5-7	In-the-wild	34	In-lab	
LIVE-APV [208]	2021	33	315	1080p, 4K	7	In-the-wild	40	In-lab	
LIVE-YT-HFR [151]	2021	16	480	4K	10	In-the-wild	85	In-lab	
LIVE-LSVQ [280]	2022	39075	39075	1080p	5-12	In-the-wild	6284	Crowd	
MSU CVQAD [10]	2022	2500	2486	360p-1080p	10, 15	Compression	10800	Crowd	
M-VCM [162]	2023	10	1628	1080p	6	In-capture	N/A	Crowd	

A summary of the benchmark datasets discussed in this section is presented in Table 2 and Figure 7 illustrates exemplar cases from several representative datasets.

**CVD2014** [170]: The CVD2014 database contains 234 video sequences with in-capture distortions across 5 different scenes. The videos come in resolutions of 720p or 480p and feature frame rates that range from 9 to 30 fps. The lengths of these videos vary between 10 and 25 seconds, and they are stored in AVI format. The authors recruited 210 participants to assess video quality, 158 of whom were female, with ages ranging from 18 to 46 years and an average age of 24. Vision tests were administered using EDTRS for acuity, F.A.C.T. for contrast, and the Farnsworth D-15 for color discrimination. The trials included a variety of video presentations, explained through example videos to mitigate central bias in scoring. The experimental setup was conducted in a dark, controlled environment featuring a 24-inch calibrated display. Participants were prepped with a demonstration of video quality before using the VQone MATLAB toolbox for the tests. The observers maintained an 80 cm viewing distance, ensured by a counterweight system. Each session lasted approximately 66 minutes, including vision testing, briefing, and training phases. Videos were shown sequentially in random order, and participants rated them using graphical sliders, with results compiled as Mean Opinion Scores (MOS).

**KoNViD-1k** [78]: KoNViD-1k is a comprehensive video quality database consisting of 1200 unique video sequences. These videos, selected from the YFCC100m dataset (Flickr), exhibit a variety of authentic distortions. Each video was clipped and resized to 540p in a landscape layout. The frame rates are either 24, 25, or 30 fps, and each video has a duration of 8 seconds. The videos are in MP4 format. The authors utilized the CrowdFlower platform for crowdsourcing subjective video quality assessments. Each participant was briefed on various video degradations and evaluated videos based on a displayed quality scale. To enhance engagement and reliability, they used 'gold standard' test questions derived from a subset of 100 videos. This process ensured higher consistency among worker responses. A 70% accuracy threshold was mandated. Overall, 642 participants from 64 countries provided a total of 136,800 ratings, with each video receiving an average of 114 assessments. The 95% confidence interval on the MOS scale did not exceed 0.5.

**LIVE-VQC** [214]: The LIVE-VQC database features 585 unique video sequences that capture a range of authentic distortions, including camera motion and night scenes. These videos have varying resolutions from 240p to 1080p and include both landscape and portrait formats. The frame rates are 20, 24, 25, and 30 fps, and each video is 10 seconds long. Stored in MP4 format, the videos were subjectively evaluated using Amazon Mechanical Turk. In this study, 4,776 participants, aged between 11 and 65 years and almost equally divided by gender, provided over 205,000 subjective evaluations of 585 videos, which translates to approximately 240 votes per video. The majority of the participants, from the U.S. and India, viewed the videos on displays that supported at least a 720p

resolution. Chosen for their high reliability from previous tasks, they used only non-mobile devices with specific browser requirements to ensure consistent performance. The study was designed with stringent procedures to ensure reliable and consistent assessments of video quality. These procedures included a detailed introduction, eligibility checks, comprehensive training, and testing phases with ongoing feedback. Additionally, the study employed strategies to discard any unreliable data from participants experiencing playback issues. It concluded with a survey to collect demographic data and information about viewing conditions. The gathered MOS values are included in the database. The methods used in this study allowed for the analysis to encompass a wide range of conditions, thus providing a robust evaluation of video quality across a diverse participant pool.

**YouTube-UGC** [255]: The YouTube-UGC database is composed of 1500 video sequences with a wide range of authentic distortions. These videos were sampled from YouTube and cover various content types, including HDR, screen content, animation, and gaming videos. The resolutions span from 360p to 4K, and frame rates include 15, 20, 24, 25, 30, 50, and 60 fps. Each video is 20 seconds long and is stored in MKV format. The subjective quality assessment was performed via crowdsourcing on Amazon Mechanical Turk, with participation from over 8000 subjects, yielding 170159 ratings, roughly 123 per video. The database provides MOS and standard deviation values.

**SPSS** [17]: This dataset comprises 14 reference videos in a resolution of 1080p and 224 videos with various distortions. The reference videos cover typical surveillance scenarios such as crowded streets, transport hubs, parking areas, and stadiums. The applied distortions are noise, uneven illumination, blur, and smoke. Noise was simulated using an additive white Gaussian noise model; uneven illumination was created using a grayscale circular fading mask; blur was generated with a motion filter to simulate movement; and smoke was blended using video editing software. Each distortion was applied at four different severity levels. The authors employed a pairwise-comparison method for subjective testing, where observers were shown pairs of distorted videos from the same category and distortion type, but with varying severity levels, to assess each. Observers repeated this for all possible video pairs, resulting in 6 pairwise comparisons per reference video. Participants could score videos equally if they appeared similar. Preferred videos scored one point, while equal assessments received 0.5 points per video. Each observer viewed the videos once, with the option to rewatch. Scores from all observers were then aggregated to calculate MOS for each video by averaging the scores over the number of observers. The authors utilized two benchmark metrics for natural images and videos—PSNR and SSIM, along with a full-reference metric, VIF. Given the absence of a ground truth in video surveillance, they also included two no-reference metrics, BRISQUE and NIQE, averaging their values across all video frames. Nineteen observers conducted these evaluations, with initial outlier detection based on non-transitivity. The authors then aggregated the remaining data to derive subjective scores and tested correlation with objective metrics using Spearman Rank Order and Pearson Linear Correlation Coefficients after nonlinear regression. They found that the metrics did not consistently correlate with subjective assessments, particularly in a video-surveillance context, which suggested a need for developing more specialized objective metrics.

**UGC-VIDEO** [125]: This dataset comprises 50 source videos from TikTok, representing diverse user-generated content categories such as selfie, indoor, outdoor, and screen content. 400 videos were randomly selected, each with a resolution of 720p, a frame rate of 30 FPS, and longer than 10 seconds. They were transcoded using H.264/AVC and H.265/HEVC at five different quantization levels (QPs: 22, 27, 32, 37, 42), resulting in 550 video sequences in total. In this study, 28 participants were recruited and followed the subject screening procedures outlined in ITU-R BT 500.13. Participants evaluated the quality of each video using a five-point scale. To reduce viewer fatigue, the study was divided into three sessions, each lasting approximately 30 minutes. Each session included the presentation of 16 or 17 original videos and their respective transcoded versions, shown in a random sequence. Additionally, each session began with 10 preliminary 'dummy presentations' of varying quality levels to calibrate the participants' responses; these initial evaluations were excluded from the final analysis. Along with the MOS, the Differential Mean Opinion Score (DMOS) was calculated to measure the quality difference between the original and transcoded videos.

**ETRI-LIVE-STSVQ** [113]: This dataset comprises 437 videos derived from 15 unique 4K, 10-bit sources subjected to various levels of space-time subsampling and compression. The sources, including selections from the Ultra Video Group, Harmonic 4K footage, and Netflix's public library, were adapted to specific video formats through cropping and chroma subsampling. Videos were processed to include spatial resolutions ranging from  $3840 \times 2160$  down to  $960 \times 540$ , using the Lanczos kernel for both downsampling and upsampling to preserve the original resolution. Temporal

subsampling involved reducing frame rates from 120 or 60 fps by dropping alternate frames and employing Linear Filter Interpolation (LFI) to ensure smooth transitions and minimize artifacts. Additionally, videos underwent compression using the HEVC Main 10 profile with the FFmpeg libx265 encoder, with compression levels adjusted through the Quantization Parameter (QP) to match the predetermined target bit rates, ensuring a range of perceptual quality levels. The videos were displayed on setups supporting 3840×2160 resolution at 60/120 fps, which helped explore optimal video coding practices under varied conditions. Thirty-four participants evaluated these videos using a Single-Stimulus Continuous Quality Evaluation (SSCQE) method with hidden references. Approximately 15,000 subjective quality evaluations were collected to analyze rate-distortion impacts. Each video, approximately 5.61 seconds in length, was rated once on a continuous scale during three 30-minute sessions spread over three days to avoid viewer fatigue. Each session included between 145 and 146 videos and 15 hidden references, with the ratings transformed into differential DMOS. To ensure unbiased video quality evaluations and minimize memory effects, several randomization strategies were implemented. The playlist was initially shuffled repeatedly to prevent any video content from appearing consecutively more than ten times. The videos were then organized into 30 groups, with each participant viewing ten groups per session. A round-robin ordering was employed to ensure each video group was presented equally across all sessions and participants, avoiding repetitive combinations. Finally, the playlist, including the undistorted reference videos and video groups, was reshuffled before each session, ensuring significant separation between videos of the same content.

**LIVE-APV** [208]: This dataset includes 315 video clips extracted from 45 source sequences, originating from 33 uncompressed, high-quality videos across 10 distinct sports categories. These source videos are available in either 1080p or 4k resolutions, are progressively scanned in the YUV 4:2:0 format, and have had their audio components stripped away. Each video is played at a consistent rate of 30 fps. For analytical precision, longer videos were manually divided into shorter segments, each roughly 7 seconds long, with careful attention to avoid overlapping or closely successive segments. Six altered versions of each original sequence were created by applying various distortion techniques including H.264 Compression, Aliasing, Judder, Flicker, Frame Drops, and Interlacing. Compression artifacts were implemented using H.264 encoding with varying Constant Rate Factor (CRF) values. Aliasing effects were created through a process of downscaling followed by upscaling without the use of anti-aliasing filters. Motion judder was simulated using a 2:3 pulldown technique, which can introduce irregular motion. Flicker was generated by alternating quantization parameters (QP) between frames. Frame loss was replicated by strategically removing clusters of frames from the video sequence. Lastly, interlacing artifacts were produced by separating even and odd lines to create interlaced frames. To ensure a wide spectrum of perceptual qualities, each distortion type was applied at multiple intensity levels, allowing for a comprehensive study of how different distortions impact video quality across various severity ranges. This approach generated a diverse array of video quality conditions for further evaluation. Subjective assessments were performed using the assembled database, collecting over 12,000 evaluations from 40 participants. This evaluation took place in the LIVE Subjective study room, where each video was rated on a numerical scale from 0 (lowest quality) to 100 (highest quality).

**LIVE-YT-HFR** [151]: This dataset encompasses 480 videos, modified to 6 frame rates and 5 compression levels to study the effects of both factors. It originates from 16 natural scene videos at 120 fps, 11 of which come from the BVI-HFR database in 4K resolution and have been downsampled to HD for public use. Additionally, five high-motion sports sequences were captured by the Fox Media Group in 4K. Each original video was transformed into 30 test sequences across frame rates ranging from 24 to 120 fps, using VP9 compression. The videos vary in resolution from 1080p to 4K, aligning with the shift toward 4K in streaming services. Quality assessment was uniform across different resolutions. Three descriptors were measured for each video: Spatial Information (SI), Temporal Information (TI), and Colorfulness (CF). SI was assessed using a Sobel magnitude, and TI was calculated from luminance differences between frames. The authors conducted a subjective evaluation involving 85 undergraduate volunteers, evenly split between 14 female and 71 male participants aged 20-30, all with normal or corrected-to-normal color vision. They participated in a study generating 19,000 quality ratings across 240 videos. Each participant was trained with 6 non-database videos to calibrate their understanding of expected video quality and instructed not to rate based on content. Ratings were collected over sessions that did not exceed 40 minutes, using the Single-Stimulus Continuous Quality Evaluation (SSCQE) method and a 5-point Likert scale via a Palette gear console. Each video received a Mean Opinion Score (MOS) based on at least 42 ratings.

Although not every video was rated by each participant, anchor videos common to all sets showed consistent MOS.

**LIVE-LSVQ** [280]: This dataset features a diverse collection of 38,811 real-world videos, varying in size, content, and distortion levels. It also includes 116,433 space-time video patches, known as ‘v-patches’, and 5.5 million human perceptual quality annotations provided by 6,284 participants. The source material was gathered from two major public user-generated content (UGC) video repositories: the Internet Archive and YFCC-100M. Each video was processed to have an average duration of seven seconds using ffmpeg. Unlike KoNViD-1k, this collection does not impose restrictions on video resolution or aspect ratio, making it more representative of real-world content. The researchers avoided applying scaling or additional processing that might alter the original quality. To ensure the videos closely resembled UGC, they employed a mixed integer programming method to match UGC feature histograms, considering 26 holistic spatial and temporal characteristics. To investigate the relationship between global and local spatio-temporal qualities, the researchers developed three distinct types of video patches (v-patches) from each video. The first type, known as the spatial v-patch (sv-patch), preserves the original temporal duration of the video but reduces its spatial dimensions to 40%. The second type, the temporal v-patch (tv-patch), maintains the original spatial dimensions but cuts the temporal length to 40%. The third type, the spatio-temporal v-patch (stv-patch), is reduced to 40% of the original dimensions across both spatial and temporal aspects. All v-patches maintain the aspect ratio of their original videos. Although each v-patch originates from the same source video, the overlap in volume between sv-patches or tv-patches and their corresponding stv-patches is limited to 25%. The authors employed Amazon Mechanical Turk (AMT) for human evaluations of both the original videos and the different types of v-patches. They conducted two separate AMT tasks: one for the full-length videos and another for the three types of v-patches. On average, each video and v-patch garnered 35 ratings from participants, providing a substantial dataset for analysis.

**MSU CVQAD** [10]: This dataset comprises approximately 2,500 video streams that have been encoded with five different compression protocols: H.264/AVC, H.265/HEVC, H.266/VVC, AV1, and VP9. The dataset includes 1,022 compressed streams across two encoding settings: 1 fps and 30 fps. Each video was encoded at three designated bitrates—1,000 kbps, 2,000 kbps, and 4,000 kbps—employing Variable Bit Rate (VBR) mode where applicable, or using specific QP/CRF values to achieve these bitrates. The researchers utilized the Subjectify.us crowdsourcing platform to gather subjective scores, amassing 766,362 valid responses from almost 11,000 participants, with a minimum of 10 responses per video pair. Specifically, the dataset was segmented into five subsets based on source videos, leading to five distinct comparison groups. Each subset comprised a collection of source videos and their compressed versions. Comparisons involved evaluating all possible pairs of compressed videos derived from a single source video, ensuring that each pair consisted only of videos from the same source. The original source videos were also included in the comparisons. Participants viewed each video pair sequentially and were asked to select the video with superior visual quality or state if the quality was identical. They had the option to replay the videos. Each participant compared 12 pairs, including two pairs with an evidently higher-quality option, which served as verification questions. Responses from participants who failed these verification questions were excluded. The Bradley-Terry model was employed to convert pairwise voting results into a score for each video. Applying the model to the pairwise ranking data yielded consistent subjective scores for each group of videos compressed from the same reference video.

**M-VCM** [162]: The dataset originated from 1628 Microsoft Teams calls conducted across 83 varied network environments, utilizing 10 source videos. Among these, eight videos featured an individual speaking directly to the camera, and the remaining two recorded conversations between two individuals. The resolution of all source videos was maintained at 1080p with a frame rate of 30 fps. The experiment simulated calls between two systems under a range of network conditions, including constant and variable bandwidths, burst losses, and cross traffic, among others. These conditions necessitated adjustments in video bitrate and resolution, influencing the overall video quality through changes in bitrate, shifts in resolution, fluctuations in frame rate, and occurrences of frozen or missing frames. Dual QR codes, positioned at the top-left and bottom-right corners of the source videos and encoding the original frame index, were employed to aid frame alignment in live video calls with the reference video. The degraded videos, recorded at 30 fps, were derived from segments of calls lasting from 6 seconds to 2 minutes. They were evaluated for quality using the Absolute Category Rating (ACR) scale as specified in ITU-T Recommendation P.910. Each video

clip, on average, received 17 valid evaluations through a crowdsourced approach employing the P.910 Toolkit.

## 4.2 Image Quality Assessment

Evaluating AI-generated videos presents a significant challenge due to the absence of reference content for comparison or guidance. To address this, no-reference quality assessment methods have been developed, allowing for the evaluation of video quality independently without relying on reference data. These techniques were initially introduced and applied in the domain of AI-generated Image Quality Assessment (IQA), and many of these early No-Reference (NR) IQA methods served as the foundation for subsequent approaches in AI-generated video quality assessment. In this context, we first review several key NR-IQA techniques that have laid the foundation for developing early-stage AI-generated video quality assessment algorithms.

Early IQA methods were primarily developed using statistical approaches [163, 165]. These methods laid the foundation for evaluating image quality by relying on statistical features extracted from natural images. With the introduction of deep learning, IQA techniques have seen substantial improvements, particularly in handling complex, real-world distortions [276]. More recently, the incorporation of advanced architectures, such as Vision Transformers (ViT) [51] and Swin-Transformers [142], has enabled the development of several methods that achieve competitive performance in the field. In the remainder of this section, we highlight several representative AI-generated IQA methods that have emerged from these advancements.

**WaDIQaM**[21]: WaDIQaM proposed several pioneer model deep neural network architectures for IQA. Characterized by its depth and versatility, it comprises 10 convolutional layers and 5 pooling layers for effective feature extraction, along with 2 fully connected layers for regression tasks. This architecture allows for end-to-end training, enabling the model to learn directly from raw input data without the need for hand-crafted features or prior domain knowledge about the human visual system. A unique aspect of the architecture is its adaptability for both no-reference and full-reference IQA settings, facilitating joint learning of local quality and local weights, which represent the relative importance of local quality to the global quality estimate. The model employs a patch-based approach, where features are extracted from image patches, and local quality estimates are aggregated to derive a global quality score. This method enhances the model's robustness against various distortions, including luminance and contrast changes, and allows for the effective pooling of local patch qualities.

**RankIQA**[138]: The RankIQA approach to NR-IQA uses ranked image datasets generated from synthetic distortions, which allows for effective training of deep learning models without the need for extensive human-annotated data. By employing a Siamese Network architecture, RankIQA learns to rank images based on their quality by leveraging the relative quality information derived from pairs of images, rather than relying on absolute quality scores. This method addresses the critical challenge of limited labeled datasets in the field, enabling the training of deeper and more complex networks that can capture intricate features relevant to image quality. Additionally, the authors introduce an efficient backpropagation technique that optimally utilizes computational resources by considering all possible pairs within a mini-batch, significantly enhancing training speed and performance. The experimental results demonstrate that RankIQA outperforms existing NR-IQA methods, showcasing its effectiveness in correlating with human judgments of image quality.

**dipIQ** [150]: dipIQ develops an Opinion-Unaware Blind Image Quality Assessment (OU-BIQA) model that can predict the quality of digital images without requiring access to their original pristine counterparts. Traditional BIQA models rely heavily on subjective testing to collect ground truth data, which is often slow, cumbersome, and expensive, leading to limited training datasets that may not adequately represent the vast image space. In contrast, the dipIQ approach leverages large-scale databases to automatically generate a substantial amount of reliable training data in the form of quality-discriminable image pairs (DIPs). Each DIP is associated with a perceptual uncertainty measure, allowing the model to learn from these pairs using a pairwise learning-to-rank algorithm, specifically RankNet. This innovative method enables the dipIQ model to achieve higher accuracy and improved robustness across various image content types, outperforming existing OU-BIQA models. The research also explores extending the framework to a listwise learning-to-rank approach, resulting in the diliQ index, which further enhances performance. Overall, dipIQ aims to address the challenges of traditional BIQA by providing a more efficient and effective means of assessing image quality objectively.

**MetaIQA**[302]: MetaIQA improves the evaluation of image quality by enabling models to learn shared prior knowledge from various distortion-specific tasks, allowing for quick adaptation to unknown distortions. Traditional methods often rely on pre-trained networks that are not specifically designed for IQA, leading to generalization issues when faced with different types of distortions. These methods typically require large amounts of annotated data, which is challenging to obtain for IQA tasks. In contrast, the proposed MetaIQA approach utilizes a bi-level gradient descent strategy to learn a meta-model from multiple NR-IQA tasks, capturing the intricate relationships between image data and human-perceived quality. This meta-model can generalize across different distortions, enabling the system to quickly adapt to new, unseen distortions with minimal training data. By leveraging the ability of humans to learn from limited examples, deep meta-learning allows for a more efficient and effective evaluation of image quality, ultimately outperforming traditional methods in both synthetic and authentic distortion scenarios. This capability is particularly valuable in real-world applications where images may undergo various types of distortions throughout their lifecycle.

**TReS**[61]: TReS integrates Convolutional Neural Networks (CNNs) with Transformers to leverage both local and global features of images, enhancing the model’s ability to capture complex perceptual qualities. Secondly, the model introduces a novel self-consistency loss mechanism that reinforces representation learning without requiring additional labels or external supervision, allowing the network to learn more robust features through transformations like flipping and rotation. This approach improves the model’s generalization capabilities across different datasets. Additionally, the authors employ a relative ranking strategy to better align the model’s predictions with human subjective assessments, thereby increasing the correlation between objective and subjective quality scores. The model is extensively evaluated on multiple publicly available datasets, particularly on challenging datasets like FBLIVE. Overall, the combination of these innovative techniques results in a powerful NR-IQA model that effectively assesses image quality in a way that closely mirrors human perception, addressing limitations found in existing methods and setting a new benchmark in the field.

**RKIQT**[123]: RKIQT significantly advances NR-IQA by introducing a novel approach that learns reference information during training, thus eliminating the need for pristine reference images during inference. Central to this framework is the Masked Quality-Contrastive Distillation (MCD) method, which enables the student model to acquire comparative knowledge from a non-aligned reference teacher network, enhancing its robustness and representation capacity. Additionally, the framework incorporates inductive bias regularization to facilitate fast convergence and mitigate overfitting, allowing the student model to fine-tune its quality-aware abilities by leveraging insights from various teacher networks. Extensive experiments conducted on eight benchmark IQA datasets validate the effectiveness and efficiency of RKIQT. Furthermore, the framework improves the feature extraction process, enabling it to convey richer quality information while utilizing less input than existing models.

**CrossScore**[260]: CrossScore distinguishes itself from traditional IQA methods primarily through its innovative use of multiple unregistered reference images captured from different viewpoints, allowing for evaluation without the need for ground truth images. This is particularly advantageous in scenarios like Novel View Synthesis (NVS), where direct reference images are often unavailable. Unlike traditional full-reference metrics, such as SSIM[257], which require a single ground truth image for comparison, CrossScore employs a neural network with a cross-attention mechanism that facilitates detailed per-pixel evaluation by focusing on relevant features across multiple views. This approach not only enhances the accuracy of the assessment but also aligns more closely with human visual perception. Additionally, while traditional IQA methods are typically limited to single-view comparisons, CrossScore is specifically designed for multi-view scenarios, making it more versatile for various applications. Overall, CrossScore addresses the limitations of existing IQA frameworks by providing a more flexible, accurate, and perceptually relevant evaluation method.

**DepictQA-Wild**[282]: DepictQA-Wild leverages Vision Language Models to enhance the evaluation of image quality through linguistic descriptions that align with human perception. The method is structured around a multi-functional task paradigm that encompasses both single-image assessment and paired-image comparison, allowing for a comprehensive evaluation of image quality across various scenarios. A significant contribution is the construction of a large-scale dataset, DQ-495K, which includes 35 types of distortions at five levels, ensuring diverse and high-quality training data. The model retains the original image resolution during training, which is crucial for accurately

perceiving resolution-related quality issues. Additionally, DepictQA-Wild incorporates a confidence estimation mechanism that filters low-quality responses, enhancing the reliability of the assessments.

### 4.3 Video Quality Assessment

Video Quality Assessment (VQA) algorithms aim to predict video quality in alignment with human perception. The rapid expansion of social media platforms, such as YouTube, Facebook, and TikTok, has driven a surge in no-reference user-generated video content. While professional-generated content has received less attention, possibly due to copyright concerns, much of the recent research in VQA has focused on no-reference user-generated content. For no-reference (NR) VQA, a basic approach involves assessing the quality of each frame using NR-IQA methods and aggregating the results to produce an overall video quality score. However, compared to NR-IQA, NR-VQA must account for temporal distortions, which adds complexity by requiring an understanding of time-dependent quality degradations.

Early NR-VQA algorithms are often designed to address specific types of distortions, such as those caused by transmission or compression artifacts [215, 240, 177, 244, 27, 95]. Their methods are usually statistical-based, which leverage handcrafted machine learning matrices, such as SVM[69], to train regression models for perceptual quality prediction[13, 215, 202, 167, 124, 213].

More contemporary methods adopt complex neural networks such as CNN[112], ViT[51], or Swin-Transformer[142] to extract a vast array of perceptually relevant features[118, 223, 29, 263, 297, 284], processing videos in an end-to-end manner. Some of these methods extract multi-scale features to capture both global and local information, which helps in modeling different levels of perceptual quality[286, 49, 297].

Given the subjective nature of video quality assessment, recent models are increasingly looking at ways to decompose a single no-reference VQA score into multiple dimensions [256, 266, 224, 41], such as aesthetics feature, semantic feature, distortion feature, motion feature, etc. They may adopt domain-fusion or knowledge transfer to incorporate information from different feature dimensions, enhancing the overall understanding of video quality [285, 160, 134]. We summarize the VQA methods discussed in this section chronologically in Table 3, while Figure 8 provides an overview of these methods categorized by their backbone architecture.

#### 4.3.1 Early-stage VQA

**VMAF[13]:** Video Multi-method Assessment Fusion framework in video quality assessment lies in its innovative integration of multiple quality-aware features and its focus on both spatial and temporal video distortions. VMAF employs a Support Vector Regression (SVR) model that aggregates elementary video quality metrics, such as Detail Loss Metric (DLM), Visual Information Fidelity (VIF), and Temporal Information (TI), to predict overall video quality. This model effectively weights the contributions of each feature, aligning objective predictions with subjective ground truth data. However, VMAF’s initial design primarily captures compression and scaling artifacts, lacking robust temporal quality measurements. To address this limitation, the authors propose two enhancements: SpatioTemporal VMAF, which incorporates strong temporal features into a unified regression model, and Ensemble VMAF, which combines predictions from multiple models to improve accuracy. Both approaches leverage efficient temporal video features, allowing for better generalization across diverse datasets while maintaining computational efficiency. The introduction of a large subjective database for training further strengthens the model’s predictive capabilities, making VMAF a powerful tool for real-time video quality assessment in various applications, including adaptive streaming and digital cinema.

**Temporal Pooling[233]:** Temporal Pooling enhances video quality assessment by integrating multiple temporal pooling strategies to produce a more reliable and accurate quality prediction. The method operates in two phases: first, it maps the input features from NR-IQA models to frame-level quality predictions, and second, it learns a regression function that fuses these temporally pooled predictions into a final quality score. This dual-phase training ensures that Temporal Pooling captures diverse aspects of perceptual quality by leveraging the strengths of various pooling methods, such as Hysteresis[234], and traditional statistical means. The model’s robustness is demonstrated through its performance across different datasets, particularly in scenarios with varying motion and quality dynamics.

Table 3: Summary of mentioned video-human perception alignment benchmarks.

Reference Type	Model	Year	Backbone	Key Words	Extracted Features	Link
NR	VMAF [13]	2018	ML-based	SVR-based	statistical, temporal	C
NR	VSFA [117]	2019	CNN-based	Gated Recurrent Unit (GRU)	content-aware, distortion	C
NR	Temporal Pooling [233]	2020	IQA model-based	ensemble learning	frame-level, temporal, spatial, statistical	C
NR	VIDEVAL [235]	2020	CNN-based	Modular learning-based, transfer learning	statistical, distortion	C
NR	Patch-VQ [281]	2020	CNN-based	local-to-global model, patch-based, end-to-end	temporal, spatial, patch	C
NR	MDTVSFA [118]	2020	CNN-based	Gated Recurrent Unit (GRU)	content-aware, temporal, distortion	C
NR	GSTVQA [29]	2020	CNN-based	Modular learning-based, Gaussian distribution regularization	multi-scale, temporal, statistical	C
NR	RAPIQUE [236]	2021	CNN-based	modular learning-based, spatial-temporal integration	statistical, temporal, spatial	C
FR+NR	CompressedVQA [223]	2021	CNN-based	Modular learning-based, end-to-end	structure, texture, statistical	C
NR	StarVQA [269]	2021	transformer-based	Modular learning-based, spatial-temporal integration	spatial, temporal, long-range	C
NR	SimpleVQA [224]	2022	CNN-based	Modular learning-based, end-to-end	spatial, motion	C
NR	HVS-5M [286]	2022	CNN-based	domain-fusion	saliency, content, edge, motion, temporal	C
NR	FANet [263]	2022	transformer-based	Local-to-global model	texture, distortion, statistical	C
NR	DOVER [266]	2022	CNN-based	Modular learning-based	aesthetics, technical, perceptual	C
NR	MD-VQA [296]	2023	CNN-based	Modular learning-based	semantic, distortion, motion	C
NR	BVQI [264]	2023	CLIP-based	Local-to-global model	temporal, spatial, semantic, localized	C
NR	Zoom-VQA [297]	2023	transformer-based	Local-to-global model	patch-level, frame-level, clip-level	C
NR	Light-VQA [49]	2023	CNN-based	Modular learning-based	semantic, motion, brightness, noise	C
NR	SB-VQA [80]	2023	transformer-based	Modular learning-based, patch stack based	spatial, temporal, global, patch-based	C
NR	VQT [284]	2023	transformer-based	Modular learning-based	temporal, distortion, long-range	C
NR	Ada-DQA [134]	2023	CNN-based	Modular learning-based, knowledge distillation	motion, distortion, content	C
NR	SSL-VQA [160]	2023	transformer-based	Knowledge transfer	temporal, spatio, statistical, distortion	C
NR	KSVQE [146]	2024	CLIP-based	Modular learning-based	distortion, semantic, quality	C
NR	BVQA [261]	2024	CNN-based	Modular deep learning-based	temporal, spatial, distortion, rectification	C
NR	COVER [41]	2024	transformer-based	domain-fusion	technical, aesthetics, semantic	C
NR	Light-VQA+ [300]	2024	transformer-based	Modular learning-based	brightness, noise, temporal, spatial	C
NR	RQ-VQA [225]	2024	CLIP-based	Knowledge transfer, vision-language guidance	motion, technical, temporal	C
NR	PTM-VQA [285]	2024	transformer-based	Knowledge transfer	visual, motion, semantic, distortion, temporal	C
NR	Priorformer [189]	2024	transformer-based	Modular learning-based	temporal, distortion, technical	C
NR	UGVQ [295]	2024	CLIP-based	domain-fusion	temporal, spatial, semantic	C
NR	ReLax-VQA [252]	2024	transformer-based	spatial-temporal integration, layer-stacking	temporal, motion, technical	C

Note: FR: Full Reference; NR: Non-Reference. ML refers to machine learning.

#### 4.3.2 CNN-based VQA

**VSFA[117]:** VSFA introduces a novel no-reference video quality assessment (NR-VQA) model that uniquely integrates knowledge from the human visual system into a deep learning framework. The model employs a modified pre-trained Convolutional Neural Network (CNN) to extract content-aware features from video frames, which are crucial for understanding the perceived quality based on video content. Following feature extraction, a Gated Recurrent Unit (GRU) is utilized to model long-term dependencies, allowing the model to predict frame quality effectively. Additionally, the method incorporates a differentiable, subjectively-inspired temporal pooling layer that accounts for temporal hysteresis effects, ensuring that the overall video quality is assessed in a manner that reflects human perception.

**VIDEVAL[235]:** VIDEVAL improves upon VQA models by employing a feature selection and fusion strategy that optimally combines the strengths of various high-performing blind VQA (BVQA) models while maintaining computational efficiency. By extracting 60 relevant features from a pool of 763 statistical features used in prior methods, VIDEVAL effectively balances performance and resource consumption, costing comparely lower computational costs. The model leverages a diverse set of features that capture different perceptual domains, enhancing its robustness across various datasets and use cases. Additionally, the integration of deep learning features through ensembling with models like ResNet-50 and KonCept512 has been shown to further elevate performance, indicating the potential of transfer learning in the UGC-VQA context.

**Patch-VQ[281]:** Patch-VQ develops two key architectures: Patch VQ (PVQ) and PVQ Mapper. PVQ employs a deep neural network that integrates both 2D and 3D feature extraction to effectively capture complex distortions in videos. By learning the relationships between global video quality and local space-time v-patch quality, PVQ achieves state-of-the-art performance on various user-generated content datasets, surpassing existing models. Its local-to-global architecture enhances generalizability across diverse video content, making it robust in predicting perceptual quality. Additionally, the PVQ Mapper introduces a novel capability to generate detailed space-time quality maps, which allow for the localization and visualization of distortions within videos. This feature not only aids in understanding the nature of video quality degradation but also provides actionable insights for content creators and streaming platforms.

**MDTVSFA[118]:** MDTVSA designs a three-stage framework to effectively tackle the complexities of assessing in-the-wild videos. The first stage involves a relative quality assessor that predicts the relative quality of videos using a monotonicity-induced loss, ensuring consistent quality rankings aligned with human perception. This stage incorporates content-aware feature extraction and models temporal-memory effects, enhancing the model's ability to account for the nuances of human visual perception. The second stage employs a nonlinear mapping module, utilizing a 4-parameter logistic function to translate relative quality scores into perceptual quality, addressing the nonlinearity inherent in human perception of video quality. Finally, the third stage focuses on dataset-specific perceptual scale alignment, which aligns the predicted perceptual quality with subjective quality assessments.

**GSTVQA[29]:** GSTVQA introduces a no-reference VQA framework that effectively assesses the perceptual quality of videos across diverse acquisition, processing, and compression techniques. It employs a multi-scale feature extraction scheme that captures quality features at different scales, enhancing the model's ability to adapt to various video characteristics. Besides, an attention module is integrated to weight the extracted features based on their importance, ensuring that the most relevant information is prioritized during quality prediction. Additionally, it incorporates a Gaussian distribution to unify the quality features of each frame, with learnable mean and variance parameters that help mitigate the domain gap caused by varying content and distortion types. Furthermore, the model utilizes a pyramid aggregation module in the temporal domain, which effectively combines features over time to improve prediction accuracy.

**RAPIQUE[236]:** RAPIQUE integrates both spatial and temporal scene statistics with deep learning features to achieve rapid and accurate predictions. It employs a two-branch framework that combines low-level scene statistics and high-level deep convolutional features, allowing for a comprehensive analysis of video quality. The model utilizes aggressive spatial and temporal sampling strategies to exploit content and distortion redundancies, enhancing efficiency without compromising performance. Additionally, RAPIQUE introduces a novel spatial NSS feature extraction module, providing a cost-effective alternative to traditional feature-based models. It also features a pioneering temporal

statistics model that captures bandpass regularities in natural videos, making it suitable for motion-intensive content.

**CompressedVQA**[223]: The proposed deep learning-based VQA framework comprises three main components: the feature extraction module, the quality regression module, and the quality pooling module. The feature extraction module fuses features from intermediate layers of a convolutional neural network (CNN), enabling the model to leverage both low-level visual information and high-level semantic features, which enhances the quality-aware feature representation for both full-reference and no-reference tasks. The quality regression module employs a fully connected layer to convert these quality-aware features into frame-level quality scores. Finally, the quality pooling module utilizes a subjectively-inspired temporal pooling strategy to aggregate frame-level scores into a comprehensive video-level score. This end-to-end learning approach allows the model to effectively capture the complex relationships between video quality and raw pixel data, outperforming existing state-of-the-art VQA models on various datasets, including the Compressed UGC VQA database.

**SimpleVQA**[224]: SimpleVQA introduces an effective and efficient deep learning-based architecture that integrates a feature extraction module, a quality regression module, and a quality pooling module. The model extracts two types of quality-aware features: spatial features for addressing spatial distortions and spatial-temporal features for capturing motion distortions, with spatial features learned directly from raw video pixels in an end-to-end manner and motion features derived from a pretrained action recognition network. Additionally, the model employs a multi-scale quality fusion strategy to effectively assess video quality across different resolutions, utilizing multi-scale weights derived from the contrast sensitivity function of the human visual system, which considers viewing environment information. This comprehensive approach not only enhances the model’s performance but also demonstrates its generalizability.

**HVS-5M**[286]: HVS-5M revisits the Human Visual System (HVS) and integrating five representative characteristics to enhance the VQA evaluation process. The model is structured around five key modules: the visual saliency module, which employs SAMNet to generate a saliency map; the content-dependency module and the edge masking module, both utilizing ConvNeXt to extract spatial features that are weighted by the saliency map; the motion perception module, which leverages SlowFast to capture dynamic temporal features; and the temporal hysteresis module, which simulates the memory mechanism of human perception. This domain-fusion design allows HVS-5M to simultaneously assess frame-level spatial features and video-level temporal features.

**DOVER**[266]: Disentangled Objective Video Quality Evaluator presents advancements, particularly for user-generated content (UGC) VQA. It develops the DIVIDE-3k dataset, which comprises 3,590 videos and 450,000 subjective quality opinions, capturing both aesthetic and technical perspectives on video quality. In addition, DOVER employs an architecture that disentangles aesthetic and technical evaluations, allowing for a more nuanced assessment of video quality. Furthermore, the model incorporates a subjectively-inspired fusion strategy that improves overall quality predictions, making it more reliable for practical applications. Finally, DOVER also provides insights into the perceptual mechanisms underlying human assessments of video quality.

**MD-VQA**[296]: MD-VQA introduces a framework specifically designed for evaluating user-generated content (UGC) live videos. It developed a large-scale UGC Live VQA database, which comprises 418 diverse source videos and 3,762 compressed versions generated under various encoding settings. This database serves as a critical resource for training and validating VQA models. The core of the MD-VQA model lies in its multi-dimensional approach, which assesses video quality through three key components: semantic features, distortion features, and motion features. Semantic features are extracted using pretrained convolutional neural networks (CNNs), capturing the content-related aspects of the videos. Distortion features are derived from specific handcrafted image distortion descriptors, addressing common quality issues such as blur and noise. Motion features are obtained through pretrained 3D-CNNs, which analyze the temporal dynamics of the video clips. By integrating these diverse feature sets, the MD-VQA model could provide a more holistic and interpretable assessment of visual quality.

**Light-VQA**[49]: Light-VQA is a specialized quality assessment framework tailored for low-light video enhancement, which integrates handcrafted features with deep-learning-based features to effectively represent quality-aware characteristics of enhanced videos. The model emphasizes the importance of brightness and noise, which significantly impact low-light video quality, by incorporating specific features such as brightness, brightness consistency, and noise into its design. Second, Light-

VQA leverages a newly constructed Low-Light Video Enhancement Quality Assessment (LLVE-QA) dataset, comprising 254 original low-light videos and 1,806 enhanced videos generated by various state-of-the-art low-light video enhancement (LLVE) algorithms, facilitating a comprehensive evaluation of enhancement techniques. This model not only enhances the assessment of low-light video quality but also serves as a foundational tool for future research in low-light video enhancement algorithms, bridging the gap between enhancement techniques and quality evaluation.

**Ada-DQA**[134]: **Adaptive Diverse Quality-aware Feature Acquisition** integrates diverse pretrained models to capture a comprehensive range of quality-related features, such as content, distortion, and motion. The framework comprises three key components: first, it selects multiple frozen pretrained models as feature extractors, which reduces computational costs while retaining essential quality information. Second, the Quality-aware Acquisition Module (QAM) adaptively aggregates features from these models, applying dynamic weights to emphasize the most relevant features for each video sample. This adaptive approach is further enhanced by a sparsity constraint that promotes focus on critical quality-related aspects. Finally, the learned quality representation serves as supplementary supervisory information during the training of a lightweight VQA model, utilizing knowledge distillation to optimize performance while minimizing computational demands during inference.

**BVQA**[261]: **Blind Video Quality Assessment** introduces a structured approach that enhances the accuracy and reliability of quality predictions across diverse video content. The model comprises three key components: a base quality predictor, a spatial rectifier, and a temporal rectifier. The base quality predictor processes a sparse set of spatially down-sampled key frames to generate an initial quality estimate. The spatial rectifier, utilizing a shallow convolutional neural network, refines this estimate by analyzing the Laplacian pyramids of the keyframes at their actual spatial resolution, adjusting the quality score based on spatial resolution changes. Similarly, the temporal rectifier employs a lightweight CNN to assess video chunks around the keyframes at the actual frame rate, further refining the quality prediction by accounting for temporal variations. This modular design not only allows for targeted adjustments based on specific distortions but also facilitates the integration of additional rectifiers for other video attributes, enhancing the model's extensibility.

#### 4.3.3 Transformer-based VQA

**StarVQA**[269]: StarVQA leverages a space-time attention network built on the Transformer architecture. It implements a unique vectorized regression loss function, which encodes mean opinion scores (MOS) into a probability vector, enhancing the training process. StarVQA effectively captures long-range spatiotemporal dependencies by incorporating space-time position information into the input, allowing it to analyze video sequences more comprehensively. The architecture alternates between divided space and time attention mechanisms, enabling the model to focus on both spatial and temporal features simultaneously. Additionally, the model's design is optimized for high-resolution videos, revealing the advantages of the Transformer architecture in this context.

**FANet**[263]: Fragment Attention Network introduces the match constraint for pooling layers, which aligns pooling operations with sampled mini-cubes, ensuring controlled pixel fusion. FANet employs a modified Video Swin Transformer backbone, enhanced with Gated Relative Position Biases (GRPB) to accurately represent pixel positions within fragments. Additionally, it incorporates an Intra-Patch Non-Linear Regression (IP-NLR) head, replacing the traditional pool-first approach, which allows for the prediction of local quality maps rather than just quality scores. The model demonstrates unprecedented efficiency, achieving up to 1612 $\times$  faster inference times compared to existing methods while maintaining competitive accuracy. Overall, FANet's GRPB and IP-NLR modules position it as a robust solution for efficient and accurate video quality assessment, particularly in the context of increasingly large and complex video data.

**Zoom-VQA**[297]: Zoom-VQA reaches high VQA performance by effectively capturing spatio-temporal features across three levels: patches, frames, and clips. The model comprises three key components: the patch attention module, which focuses on region-of-interest in the spatial dimension; the frame pyramid alignment, which addresses multi-level feature information; and the clip ensemble strategy, which integrates distortions over the temporal dimension. This comprehensive design allows Zoom-VQA to analyze videos in a hierarchical manner, enhancing its ability to assess quality by considering both local and global information. The architecture includes two branches: an image quality assessment (IQA) branch that processes individual frames for global insights, and a video quality assessment (VQA) branch that utilizes spatio-temporal information from video segments.

**SB-VQA**[80]: Stack-Based Video Quality Assessment advances the understanding and application of VQA in video enhancement and restoration. It introduces a scalable stack-based architecture that enhances the evaluation of video quality by effectively capturing both spatial and temporal features. Additionally, the framework is fine-tuned on a newly constructed dataset, PGCVQ, which consists of professionally generated content, addressing the gap in VQA research focused on such videos. This dataset allows for a more accurate assessment of video quality in real-world applications. Furthermore, SB-VQA incorporates a novel approach to analyze the influence of video content on subjective quality perception, utilizing heatmaps to explore the relationship between VQA algorithms and video characteristics. This multifaceted analysis not only validates the effectiveness of existing VQA methods on professionally generated content but also highlights the importance of content characteristics, such as resolution and distortion, in human quality assessment.

**VQT**[284]: Visual Quality Transformer introduces a Sparse Temporal Attention (STA) mechanism that efficiently selects keyframes by analyzing the temporal correlation between frames, significantly reducing computational complexity from  $O(T^2)$  to  $O(T \log T)$ . This allows VQT to focus on frames that contain critical distortions, enhancing the model’s ability to perceive multi-distortion characteristics in videos. Secondly, VQT employs a Multi-Pathway Temporal Network (MPTN) that stacks multiple STA modules with varying sparsity levels, enabling simultaneous capture of different distorted features. This dual-component architecture not only improves the model’s performance on non-reference VQA tasks but also demonstrates superior results compared to state-of-the-art methods, achieving notable increases in performance metrics like PLCC on various datasets. Additionally, VQT exhibits good generalization capabilities, making it applicable to other computer vision tasks, such as video classification, while maintaining lower computational costs compared to traditional dense attention mechanisms.

**SSL-VQA**[160]: SSL-VQA leverages a self-supervised Spatio-Temporal Visual Quality Representation Learning (ST-VQRL) framework, which serves as a robust feature extractor for video quality. The method is designed to operate effectively with limited labeled data, addressing the challenges posed by user-generated content (UGC) videos that often lack reference quality. SSL-VQA employs a dual-model architecture that facilitates knowledge transfer between two quality prediction models, enhancing the learning process through a combination of supervised and semi-supervised learning techniques. The first stage focuses on self-supervised learning to capture rich spatio-temporal features from unlabelled videos, while the second stage utilizes these features in a semi-supervised framework to optimize performance using a small set of labeled data. By incorporating a novel statistical contrastive learning loss, SSL-VQA improves the robustness of the learning process.

**COVER**[41]: COmprehensive Video quality EvaluatoR integrates three distinct branches that evaluate video quality from technical, aesthetic, and semantic perspectives. The architecture comprises a Swin Transformer backbone for the technical branch, which processes spatially sampled crops to assess technical quality. The aesthetic branch employs a ConvNet that analyzes subsampled frames to derive aesthetic quality, while the semantic branch utilizes a CLIP image encoder to extract high-level semantic information from resized frames. A key innovation of COVER is the Simplified Cross-Gating Block (SCGB), which facilitates interaction between the branches, allowing for effective feature fusion before the final quality prediction. This multi-faceted approach enables COVER to generate a comprehensive quality score by averaging the scores from each branch, thus providing a more holistic evaluation of video quality. The model’s design not only enhances the accuracy of quality assessments but also ensures real-time processing capabilities for high-definition videos, making it suitable for large-scale applications in video streaming platforms.

**Light-VQA+**[300]: Light-VQA+ enhances the capability to evaluate exposure-corrected videos through a multi-dimensional approach that integrates vision-language guidance. Unlike its predecessor, Light-VQA[49], which relied on traditional handcrafted feature extraction methods, Light-VQA+ leverages advanced deep learning techniques, specifically utilizing multimodal large language models (MLLMs) like CLIP for improved feature extraction related to brightness, noise, and brightness consistency. The model divides input videos into clips, extracting both spatial information (via Swin-Transformer and CLIP) and temporal information (using the SlowFast Model and CLIP). A cross-attention module fuses these diverse features, followed by fully connected layers that regress the fused features into a unified quality score. Additionally, Light-VQA+ introduces trainable attention weights to mimic the Human Visual System (HVS), allowing the model to prioritize certain video clips based on perceived importance. This results in a more accurate and efficient assessment of video quality, particularly for low-light and over-exposed videos.

**PTM-VQA**[285]: PTM-VQA utilizes a novel approach of extracting features from multiple pretrained models with fixed weights, allowing for the integration of various knowledge domains without incurring significant computational costs. This is particularly beneficial as it enables the model to capture distinct characteristics related to video quality, such as content attractiveness and distortion types. Secondly, PTM-VQA introduces an Intra-Consistency and Inter-Divisibility (ICID) loss function, which imposes constraints on the extracted features to ensure they reside in a unified quality-aware latent space while promoting separation among different quality clusters. This dual constraint mechanism enhances the model’s ability to generalize across different datasets.

**Priorformer**[180]: PriorFormer enhance adaptability and representation capabilities. Central to this approach is the Content Prior Features Extractor, which leverages the CLIP model to generate detailed content embeddings that capture the semantic essence of the video. Complementing this is the Distortion Prior Feature Extractor, which constructs a distortion graph to identify and represent various distortion types and levels present in UGC videos. These two sets of embeddings are utilized as adaptive prior tokens within the vision transformer architecture, allowing the model to effectively incorporate both content and distortion information into the quality assessment process. Furthermore, PriorFormer includes a temporal feature fusion module that employs gated recurrent units (GRU) to aggregate frame-level quality assessments over time, addressing the temporal dynamics inherent in video content. This combination of content and distortion priors, along with the temporal perception mechanism, enables PriorFormer to provide a more nuanced and accurate evaluation of video quality.

**ReLaX-VQA**[252]: ReLaX-VQA employs a spatio-temporal fragment sampling module that effectively captures the quality-aware features from successive video frames by utilizing frame differencing and optical flow techniques. This innovative approach allows the model to analyze both spatial and temporal aspects of video quality, addressing the complexities inherent in user-generated content (UGC). Secondly, the model incorporates a deep neural network (DNN) layer stack module, which fuses multi-layered features extracted from the video frames, enhancing the model’s ability to recognize and assess various quality distortions. Finally, the quality regression module translates the extracted features into a quality score, providing a robust evaluation of video quality without the need for reference content.

#### 4.3.4 CLIP-based VQA

**BVQI**[264]: Blind Video Quality Index advances video quality assessment (VQA) by integrating the Semantic Affinity Quality Index (SAQI) and its localized variant, SAQI-Local. BVQI leverages the capabilities of Contrastive Language-Image Pre-training (CLIP) to evaluate video quality based on semantic content, allowing for a robust assessment of both aesthetic and authentic distortions without the need for human-annotated quality scores. The SAQI component focuses on measuring the affinity between visual features and textual prompts, effectively capturing semantic-related quality perceptions. Additionally, the BVQI incorporates two technical naturalness metrics: the Spatial Naturalness Index and the Temporal Naturalness Index, which enhance the overall quality prediction by considering both spatial and temporal aspects of video content. The method also introduces an efficient fine-tuning strategy that optimizes text prompts and fusion weights.

**KSVQE**[146]: Kaleidoscope Short-form Video Quality Evaluator is an efficient VQA framework, particularly for short-form user-generated content (S-UGC). Firstly, it introduces a large-scale kaleidoscopic short-form video database, termed KVQ, which comprises 4200 videos collected from popular platforms, addressing the unique challenges posed by diverse content creation modes and sophisticated processing workflows. KSVQE incorporates innovative components such as the Quality-Aware Region Selection module (QRS) and Content-Adaptive Modulation (CaM), which leverage the capabilities of the pre-trained vision-language model CLIP to enhance content understanding and identify quality-determined regions. Additionally, the model integrates a Distortion-Aware Modulation (DaM) module, utilizing the CONTRIQUE model to improve distortion identification, thereby addressing the indistinguishability of distortions in S-UGC videos. The combination of these modules allows KSVQE to effectively capture quality-aware content and complex distortions.

**RQ-VQA**[225]: RQ-VQA is specifically tailored for social media videos, which often exhibit unique challenges due to diverse content and complex processing workflows. The method enhances the Simple VQA framework by integrating rich quality-aware features extracted from various pre-trained models, including both blind image quality assessment (BIQA) and BVQA models. Key components of the proposed model include a trainable spatial quality module and a fixed temporal quality module,

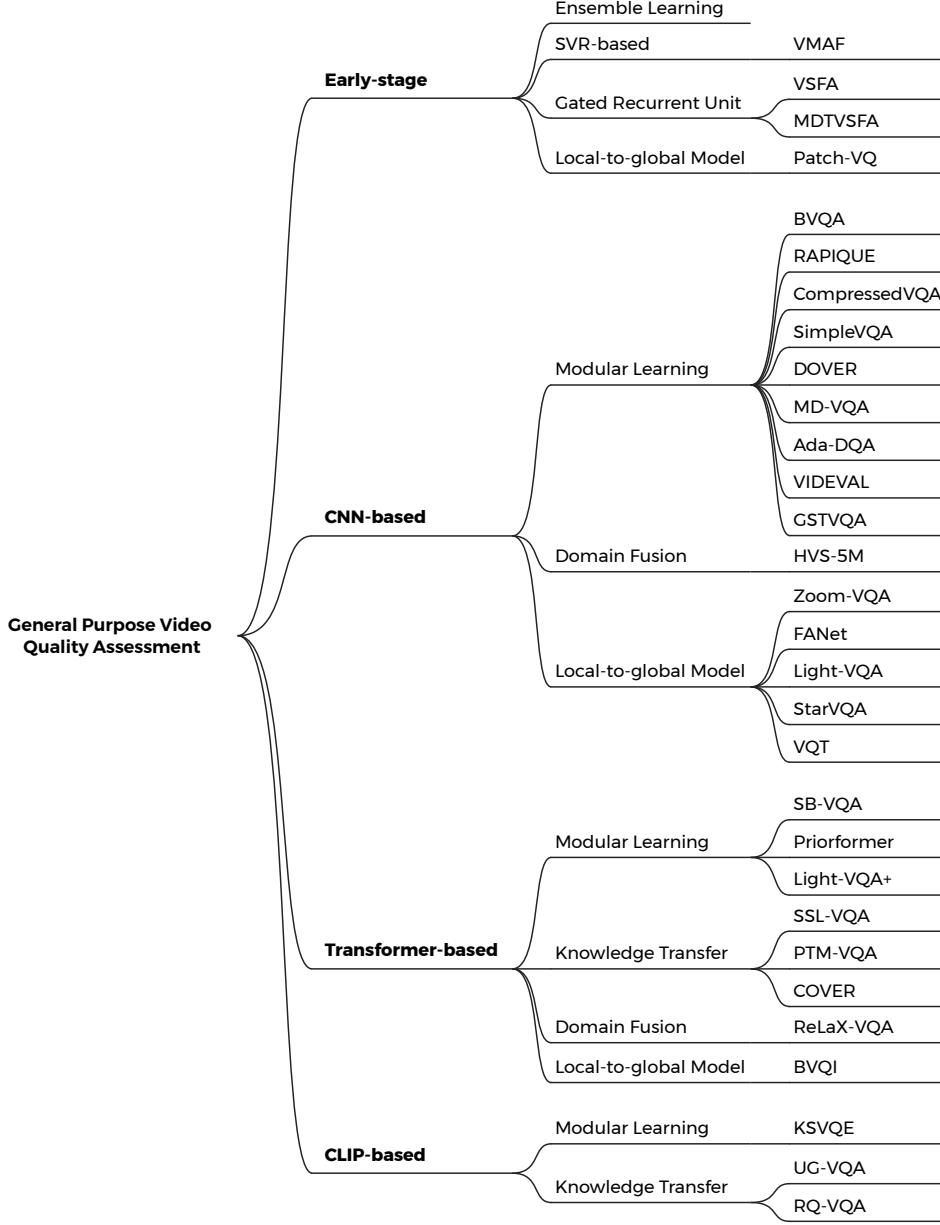


Figure 8: Summary of mentioned video-human perception alignment benchmarks by backbone.

which work in tandem to capture both spatial and temporal distortions effectively. Additionally, the model incorporates features from deep learning architectures, such as motion and jerkiness metrics, to improve quality prediction accuracy. Overall, the contributions lie in the innovative integration of multiple quality-aware features and the adaptability of the Simple VQA framework to address the specific needs of social media video quality evaluation.

**UGVQ[295]: Unified Generative Video Quality** provides a structured framework that comprehensively evaluates the quality of AI-generated content (AIGC) videos across three critical dimensions: spatial quality, temporal quality, and text-to-video alignment. The model consists of three key modules: a feature extraction module, a text-visual feature fusion module, and a quality regression module. The feature extraction module captures spatial features from keyframes and temporal features from video sequences, ensuring a robust representation of the video content. The text-visual feature fusion module integrates visual features with textual semantics derived from the prompts, enhancing the model's ability to assess alignment between generated videos and their descriptions. Finally, the quality regression module synthesizes these features to produce a comprehensive quality score. By

leveraging advanced techniques such as CLIP for visual and textual feature extraction and SlowFast for motion representation, UGVQ addresses the unique challenges posed by AIGC videos.

#### 4.4 Evaluation Metrics

The algorithms mentioned in previous sections are usually evaluated based on the correlation of subjective and objective ratings. Among various statistical indices, SRCC, KRCC, PLCC, RMSE and MAE are five frequently used metrics to highlight various facets of the VQA model's performance. The correlation between the subjective quality ratings and the objective predicted scores is calculated by SRCC, KRCC, and PLCC, which show the prediction monotonicity. The error between the subjective quality ratings and the objective predicted scores is computed by RMSE and MAE, which shows the prediction accuracy. Better performance is indicated by greater (near to 1) SRCC, KRCC, and PLCC values and lower (closer to 0) RMSE and MAE values. The details of these five metrics are introduced as follows:

**Spearman Rank-order Correlation Coefficient (SRCC)** The Spearman Rank-Order Correlation Coefficient is a non-parametric measure used to evaluate the strength and direction of the monotonic relationship between two ranked variables. Unlike Pearson's correlation, which assesses linear relationships, SRCC is useful when data do not meet the assumptions of normality or when the relationship is not linear. It works by ranking the data points and then applying Pearson's correlation formula to these ranks. It could be represented as:

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

where  $N$  represents the total number of test movies and  $d_i$  represents the difference value between the subjective and objective scores for the  $i$ -th video. SRCC is commonly used in situations where ordinal data or nonlinear associations are present, making it a versatile tool in statistical analysis.

**Kendall Rank-order Correlation Coefficient (KRCC)** Kendall Rank-Order Correlation Coefficient is a non-parametric statistic that measures the ordinal association between two variables. It evaluates how well the relationship between the variables can be described using a monotonic function. Unlike SRCC, which ranks the data and then applies Pearson's correlation to the ranks, Kendall's coefficient is based on the number of concordant and discordant pairs of observations. A concordant pair occurs when the order of the ranks of both variables is consistent, while a discordant pair occurs when the order is reversed. It could be represented as:

$$\text{KRCC} = \frac{N_c - N_d}{\frac{1}{2}N(N - 1)}$$

where the number of concordant pairs is  $N_c$ , and the number of discordant pairs is  $N_d$ . KRCC is particularly useful for small datasets or when there are many tied ranks, providing a robust method for assessing ordinal relationships.

**Pearson Linear Correlation Coefficient (PLCC)** Pearson Linear Correlation Coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Developed by Karl Pearson, it assumes that the relationship between the variables is linear and both variables are normally distributed. The coefficient value ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 suggests no linear relationship. It is formulated as:

$$\text{PLCC} = \frac{\sum_i^N (q_i - \bar{q}) \cdot (o_i - \bar{o})}{\sqrt{\sum_i^N (q_i - \bar{q})^2 \cdot (o_i - \bar{o})^2}}$$

where  $o_i$  and  $q_i$  represent the subjective opinion score and the nonlinear-fitted objective score for the  $i$ -th video,  $\bar{o}$  and  $\bar{q}$  indicate the mean values of all  $o_i$  and  $q_i$  scores. It is widely used in various fields such as science, economics, and social sciences to assess how one variable changes in relation to another. However, it is sensitive to outliers and should only be used when a linear relationship is expected.

**Root Mean Square Error (RMSE)** Root Mean Square Error is a commonly used metric to measure the difference between predicted and actual values in a dataset. It provides an indication of the

Table 4: Summary of the mentioned video-human instruction alignment benchmark datasets.

Dataset	Year	Instruction Type	Domain	#Video	#Video Clips	#Sentence	Duration(hrs)	Resolution	Link
MSVD [30]	2011	text	multi-category	-	1,970	70,028	5.3	-	
UCF101 [218]	2012	text	101 action classes	-	13,320	-	27	320p	
MPII-MD [197]	2015	text	movie	94	68,337	68,337	73.6	720p	
MSR-VTT [272]	2016	text	multi-category	7,180	10,000	200,000	41.2	240p	
Kinetics [93]	2017	text	400 action classes	-	306,245	-	~ 850	340p/128p	
YouCook2 [299]	2018	text	cooking	2,000	14,000	14,000	176	-	
TACos-Multi-Level [196]	2014	text	cooking	273	14,105	52,593	176	-	
HowTo100M [158]	2019	text	instruction	1.22M	136M	136M	134,472	240p	
VATEX [251]	2019	text	open	41,269	41,269	825,380	~ 115	-	
Webvid-2M [12]	2021	text	instruction	-	2.5M	2.5M	13K	360p	
InternVid [254]	2023	text	open	7.1M	234M	234M	760.3	720p	
Panda-70M [34]	2024	text	open	3.8M	70.8M	70.8M	166.8K	720p	
VAST-27M [33]	2024	text, audio	open	3.3M	27M	297M	75K	720p	

magnitude of prediction errors by calculating the square root of the average of the squared differences between predicted and observed values. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_i^N (q_i - o_i)^2}$$

where  $q_i$  represents the observed values,  $o_i$  the predicted values, and  $N$  the number of data points. RMSE is sensitive to large errors due to the squaring process, making it particularly useful when large deviations from the expected values are of greater concern. A lower RMSE value indicates better model performance and a closer fit between predictions and actual outcomes. RMSE is widely used in regression models and forecasting, helping to evaluate the accuracy and reliability of a model's predictions.

**Mean Absolute Error (MAE)** Mean Absolute Error is a widely used metric in regression analysis to measure the average magnitude of errors between predicted and actual values without considering their direction. It is calculated as:

$$\text{MAE} = \frac{1}{N} \sum_i^N |q_i - o_i|$$

where  $N$  represents the number of data points. Unlike Root Mean Square Error (RMSE), MAE does not give extra weight to large errors, as it simply takes the absolute value of the residuals, making it less sensitive to outliers. This makes MAE a more robust and interpretable measure of overall prediction accuracy, especially in cases where all errors are treated with equal importance. A lower MAE value indicates better model performance. It is commonly used in various fields, such as machine learning, time series analysis, and economics, to assess the accuracy of predictive models.

## 5 Alignment with human Instructions

Generating videos based on specified human instructions has always been an important topic in video generation. It is also an essential aspect that many current AI-generated video (AIGV) studies strive to improve. Over the past decade, numerous benchmark datasets and evaluation methods have been proposed to assess the alignment of generated videos with human instructions (i.e., text, audio). This section introduces representative benchmark datasets and evaluation methods.

### 5.1 Benchmark Datasets

In this section, we introduce several representative benchmark datasets widely used before and during the era of AIGV to assess the alignment between video and human instruction. Specifically, the general data format of the benchmark datasets can be defined as  $D = \{V, I\}$ , where  $V$  represents the

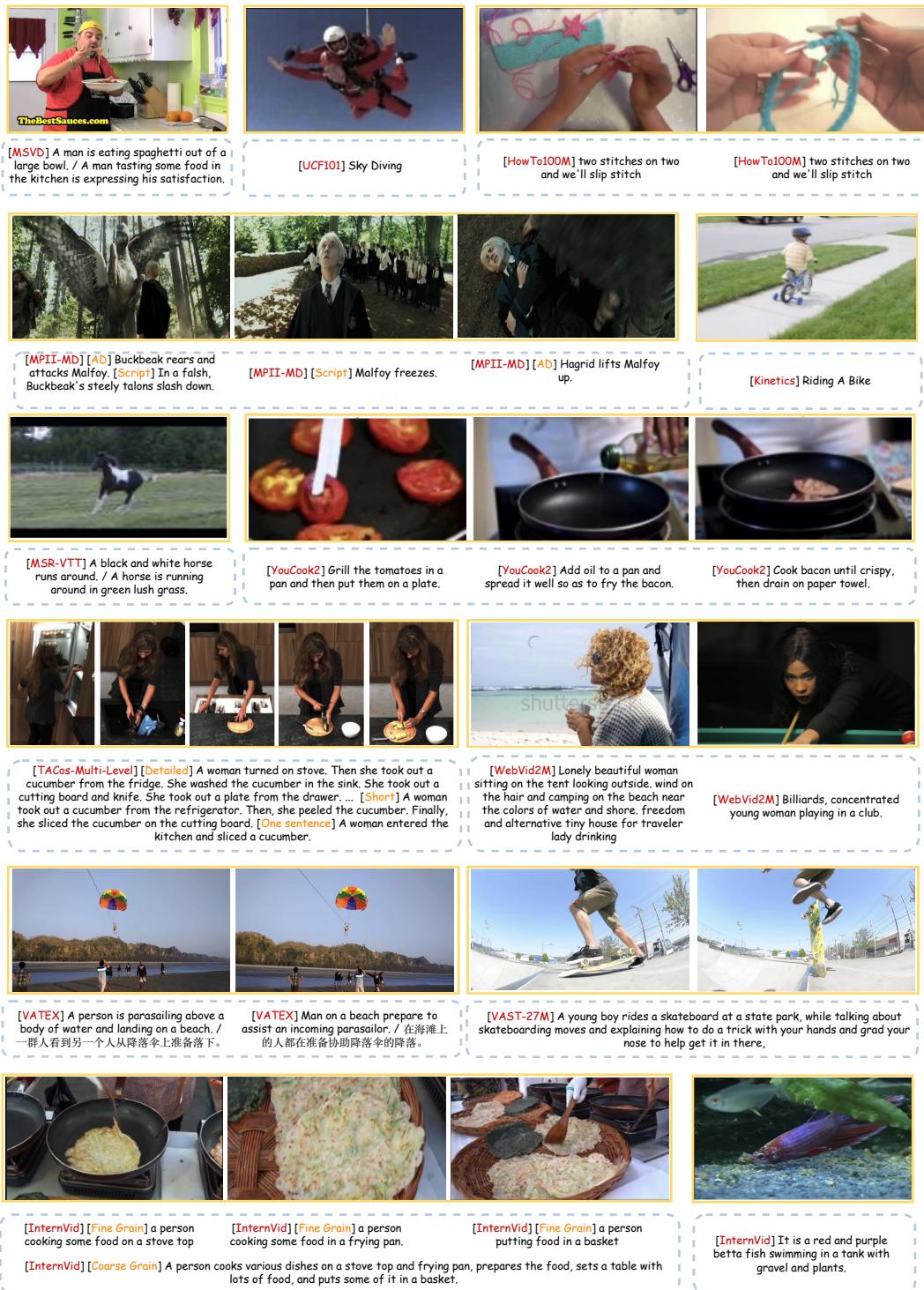


Figure 9: Exemplar Cases from Alignment with Human Instructions Benchmark Datasets. Here we intercept one or several frames of the video clip as a representation of the video within each dataset, along with the different corresponding textual representations.

video clip and  $I$  represents the human instructions, which are typically the textual descriptions (in addition to VAST-27M [33], which also contains the audio descriptions). Each benchmark dataset contains thousands to millions of video clips, along with the corresponding descriptions that are either manually annotated or generated by multimodal models. Table 4 provides an overview of the key details for each benchmark dataset mentioned, while Figure 9 illustrates an exemplar case from each dataset.

**MSVD** [30]: The Microsoft Research Video Description Corpus (MSVD) is one of video description’s most popular early benchmark construction efforts. During the data collection phase of MSVD, the Amazon Mechanical Turk workers are asked to find short video segments depicting single and unambiguous events from YouTube and then mute the video to write at least one sentence to summarize the single video events. As a result, the outcome MSVD dataset consists of 70,028 English descriptions collected from 1,970 short video snippets.

**UCF101** [218]: UCF101 is a classic and widely used benchmark dataset for evaluating the alignment between human instructions and video content regarding human actions. UCF101 is a further refinement of the previous precursor work (i.e., UCF Sports [195], UCF11 [135], and UCF50 [191]). It comprises 101 action classes, over 13k clips, and 27 hours of video data. The 101 action classes in UCF101 can be divided into five types: *Human-Object Interaction*, *Body-Motion Only*, *Human-Human Interaction*, *Playing Musical Instruments*, and *Sports*. The source videos of UCF101 are downloaded from YouTube and feature challenges like poor lighting, cluttered backgrounds, and severe camera motion. For one specific action, the corresponding clips are divided into 25 groups. Each contains 4-7 clips with common features, such as the background and actors. All clips have a fixed frame rate and resolution of 25 FPS and  $320 \times 240$ , respectively.

**TACos-Multi-Level** [196]: TACos-Multi-Level introduces three levels of detailed text descriptions of each single video and addresses the limitation that most of the existing benchmarks for automatically evaluating the video alignment with human instructions focus on a fixed level of detail (i.e., the ground-truth instructions are mainly described as a single sentence). The data is collected via Amazon Mechanical Turk based on the TACoS corpus [193], which contains various cooking videos of 26 different dishes and aligned text descriptions. The human annotators are asked to describe the videos from the TACos corpus in three ways: 1) a detailed description with at most 15 sentences, 2) a short description with 3-5 sentences, and 3) a single sentence. TACos-Multi-Level then employs an intermediate semantic representation (SR) to split each video into video snippets, thereby better aligning the multi-level video descriptions. As a result, TACos-Multi-Level extends to 273 videos, with 14,105 video clips and 52,593 multi-grain associated caption sentences. This benchmark also reveals that as the length of the annotated description decreases, the verbalized information will be more ‘compressed’ according to the topic of the video.

**Kinetics** [93]: The Kinetics dataset can be seen as an extension of the previous human actions video datasets UCF101, as it bridges for UCF101’s lack of data scale and sufficient variation. Similar to UCF101, the human actions in Kinetics data can be divided into several types: *Person Actions (singular)* (i.e., drawing, laughing), *Person-Person Actions* (i.e., hugging, kissing), *Person-Object Actions* (i.e., opening present, mowing lawn). Specifically, Kinetics contains 400 human action classes. For each action, there are 400-1150 at least 10s lasting video clips, with a total number of 306,245. Unlike the video clips from UCF101, which may be from the same source video and have a fixed resolution, each video clip from Kinetics is taken from a different YouTube source video and has variable resolution and frame rate, ensuring the content variation of this benchmark. In addition, several successor datasets have been constructed based on the original Kinetics dataset, such as Kinetics-600 [24] and Kinetics-700 [25], which are larger in terms of human action class.

**MPII-MD** [197]: MPII-MD (MPII Movie Description) dataset is proposed to evaluate the video’s alignment with human instructions in terms of the movie domain. Inspired by the utilization of audio descriptions (ADs) for blind or visually impaired people, MPII-MD collects Blu-ray movies with ADs and selects a set of 55 films of diverse genres. The audio description for each movie is first segmented using the Fast Fourier Transform (FFT) and transcribed by a crowd-sourced transcription service. MPII-MD then splits each movie into video clips according to the time stamps for each spoken sentence provided by audio segments, thereby constructing the video-caption pairs. In addition to collecting the video-caption pairs by ADs, MPII-MD also collects several movies by mining the script web resources, resulting in 94 unique films. These movie data are automatically processed via [42, 111] to align the script caption and video clip content. Specifically, MPII-MD ends up with 94

HD movies with a total duration of 73.6 hours, containing 68,337 video clips and the corresponding sentences. Different from most of the prior video-text alignment benchmark datasets, such as TACoS-Multi-Level [196], which focus on short video clips and are limited in video duration, this benchmark provides a novel and much longer data format to evaluate the text description alignment with the video content.

**YouCook2** [299]: YouCook2 contains 2,000 videos from 89 recipes with a total length of 176 hours. The recipes are from four major cuisines (i.e., Africa, Americas, Asia, and Europe) and are various in cooking styles, methods, ingredients, and cookware. The source videos of the YouCook2 dataset are from YouTube, which is of various challenges, such as fast camera motion, camera zooms, and scene-type changes. In specific, each video in YouCook2 contains 3-16 procedure segments. Compared with its precursor work, YouCook [44], which does not have the *Procedure Annotation* ability, each procedure segment in YouCook2 has time boundary annotations and is described by English sentences.

**MSR-VTT** [272]: MSR-VTT (MSR-Video to Text) dataset is proposed to give a more comprehensive video benchmark for video understanding, especially for translating video to text, instead of focusing on specific fine-grained domains with simple descriptions. MSR-VTT contains 257 popular queries corresponding to 20 main categories: music, people, gaming, sports, news, etc. Each query has a corresponding video list with a length of 118. Specifically, MSR-VTT contains 10k web video clips with 41.2 hours and 200k clip-sentences pairs, and clip in the dataset is annotated with about 20 natural sentences by 1,327 Amazon Mechanical Turk workers, providing a larger-scale video description with a more comprehensive topics dataset benchmark in earlier days.

**HowTo100M** [158]: HowTo100M dataset is a large-scale dataset that consists of 136M video clips from 1.22M narrated instructional videos depicting humans performing and describing over 23,000 visual tasks, including various activities such as cooking, handcrafting, personal care, gardening, etc. These activities tasks are first selected from an extensive list of activities using *WikiHow*, which contains 120,000 articles on *How to...* for various domains, then limited the list to "visual tasks" and filtered by restricting the primary verb to physical actions. Based on these tasks, the original HowTo100M videos are searched and collected on YouTube. Each source video has corresponding narration subtitles, either handwritten or from the output of the Automatic Speech Recognition system. Different from the manually annotated datasets such as MSR-VTT [272], the video clips from HowTo100M are automatically captioned through narration, which could be thought of as *weakly paired*. Specifically, HowTo100M selects each line of video subtitles as one caption and pairs it with the corresponding time interval segmented video as the video clip. Generally, a source video can be segmented into 110 clip-caption pairs on average, with an average duration of 4 seconds per clip and four words per caption. This benchmark proposes an automatic data collection method to construct instructional video-caption pairs, which is faster and consumes fewer resources.

**VATEX** [251]: Compared to the previous widely-used large-scale datasets like MSR-VTT [272], VATEX is a larger, linguistically complex, and more diverse dataset benchmark for both video and text instruction description. Moreover, VATEX supports multilingual studies, and the video contents in VATEX are described in both English and Chinese. Specifically, this dataset benchmark covers 600 human activities and contains over 41,250 videos and 825,000 captions in two languages, and it also has over 206,000 English-Chinese parallel translation pairs. The source of its data collection is originally from Kinetics-600 [24], VATEX reuses 41,269 video clips from Kinetics-600 [24], and collects the corresponding bilingual descriptions for each clip via Amazon Mechanical Turk for English description and Bytedance Crowdsourcing for Chinese. The emergence of this dataset benchmark further supports the task of multilingual video and text description alignment and the task of video-guided machine translation.

**Webvid-2M** [12]: Compared to the prior benchmark datasets in video alignment with text instruction, WebVid-2M is an order of magnitude large-scale open-domain dataset comprising over 2,500,000 video-text pairs. The average length of the video clips from Webvid-2M is 18 seconds, with a total time length of 13,000 hours. The data of WebVid-2M is scraped from the web following a similar method of Google Conceptual Captions (CC3M) [210]. That is, first conduct the video-based filtering (i.e., keeping the specific video format data) and then the text-based filtering process (i.e., analyze candidate Alt-text using part-of-speech, sentiment/polarity and pornography/profanity annotations) to construct the dataset from the extracted candidates. In addition, the styles of text descriptions of

each video clip in Webvid-2M vary: both longer and poetic descriptions and succinct captions are present in this benchmark.

**InternVid** [254]: InternVid dataset contains over 7,000,000 videos lasting nearly 760K hours, resulting in 234,000,000 video clips with an average length of 10 seconds. The segmented video clips have detailed text descriptions of a total of 4.1B words, covering 16 scenarios and 6,104 motion descriptions collected from American Time Use Survey (ATUS) [22], public video datasets (i.e., Kinetics [93], UCF101 [218], etc.), and text corpus. Meanwhile, InternVid employs a multi-scale method with two types of captions. The finer scale descriptions are constructed by the lightweight image caption model Tag2Text [81], concentrating on the common objects, actions, and scene descriptions within the video clip. For the coarser scale, InternVid uses BLIP2 [119] to exclusively caption the middle frame of the clip. The videos from InternVid vary in language, length, and resolution. The videos are collected from countries with different languages, including the UK, USA, China, Japan, Korea, Russia, and France. Among these, 85% are high-resolution (i.e., 720P), with the rest ranging from 360p to 720p. For video duration, about 49% are under 5 minutes, 26% are 5-10 minutes, and only 8% exceed 20 minutes. Moreover, it also shows diverse clip durations and caption lengths according to the segmented clip level. That is, most video clips are 0-10 seconds long, accounting for 85% of the total, with half featuring captions of 10-20 words, one-third having fewer than ten words, and 11% containing more than 20 words. This large-scale, multilingual, and multi-level benchmark enhances the development of robust and transferable video-text representations for advanced multimodal understanding and generation.

**Panda-70M** [34]: Panda-70M uses an automatic method to construct a high-quality video-text instruction open-domain dataset. Specifically, Panda-70M collects 3.8M high-resolution videos from the HD-VILA-100M dataset [238], then splits these videos into semantically coherent clips via a two-stage semantics-aware splitting algorithm. This algorithm first cuts the video based on shot boundary detection [26] and then uses ImageBind [60] to extract video frames' embeddings and merge the adjacent clips if their embeddings are quite similar. Panda-70M also introduces *Max Running LPIPS*, counting the maximum perceptual similarity [291] among the given video clip to highlight the most significant perceptual change. Driven by the various original video captions from HD-VILA-100M dataset (i.e., useful texts and images), Panda-70M proposes to use multiple cross-modality teacher models, including Video-LLaMA [289], VideoChat [121], VideoChat Text [121], BLIP-2 [119], and MiniGPT4 [301], to obtain captions from each video clip. Based on the generated candidate captions for each video, Panda-70M further uses a fine-tuned retrieval model to select the one that best aligns the video content, thereby getting the whole 70M high-quality video-text caption pairs. This benchmark is designed to provide a large-scale but fine-grained dataset where the video captions can accurately express the video semantics content without ambiguity.

**VAST-27M** [33]: VAST-27M dataset explores the connections between videos and types of human instructions beyond text, including vision, audio, and subtitles. It selects 27M video clips from the HD-VILA-100M [238] dataset based on the clip length and the completeness of modalities (i.e., vision, audio, and subtitle) and covers over 15 categories, including music, gaming, education, entertainment, animals, etc. During construction of the VAST-27M dataset, a vision and audio captioner is trained to generate captions based on the input video clip. Specifically, the vision caption model is pre-trained on large-scale image-text corpora, including CC4M [271], CC12M [28], and LAION-400M [206]. Then, it is fine-tuned on manually labeled image and video caption datasets such as MSCOCO [131], VATEX [251], MSR-VTT [272], and MSVD [30]. This two-step training pipeline gives the vision caption model the capabilities for perceiving static objects and dynamic actions, thus generating high-quality captions. For the audio caption task, an audio caption model is trained using large-scale audio-text corpora, including VALOR-1M [32] and WavCaps [155] datasets. After leveraging vision and audio captioners to generate two types of captions from different modalities, an off-the-shelf Vicuna-13b model is used as the omni-modality captioner to generate the omni-modality caption. Vicuna-13b is an open-source large language model (LLM) trained through fine-tuning of LLaMA [232]. The omni-modality caption generated by Vicuna-13b is designed to effectively fuse the visual, audio, and subtitle contents while adhering to a natural human caption style at the same time. The purpose of Vicuna-13b is to fill in the gaps in training and evaluate the alignment between video and human instructions beyond text. It could further be employed in a wide range of multi-modal video-related tasks, including retrieval, captioning, and question-answering.

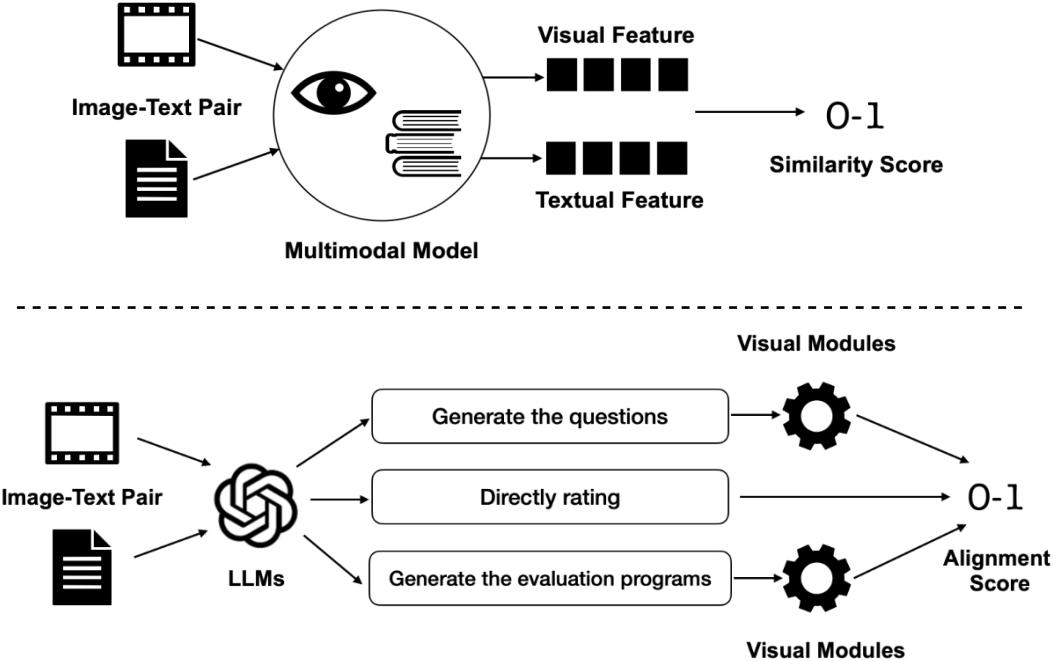


Figure 10: Compared with the previous multimodal evaluation backbone (top), the emergence of LLMs helps visual-textual alignment evaluation tasks evolve towards diversity, making the evaluation process more interpretable (bottom).

## 5.2 Evaluation Methods

In this section, we introduce several recent representative video-text alignment evaluation models. Different from the traditional evaluation metrics [8, 243, 14, 127, 179], which based solely on assessing the accuracy and fluency of generated video captions, the following studies consider both visual and textual elements, which could better align with the era of AI video generation. In addition, knowing that the current mainstream methods for modeling evaluation are frame-based, which measures the alignment between key video frame images and human instructions. The studies in the collection are different from those from benchmark evaluations, which directly use videos as visual representations. These modeling evaluation works use images as visual representations. As shown in Figure 10, early modeling evaluation studies primarily relied on multimodal pre-trained models as their core backbone. However, with recent advancements, there has been a shift towards the use of large language models (LLMs), which have become increasingly prominent in more recent evaluation methods.

### 5.2.1 Multimodal-based Methods

**TIGER** [88]: The Text-to-Image Grounding based metric for image caption Evaluation (TIGER) aims to mitigate the negative impact of evaluating only based on the text matching between reference captions and machine-generated captions in image-text alignment. TIGER assesses the alignment between text and image not only relies on the text but also takes image data into account, achieving a higher consistency with human judgments than the traditional rule-based metrics [8, 243, 14, 127, 179]. Specifically, the first stage of TIGER is *text-image grounding*, where TIGER computes a *grounding scores* for each image-text pair via a pre-trained Stacked Cross Attention Neural Network (SCAN) model [115]. Then, the TIGER framework goes through its second stage *grounding vector comparison*, where it computes the grounding vector  $s(V, C)$  between image  $V$  and the generated caption  $C$ , and the grounding vector  $s(V, R)$  between image  $V$  and the generated caption  $R$ . The higher the similarity between  $s(V, R)$  and  $s(V, C)$ , the higher the quality of  $C$  is. This quality of  $C$  is measured from two aspects: the first one evaluates the similarity in terms of image regions between these two vectors based on their grounding scores via Region Rank Similarity (RRS); the second one assesses the similarity in how attention is distributed across different regions of the image in the

two vectors, which is also indicated by grounding scores, via Weight Distribution Similarity (WDS). Then, the final TIGEr score is defined as the average value of the RRS and WDS scores.

**ViLBERTScore** [114]: ViLBERTScore is another evaluation metric that uses visiolinguistic representations rather than solely relying on textual representations that also demonstrate better alignment with human judgments than the traditional rule-based metrics [8, 243, 14, 127, 179]. ViLBERTScore first uses the ViLBERT model [144] to compute the contextual embeddings of generated caption  $\hat{x}$  and reference caption  $x$  with the target image  $I$ , respectively. Specifically, ViLBERTScore extracts the N region-level features  $V = (v_1, \dots, v_N)$  via the pre-trained object detection model [66] for each target image  $I$ , and feeds each pair of image and caption embeddings  $(X, V)$  and  $(\hat{X}, V)$  into pre-trained ViLBERT model to extract the text embedding segments  $H_{\hat{X}}^V$  and  $H_X^V$  from each entire output embedding. The pairwise cosine similarity between the generated caption and reference caption embeddings is then computed, and the final ViLBERTScore is formulated by the greedy matching results among the pair of tokens from reference and generated captions to find the most similar token-match segment based on the similarity score.

**COSMic** [84]: COHerence-Sensitive Metric of image captions (COSMic) introduces the first discourse-aware learned generation metric for evaluating image captions to mitigate the limitations that the prior existing metrics have struggled to differentiate reasonable generated image captions that deviate from the reference output in terms of goals or perspective, as the effective image descriptions may be various and present based on different goals and contexts. COSMic learns to accommodate diverse discourse goals without penalizing captions for different purposes from a new proposed *COIN dataset*, whose image descriptions are labeled with different coherence labels. Specifically, the COIN dataset contains 4,000 image-caption pairs collected from the Conceptual Captions (CC) training dataset [168]. The human annotators then select a coherence label for each pair from *Meta*, *Visible*, *Subjective*, and *Story*, and rate the quality of the captions. These collected data, named RaCCoOn (Ratings for Conceptual Caption), is the training dataset for coherence-aware captioning metric (i.e., COSMic), and the goal for metric training is to output a score for the generated caption given the image, reference caption, and the coherence-labels for both two captions. The metric has two different flavors: 1) a ViLBERT-based model pre-trained on large multimodal data and 2) a baseline Vanilla version. The former uses the pre-trained ViLBERT model [144], embedding both image and text inputs to take the vision features into account. Meanwhile, the latter independently embeds the input image and text using the BERT model and ResNet and later combines the output features for score computation. This metric introduces the coherence concept in training the image caption evaluation model, giving novel criteria closer to human judgments and more aligned with image contents.

**CLIPScore** [71]: CLIPScore is an evaluation metric based on the CLIP model [186], which is an efficient two-tower cross-modal model pre-trained on 400M image-caption pairs. CLIPScore uses the pre-trained CLIP model to fuse the image and text feature in the same dimension separately without references, and then compute the cosine similarity between the image visual CLIP embedding  $c$  and the candidate caption textual CLIP embedding  $v$ , as the corpus-level alignment score CLIP-S. In addition, the CLIPscore can further be extended to incorporate references (if available), named RefCLIPScore. RefCLIPScore first extracts the reference representation  $r$  through the CLIP model, and then is computed as a harmonic mean of CLIP-S and the maximal reference cosine similarity. This assessment method leverages the high-performing cross-modal pre-train models, offering a new approach to integrating vision and textual features to evaluate their content consistency.

**MID** [98]: Mutual Information Divergence (MID) also leverages the vision-and-language pre-trained model (i.e., the CLIP model) and uses the negative Gaussian cross-mutual information as a unified metric. MID first considers the continuous mutual information for condition  $x$  and the generation  $y$ , where the probability and joint probability distributions are multivariate Gaussian defined as the maximum entropy distribution for the given mean  $\mu$  and covariance  $\Sigma$ . Given this continuous mutual information, MID derives the point-wise mutual information for pair-wise evaluation (i.e., PMI). Then, MID further uses the expectation of PMI for the evaluating sample pair to measure the divergence from the reference samples, which becomes the final unified metric to measure the alignment of conditional generation. For the specific text-to-image generation alignment, MID uses the CLIP’s pre-trained image and text encoders [186] to encode both visual and textual information. This method bridges the gap by automatically measuring video-text alignment via mutual information (MI) based on multimodal models.

**PickScore** [101]: PickScore is an evaluation model similar to CLIPScore, and it also leverages the idea of the reward model objective from InstructGPT [176] in training. Before training the PickScore, a new dataset named Pick-a-Pic is built. Specifically, Pick-a-Pic is created by logging user interactions with the Pick-a-Pic web application for text-to-image generation, containing over 500,000 examples and 35,000 distinct prompts. Each instance in Pick-a-Pic has a prompt and two generated images, with a label representing the preferred image to reveal the real users’ preferences. PickScore follows the architecture of CLIP [186]: given a text prompt  $t$  and an image  $y$ , PickScore returns the inner product  $s(x, y)$  of the text embedding  $E_{txt}(x)$  and the image  $E_{img}(y)$  for training objectives. Then, PickScore is trained following the InstructGPT’s reward model objective that aims to minimize the natural human preference and the predicted preference distribution, which is computed by N-pair loss [217] using  $s(x, y_1)$  and  $s(x, y_2)$  for one training example. This work introduces a new large dataset with human preferences over user-prompted model-generated images, leveraging this feature to further fine-tune the CLIPScore-based evaluation model with the Reinforcement Learning from Human Feedback (RLHF) method.

**ImageReward** [270]: ImageReward is a general-purpose text-to-image human preference reward model that is able to encode human preferences. Specifically, the ImageReward uses BLIP [119] as the backbone model, and it is trained based on a systematic pipeline, including dataset collection and human annotation. The original data is collected via a diverse selection of real user prompts from DiffusionDB [259], resulting in 10,000 text prompt candidates (each of them is accompanied by 4-9 sampled images). Meanwhile, the human annotation process comprises three stages: Prompt Annotation, Text-Image Rating, and Text-Image Ranking. For each text prompt and its corresponding images, the annotators are asked to point out the missing part in the images, rating and ranking each image according to its alignment with the text prompt. As a result, the ImageReward model is trained on 8,878 valid text prompts and their 136,892 compared pairs via the reward model (RM) training [219, 176]. For the practical use of the ImageReward (i.e., as the metric for evaluating human preference on text-to-image models), the researcher annotation (i.e., by authors) is conducted across six popular high-resolution and available text-to-image models: CogView 2 [47], Versatile Diffusion [274], Stable Diffusion (both 1.4 and 2.1-base [198]), DALL-E 2 [189], and Openjourney [182]. This work extends the RLHF method in multimodal evaluation model tuning to a general and more direct optimization way.

**TITScore** [85]: TITScore solves the *long-tailed effect* in the existing text-to-image evaluation metrics, such as CLIPScore [72]. This issue arises from the non-essential elements in the text prompts, leading to a distinct difference between the knowledge representation and the embedding dimensionality when encoding the entire text prompt. To solve this problem, TITScore uses a symbolic-level understanding evaluation paradigm by explicitly embedding mixture-of-experts (MOEs) large vision models (LVMs) while maintaining the neuro-level reasoning capability. Specifically, TITScore integrates three primary components: prompt curation, MOE, and knowledge gathering. In the prompt curation process, a classifier model (i.e., a pre-trained Robustly Optimized BERT model [140]) yields a set of related evaluation aspects based on the given input prompt. The original prompt embedding  $E_p$  will be integrated with the embeddings of the decomposed tokens for each identified aspect  $E_D$  to maximize the conditional probability  $p(E_D|E_p)$ , thereby enhancing the final prompt curation output. Then, in the MOE process, the evaluation aspects will be divided into two paradigms (i.e., the explicit symbolic level and the implicit neuro-level). TITScore uses a variety of visual reasoning models (e.g., segmentation model [100] and detection model [175]) for the explicit symbolic level evaluation and adapts multimodal models (e.g., fine-tuning an adapter after the ViTLarge model [50]) for the implicit neuro-level assessment. The final TITScore is computed by comparing the MOE evaluation embeddings with the curated prompt embeddings. In addition, a new benchmark dataset *TITBench* is proposed to facilitate the semantically rich text-to-visual evaluation studies, which contains over 2,400 diverse prompts across 16 evaluation aspects, including alignment, category, etc.

### 5.2.2 LLM-based Methods

**GPT-4V Eval** [293]: GPT-4V Eval explores the potential of GPT-4V [1], a high-performance multi-modal transformer language model, in evaluating the vision-language tasks. GPT-4V Eval introduces two evaluation pipelines, *single-answer grading* and *pairwise comparison*, to systematically validate the capabilities of GPT-4V as an evaluator and how well it aligns with human performance in vision-language tasks, including image-to-text captioning, text-to-image generation, text-guided image editing, and multiple images to text alignment tasks. Specifically, for *single-answer grading*,

GPT-4V is asked to generate a score on a scale of 1-100 to evaluate the quality and alignment of the input-output pair. Meanwhile, in *pairwise comparison*, GPT-4V is asked to generate an answer to determine the best choice (select one or response 'Tie') of a pair of generated outputs. This work validates the effectiveness of LLMs (especially the GPT-4) as the evaluator, relying on its content comprehension and reasoning capabilities to directly score the alignment of text-image pairs.

**LLMScore** [148]: LLMScore leverages the powerful reasoning capabilities of large language models (LLMs) to assess the alignment between the images and captions. Inspired by the human evaluation process of measuring image-text alignment, LLMScore aims to imitate human decisions' key points, such as checking the correctness of objects and the specific attributes in the generated image based on the given text prompt. Specifically, LLMScore first uses image captioning model BLIPv2 [119] and GRiT [268] as the Multi-Granularity Visual Descriptors, transforming the image into multi-granularity descriptions (i.e., image-level global and object-level local) to capture the compositional objects in text format. Then, the above descriptions and the text prompts will be fed into LLMs, such as GPT-4 [174], to serve as a Text-to-Image Evaluator for reasoning the consistency between the text prompts and images. Given the visual descriptions above, LLM first rates the quality based on the specific instructions regarding overall semantics, error counting, etc. LLMScore combines all the rating results from different aspects and finally derives the final evaluation score. This work leverages the LLMs as text-image alignment evaluators as well. Meanwhile, it focuses more on the multi-granularity of the image compositionality when rating.

**VIEScore** [107]: VIEScore also leverages the multimodal large language models (MLLMs) as the backbone evaluators but focuses more on the reasoning explainability and task awareness of the evaluation framework. Specifically, VIEScore proposes a rating instruction based on the design of human evaluation scores from ImagenHub, [110], considering both the semantic consistency (SC) and perceptual quality (PQ), and the SC contains multiple types of scores according to the specific tasks. These rating instructions, along with the text and image (based on the specific tasks), will be fed into the MLLM together, and then the generated responses will be parsed and yield the sub-scores for SC and PQ access, respectively. VIEScore assumes each sub-score weights the same, deriving the root of the product of the SC score and PQ score as the final evaluation score. VIEScore employs various MLLMs, including GPT-4o [5], GPT-4v [1], Gemini-Pro [194], LLaVA [133], as the evaluation backbone and conducts empirical experiments on ImagenHub human evaluation dataset [110].

**TIFA** [79]: Text-to-Image Faithfulness evaluation with question Answering (TIFA) measures the faithfulness of a generated image to its text prompt based on the questioning generation and answering process (QG/A). Specifically, TIFA generates several binary question-answering pairs based on the given text prompt via LLMs. Then, the generated image's faithfulness and alignment are calculated by checking whether the VQA system can answer the corresponding question. The final TIFA metric score is defined as the mean value of the number of correct answers provided by the VQA models. In practice, TIFA uses GPT-3 as the element extraction and question generation model, then the generated questions are verified via UnifiedQA [97], filtering out the ones that UnifiedQA can not agree with. In addition, TIFA uses open-domain pre-trained vision-language models [116, 247, 99, 248] as the VQA models to answer the generated questions, as they contain various visual elements (e.g., activity, art style). Based on TIFA, a new benchmark dataset, TIFA v1.0 is also proposed, which contains 4k diverse text descriptions and 25K questions across 12 categories.

**VQ2** [279]: VQ2 is an image-text alignment evaluation approach that is also based on question generation and answering (QG/A). Unlike TIFA, which compares the contents of the text-image pair, VQ2 checks if the textual answer is accurate based on the image in the QA process. For example, given an image-text pair (i.e.,  $\{I, T\}$ ), VQ2 first extracts a set of candidate answers for the given text  $T$  via SpaCy [77], then uses a QG model (i.e., a T5-XXL model [187] fine-tuned on SQuAD1.0 [188]) to generate a question based on each candidate answer. Each generated question-answer pair  $(q_j, a_j)$  will be re-written into a new yes-no format  $q'_j$  (i.e., "is  $\{a_j\}$  true for  $\{q_j\}$ "). The VQA (i.e., PaLI-17B model [35]) model then gives an answer to  $q'_j$ . VQ2 defines the alignment score  $s_j = VQA(q_j, q'_j, I)$  as the probability of the model for answering 'yes', and the final VQ2 score is denoted as the average value over all  $s_j$  scores for all the generated question-answer pairs. In addition to VQ2, a new evaluation dataset *SeeTrue* is also proposed to facilitate the studies in text-to-image generation evaluation studies, which encompasses 31,855 real and synthetic images and text examples with human judgments for whether a given image-text pair is semantically aligned.

**DSG** [38]: Davidsonian Scene Graph (DSG) is an empirically grounded evaluator that is also based on the question generation and answering (QG/A) framework. DSG follows the QG/A methodology like TIFA [79], and VQ2 [279] but applies the idea from formal semantics [45] to address several reliability issues (i.e., duplicated and non-atomic question). Specifically, in text-to-image alignment evaluation, DSG first goes through the question generation (QG) process, generating atomic and unique questions (i.e., each question covers the smallest possible semantic unit) from the text prompts. These generated questions should cover all contents and only the contents of the given text prompt, and each question’s content should be unique. DSG then conducts the question-answering (QA) process, answering the generated questions based on the corresponding image via the VQA models. In practice, DSG implements the QG step using a Directed Acyclic Graph (DAG), where each node here represents an atomic question, and each directed edge represents the semantic dependencies between the questions. DSG leverages the high-performance LLMs, including PaLM2 [9] and GPT3.5/4 [5], for the QG stage. In the QA stage, the questions are processed by the VQA modules [116, 35, 43] according to the given DAG. To further facilitate the studies in text-to-image alignment evaluation, a fine-grained human-annotated benchmark dataset DSG-1k is also proposed to conduct the empirical experiments; DSG-1k is based on the dataset TIFA v1.0 but contains a more well-balanced mix of semantic categories and styles.

**VPEVAL** [39]: VPEVAL is an evaluation framework for text-to-image generation based on interpretable visual programming. Specifically, VPEVAL generates evaluation programs with LLMs via in-context learning, and the evaluation programs can be 1) skill-based evaluation and 2) open-ended evaluation. In skill-based evaluation, VPEVAL defines five image generation skills (i.e., object, count, spatial, scale, and text rendering) and creates a set of skill-specific text prompts. Given these prompts, the evaluation programs measure the text-image alignment scores (in binary form) by calling the related visual modules. In open-ended evaluation, VPEVAL uses a diverse of skill prompts and leverages LLMs (i.e., ChatGPT [171]) to generate the corresponding evaluation program for each skill prompt dynamically, and then adapts a set of specific visual evaluation modules for different tasks for each evaluation program measurement. For example, the object detection models Grounding DINO [136] and DPT [190] are used as the modules for object skill measurement, and the BLIP-2 [119] is used as the visual question answering module. This work provides a new perspective in textual and visual alignment evaluation, leveraging programming to invoke diverse visual modules to evaluate diverse image generation skills.

## 6 Future Prospects

As we stand on the point of significant advancements in AIGVE, we are facing numerous opportunities and unidentified challenges. This section outlines several key future prospects that aim to assist the development of this emerging field.

**Integrating Vision Language Models:** As Vision Language Models (VLMs) such as Qwen [11], LLaVA [132], and Chameleon [227] become increasingly sophisticated, their support for video inputs marks a significant advancement. These models leverage transformer [242] architectures, which enable them to process complex interactions within video frames and between accompanying textual data efficiently. This capability allows VLMs to interpret dynamic visual content while correlating it with relevant text, enhancing their ability to evaluate alignment with human perception and instructions accurately.

Moreover, the ability of VLMs to process text inputs opens opportunities for creating on-demand video evaluation models. These models could be tailored to assess general qualities or focus on specific attributes based on the user’s prompts. This capability enhances the versatility and utility of VLMs in diverse applications, from research to professional media production.

**Improving Score Interpretability:** As the evaluation methods for video quality transition from utilizing collections of scores to deploying single models that generate a unified score, the inherent "black box" nature of neural networks presents both challenges and opportunities [62]. Researchers must carefully navigate the trade-offs between model integration and score interpretability [58, 249]. Some recent research starts the early exploration of this field [79, 147, 56, 108]. This transparency not only aids in debugging and refining the models but also ensures that the outcomes are transparent, justifiable, and reproducible in diverse applications, especially as they potentially could become widely applied on platforms that determine creators’ rights and revenues. Enhancing the clarity

of how scores are derived can help maintain trust and fairness, ensuring that these models support creators fairly and consistently across various platforms.

**Ethical and Safety Considerations:** As AI-generated content becomes increasingly commonplace, the importance of embedding ethical standards and safety mechanisms within AIGVE frameworks cannot be overstated. These frameworks must be designed to actively prevent the propagation of biases or misinformation, which can have significant societal impacts. This involves implementing robust algorithms that can detect and mitigate biased data inputs or skewed results that may perpetuate stereotypes or unequal representations.

Moreover, the training data for these models often come from human-scored inputs, which inherently contain subjective biases based on the scorers' backgrounds and experiences [221, 161]. To enhance the fairness and inclusivity of AI evaluations, it is crucial to diversify the sources of human-scored data and implement strategies that can identify and correct biases in the training datasets. This might include using techniques such as stratified sampling to ensure that training data covers a wide range of demographics and viewpoints [52, 90], as well as employing fairness-aware machine learning algorithms [304] that can adjust for identified biases. Additionally, careful collection and curation of training data are essential. By deliberately selecting data that represent diverse user interactions and contexts, developers can further minimize the risk of biased outcomes.

In addition to bias mitigation, ethical AIGVE frameworks should incorporate privacy protections, especially when handling sensitive or personal content. Ensuring data anonymization and securing user data against unauthorized access are vital steps in maintaining user trust and complying with global privacy regulations.

## 7 Conclusion

This survey highlights the importance of AI-Generated Video Evaluation (AIGVE) as a distinct research area, focusing on aligning AI-generated videos with human perception and instructions. By reviewing existing methodologies from video quality assessment, multimodal text-visual alignment, and recent comprehensive evaluation approaches, we provide a structured overview of the current landscape. As AI-generated video technology continues to advance, there is a critical need for developing more robust evaluation frameworks that effectively capture the complexity of both spatial and temporal dimensions in video content while ensuring alignment with human needs. We hope this survey serves as a foundational resource for researchers, supporting the advancement of this evolving field.

## References

- [1] Gpt-4v(ision) system card. 2023. URL <https://api.semanticscholar.org/CorpusID:263218031>.
- [2] Pika, 2023.
- [3] Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023.
- [4] Zeroscope, 2023.
- [5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [6] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans, 2018. URL <https://arxiv.org/abs/1810.01325>.
- [7] Yazan Albadarin, Mohammad Saqr, Nathan Pope, et al. A systematic literature review of empirical research on chatgpt in education. *Discovery Education*, 3:60, 2024. doi: 10.1007/s44217-024-00138-2. URL <https://doi.org/10.1007/s44217-024-00138-2>.
- [8] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

- [9] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [10] Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Aleksandr Gushchin, Dmitriy Vatolin, and Dmitriy Kulikov. Video compression dataset and benchmark of learning-based video-quality metrics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13814–13825. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/59ac9f01ea2f701310f3d42037546e4a-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/59ac9f01ea2f701310f3d42037546e4a-Paper-Datasets_and_Benchmarks.pdf).
- [11] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [12] Max Bain, Arsha Nagrani, Gü̈l 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*, pages 1728–1738, 2021.
- [13] Christos G. Bampis, Zhi Li, and Alan C. Bovik. Spatiotemporal feature integration and model fusion for full reference video quality assessment, 2018. URL <https://arxiv.org/abs/1804.04813>.
- [14] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- [15] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.
- [16] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. URL <https://arxiv.org/abs/2401.12945>.
- [17] Ismail Bezzine, Zohaib Amjad Khan, Azeddine Beghdadi, Noor Al-Maadeed, Mounir Kaaniche, Somaya Al-Maadeed, Ahmed Bouridane, and Faouzi Alaya Cheikh. Video quality assessment dataset for smart public security systems. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–5, 2020. doi: 10.1109/INMIC50486.2020.9318149.
- [18] Prateep Bhattacharjee and Sukhendu Das. Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks. *advances in neural information processing systems*, 30, 2017.
- [19] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL <https://arxiv.org/abs/2311.15127>.
- [20] 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, 2023. URL <https://arxiv.org/abs/2304.08818>.
- [21] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, January 2018. ISSN 1941-0042. doi: 10.1109/tip.2017.2760518. URL <http://dx.doi.org/10.1109/TIP.2017.2760518>.

- [22] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.
- [23] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL <https://arxiv.org/abs/2104.14294>.
- [24] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.
- [25] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.
- [26] Brandon Castellano. PySceneDetect. URL <https://github.com/Breakthrough/PySceneDetect>.
- [27] Jorge Caviedes and Franco Oberti. No-reference quality metric for degraded and enhanced video. volume 5150, pages 621–632, 01 2003. ISBN 978-0-8247-2777-2. doi: 10.1201/9781420027822.ch10.
- [28] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [29] Baoliang Chen, Lingyu Zhu, Guo Li, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment, 2021. URL <https://arxiv.org/abs/2012.13936>.
- [30] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <https://aclanthology.org/P11-1020>.
- [31] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. URL <https://arxiv.org/abs/2401.09047>.
- [32] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.
- [33] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36, 2024.
- [34] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.
- [35] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [36] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos, 2024. URL <https://arxiv.org/abs/2406.06087>.
- [37] Iya Chivileva, Philip Lynch, Tomas E. Ward, and Alan F. Smeaton. Measuring the Quality of Text-to-Video Model Outputs: Metrics and Dataset, September 2023. URL <http://arxiv.org/abs/2309.08009> [cs].
- [38] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.
- [39] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

- [40] Lark Kwon Choi and Alan C. Bovik. Flicker sensitive motion tuned video quality assessment. In *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 29–32, 2016. doi: 10.1109/SSIAI.2016.7459167.
- [41] Marcos V. Conde, Saman Zadtootaghaj, Nabajeet Barman, Radu Timofte, Chenlong He, Qi Zheng, Ruoxi Zhu, Zhengzhong Tu, Haiqiang Wang, Xiangguang Chen, Wenhui Meng, Xiang Pan, Huiying Shi, Han Zhu, Xiaozhong Xu, Lei Sun, Zhenzhong Chen, Shan Liu, Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, Wei Sun, Yuqin Cao, Yanwei Jiang, Jun Jia, Zhichao Zhang, Zijian Chen, Weixia Zhang, Xiongkuo Min, Steve Göring, Zihao Qi, and Chen Feng. Ais 2024 challenge on video quality assessment of user-generated content: Methods and results, 2024. URL <https://arxiv.org/abs/2404.16205>.
- [42] Timothee Cour, Chris Jordan, Eleni Miltzakaki, and Ben Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part IV 10*, pages 158–171. Springer, 2008.
- [43] 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. URL <https://arxiv.org/abs/2305.06500>.
- [44] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013.
- [45] Donald Davidson. Theories of meaning and learnable languages. *Inquiries into truth and interpretation*, pages 3–16, 2001.
- [46] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL <https://arxiv.org/abs/2105.05233>.
- [47] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.
- [48] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers, 2022. URL <https://arxiv.org/abs/2204.14217>.
- [49] Yunlong Dong, Xiaohong Liu, Yixuan Gao, Xunchu Zhou, Tao Tan, and Guangtao Zhai. Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement, 2023. URL <https://arxiv.org/abs/2305.09512>.
- [50] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL <https://arxiv.org/abs/2010.11929>.
- [52] Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. *ArXiv*, abs/2105.11570, 2021.
- [53] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL <https://arxiv.org/abs/2403.03206>.
- [54] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhui Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024.
- [55] Gereon Fox, Ayush Tewari, Mohamed Elgarib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan, 2021. URL <https://arxiv.org/abs/2107.07224>.

- [56] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.
- [57] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL <https://arxiv.org/abs/2203.13131>.
- [58] Lei Gao and Ling Guan. Interpretability of machine learning: Recent advances and future prospects, 2023. URL <https://arxiv.org/abs/2305.00537>.
- [59] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer, 2022. URL <https://arxiv.org/abs/2204.03638>.
- [60] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.
- [61] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency, 2022. URL <https://arxiv.org/abs/2108.06858>.
- [62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [63] Fengbin Guan, Xin Li, Zihao Yu, Yiting Lu, and Zhibo Chen. Q-mamba: On first exploration of vision mamba for image quality assessment, 2024.
- [64] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.
- [65] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.
- [66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [67] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhui Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252, 2024. URL <https://arxiv.org/abs/2406.15252>.
- [68] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022.
- [69] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. doi: 10.1109/5254.708428.
- [70] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.
- [71] 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 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL <https://aclanthology.org/2021.emnlp-main.595>.
- [72] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL <https://arxiv.org/abs/2104.08718>.

- [73] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL <https://arxiv.org/abs/1706.08500>.
- [74] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. URL <https://arxiv.org/abs/2210.02303>.
- [75] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL <https://arxiv.org/abs/2204.03458>.
- [76] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. URL <https://arxiv.org/abs/2205.15868>.
- [77] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.
- [78] 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 *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017. doi: 10.1109/QoMEX.2017.7965673.
- [79] 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, 2023. URL <https://arxiv.org/abs/2303.11897>.
- [80] Ding-Jiun Huang, Yu-Ting Kao, Tieh-Hung Chuang, Ya-Chun Tsai, Jing-Kai Lou, and Shuen-Huei Guan. Sb-vqa: A stack-based video quality assessment framework for video enhancement, 2023. URL <https://arxiv.org/abs/2305.08408>.
- [81] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.
- [82] 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, November 2023. URL <http://arxiv.org/abs/2311.17982>. arXiv:2311.17982 [cs].
- [83] 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, November 2023. URL <http://arxiv.org/abs/2311.17982>. arXiv:2311.17982 [cs].
- [84] Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. Cosmic: a coherence-aware generation metric for image descriptions. *arXiv preprint arXiv:2109.05281*, 2021.
- [85] Pengliang Ji and Junchen Liu. Tltscore: Towards long-tail effects in text-to-visual evaluation with generative foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5302–5313, 2024.
- [86] Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5325–5335, June 2024.
- [87] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning, 2024. URL <https://arxiv.org/abs/2405.01483>.
- [88] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019.

- [89] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis, 2023. URL <https://arxiv.org/abs/2303.05511>.
- [90] Umutcan Karakas and Ayse Tosun Misirli. Automated fairness testing with representative sampling. *Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2023. doi: 10.1145/3617555.3617871.
- [91] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. URL <https://arxiv.org/abs/1710.10196>.
- [92] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL <https://arxiv.org/abs/1812.04948>.
- [93] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [94] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL <https://arxiv.org/abs/2108.05997>.
- [95] Christian Keimel, Tobias Oelbaum, and Klaus Diepold. No-reference video quality evaluation for high-definition video. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1145–1148, 2009. doi: 10.1109/ICASSP.2009.4959791.
- [96] Levon Khachatryan, Andranik Moysisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.
- [97] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.
- [98] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35:35072–35086, 2022.
- [99] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
- [100] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [101] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- [102] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. URL <https://arxiv.org/abs/2312.14125>.
- [103] Anton Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, January 2023. doi: 10.1257/jel.20231736. URL <https://www.aeaweb.org/articles?id=10.1257/jel.20231736>.
- [104] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM ’23, page 1066–1076, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3611860. URL <https://doi.org/10.1145/3581783.3611860>.

- [105] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment, March 2024. URL <http://arxiv.org/abs/2403.11956>. arXiv:2403.11956 [cs].
- [106] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment, May 2024. URL <http://arxiv.org/abs/2403.11956>. arXiv:2403.11956 [cs].
- [107] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhua Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.
- [108] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhua Chen. VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation, December 2023. URL <http://arxiv.org/abs/2312.14867>. arXiv:2312.14867 [cs].
- [109] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhua Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2024. URL <https://arxiv.org/abs/2312.14867>.
- [110] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhua Chen. Imagenhub: Standardizing the evaluation of conditional image generation models, 2024. URL <https://arxiv.org/abs/2310.01596>.
- [111] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.
- [112] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.
- [113] Dae Yeol Lee, Somdyuti Paul, Christos G. Bampis, Hyunsuk Ko, Jongho Kim, Se Yoon Jeong, Blake Homan, and Alan C. Bovik. A subjective and objective study of space-time subsampled video quality. *IEEE Transactions on Image Processing*, 31:934–948, 2022.
- [114] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, 2020.
- [115] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.
- [116] 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.
- [117] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia, MM ’19*. ACM, October 2019. doi: 10.1145/3343031.3351028. URL <http://dx.doi.org/10.1145/3343031.3351028>.
- [118] 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, January 2021. ISSN 1573-1405. doi: 10.1007/s11263-020-01408-w. URL <http://dx.doi.org/10.1007/s11263-020-01408-w>.
- [119] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [120] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL <https://arxiv.org/abs/2301.12597>.
- [121] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

- [122] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023.
- [123] Xudong Li, Jingyuan Zheng, Xiawu Zheng, Runze Hu, Enwei Zhang, Yuting Gao, Yunhang Shen, Ke Li, Yutao Liu, Pingyang Dai, Yan Zhang, and Rongrong Ji. Less is more: Learning reference knowledge using no-reference image quality assessment, 2023. URL <https://arxiv.org/abs/2312.00591>.
- [124] Xuelong Li, Qun Guo, and Xiaoqiang Lu. Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, 25(7):3329–3342, 2016. doi: 10.1109/TIP.2016.2568752.
- [125] Y. Li, S. Meng, X. Zhang, S. Wang, Y. Wang, and S. Ma. UGC-VIDEO: Perceptual Quality Assessment of User-Generated Videos. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38, Shenzhen, China, 2020. doi: 10.1109/MIPR49039.2020.00015.
- [126] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*, 2024.
- [127] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- [128] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [129] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model, 2024. URL <https://arxiv.org/abs/2404.09967>.
- [130] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers, 2021. URL <https://arxiv.org/abs/2106.04554>.
- [131] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [132] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [133] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.
- [134] Hongbo Liu, Mingda Wu, Kun Yuan, Ming Sun, Yansong Tang, Chuanchuan Zheng, Xing Wen, and Xiu Li. Ada-dqa: Adaptive diverse quality-aware feature acquisition for video quality assessment, 2023. URL <https://arxiv.org/abs/2308.00729>.
- [135] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos “in the wild”. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1996–2003. IEEE, 2009.
- [136] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [137] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, page 546–554, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240643. URL <https://doi.org/10.1145/3240508.3240643>.
- [138] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment, 2017. URL <https://arxiv.org/abs/1707.08347>.
- [139] 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.

- [140] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [141] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player?, April 2024. URL <http://arxiv.org/abs/2307.06281> [cs]. arXiv:2307.06281 [cs].
- [142] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL <https://arxiv.org/abs/2103.14030>.
- [143] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL <https://arxiv.org/abs/2201.03545>.
- [144] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [145] Yiting Lu, Xin Li, Bingchen Li, Zihao Yu, Fengbin Guan, Xinrui Wang, Ruling Liao, Yan Ye, and Zhibo Chen. Aigc-vqa: A holistic perception metric for aigc video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6384–6394, June 2024.
- [146] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos, 2024. URL <https://arxiv.org/abs/2402.07220>.
- [147] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation, May 2023. URL <http://arxiv.org/abs/2305.11116>. arXiv:2305.11116 [cs].
- [148] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [149] 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, 2023. URL <https://arxiv.org/abs/2303.08320>.
- [150] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, August 2017. ISSN 1941-0042. doi: 10.1109/TIP.2017.2708503. URL <http://dx.doi.org/10.1109/TIP.2017.2708503>.
- [151] Pavan C. Madhusudana, Xiangyu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Subjective and objective quality assessment of high frame rate videos. *IEEE Access*, 9: 2169–3536, 2021.
- [152] K. Manasa and Sumohana S. Channappayya. An optical flow-based full reference video quality assessment algorithm. *IEEE Transactions on Image Processing*, 25(6):2480–2492, 2016. doi: 10.1109/TIP.2016.2548247.
- [153] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention, 2016. URL <https://arxiv.org/abs/1511.02793>.
- [154] M. Masry, S.S. Hemami, and Y. Sermadevi. A scalable wavelet-based video distortion metric and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(2): 260–273, 2006. doi: 10.1109/TCSVT.2005.861946.
- [155] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [156] The Movie Gen Team @ Meta. Movie gen: A cast of media foundation models. <https://ai.meta.com/static-resource/movie-gen-research-paper>, 2024. Accessed: 2024-10-14.

- [157] Yibo Miao, Yifan Zhu, Yinpeng Dong, Lijia Yu, Jun Zhu, and Xiao-Shan Gao. T2vsafetybench: Evaluating the safety of text-to-video generative models, 2024. URL <https://arxiv.org/abs/2407.05965>.
- [158] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.
- [159] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual Video Quality Assessment: A Survey, February 2024. URL <http://arxiv.org/abs/2402.03413>. arXiv:2402.03413 [cs, eess].
- [160] Shankhanil Mitra and Rajiv Soundararajan. Knowledge guided semi-supervised learning for quality assessment of user generated videos, 2023. URL <https://arxiv.org/abs/2312.15425>.
- [161] Gabriel Mittag, Saman Zadtootaghaj, Thilo Michael, Babak Naderi, and S. Möller. Bias-aware loss for training image and speech quality prediction models from multiple datasets. *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 97–102, 2021. doi: 10.1109/QoMEX51781.2021.9465384.
- [162] Gabriel Mittag, Babak Naderi, Vishak Gopal, and Ross Cutler. Lstm-based video quality prediction accounting for temporal distortions in videoconferencing calls. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10095711.
- [163] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.
- [164] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.
- [165] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.
- [166] Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.
- [167] Anish Mittal, Michele A. Saad, and Alan C. Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2016. doi: 10.1109/TIP.2015.2502725.
- [168] Edwin G Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020.
- [169] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL <https://arxiv.org/abs/2112.10741>.
- [170] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen. CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, July 2016. doi: 10.1109/TIP.2016.2562513.
- [171] OpenAI. Chatgpt. <https://openai.com/research/chatgpt>, 2023.
- [172] OpenAI. Gpt-4 technical report, 2023.
- [173] OpenAI. Sora. <https://openai.com/index/sora/>, 2024.
- [174] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman,

Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.

- [175] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [176] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [177] Katerina Pandremenou, Muhammad Shahid, Lisimachos P Kondi, and Benny Lövström. A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses. In *Human vision and electronic imaging XX*, volume 9394, pages 486–497. SPIE, 2015.
- [178] 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 on Association for Computational Linguistics, ACL ’02*, page 311–318, USA, 2002. As-

- sociation for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.
- [179] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. URL <https://api.semanticscholar.org/CorpusID:11080756>.
  - [180] Yajing Pei, Shiyu Huang, Yiting Lu, Xin Li, and Zhibo Chen. Priorformer: A ugc-vqa method with content and distortion priors, 2024. URL <https://arxiv.org/abs/2406.16297>.
  - [181] Peng Peng, Danping Liao, and Ze-Nian Li. An efficient temporal distortion measure of videos based on spacetime texture. *Pattern Recognition*, 70:1–11, 2017. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2017.04.031>. URL <https://www.sciencedirect.com/science/article/pii/S0031320317301826>.
  - [182] PromptHero. Openjourney. <https://openjourney.art>, 2023.
  - [183] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription, 2019. URL <https://arxiv.org/abs/1903.05854>.
  - [184] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.
  - [185] Bowen Qu, Xiaoyu Liang, Shangkun Sun, and Wei Gao. Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap, 2024. URL <https://arxiv.org/abs/2404.13573>.
  - [186] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
  - [187] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL <https://arxiv.org/abs/1910.10683>.
  - [188] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
  - [189] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022. URL <http://arxiv.org/abs/2204.06125> [cs]. arXiv:2204.06125 [cs].
  - [190] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
  - [191] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013.
  - [192] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016. URL <https://arxiv.org/abs/1605.05396>.
  - [193] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.
  - [194] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricu, Angeliki Lazaridou, Orhan Firat, Julian Schrittweiser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
  - [195] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.
  - [196] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer, 2014.

- [197] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.
- [198] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [199] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL <https://arxiv.org/abs/2112.10752>.
- [200] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL <http://arxiv.org/abs/2112.10752>. arXiv:2112.10752 [cs].
- [201] runway. Gen-2. <https://runwayml.com/research/gen-2>, 2024.
- [202] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014. doi: 10.1109/TIP.2014.2299154.
- [203] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL <https://arxiv.org/abs/2205.11487>.
- [204] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. URL <https://arxiv.org/abs/1606.03498>.
- [205] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.
- [206] C Schuhmann, R Vencu, R Beaumont, R Kaczmarczyk, C Mullis, A Katta, T Coombes, J Jitsev, and A Laion Komatsuzaki. 400m: Open dataset of clip-filtered 400 million image-text pairs. arxiv 2021. *arXiv preprint arXiv:2111.02114*.
- [207] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [208] Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31:1027–1041, 2022. doi: 10.1109/TIP.2021.3136723. Epub 2022 Jan 11.
- [209] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [210] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [211] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3441–3452, 2006. doi: 10.1109/TIP.2006.881959.
- [212] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. URL <https://arxiv.org/abs/2209.14792>.
- [213] Zeina Sinno and Alan C. Bovik. Spatio-temporal measures of naturalness. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1750–1754, 2019. doi: 10.1109/ICIP.2019.8803115.

- [214] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019. doi: 10.1109/TIP.2018.2869673.
- [215] Jacob Søgaard, Søren Forchhammer, and Jari Korhonen. No-reference video quality assessment using codec analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 25:1637–1650, 2015. URL <https://api.semanticscholar.org/CorpusID:11290054>.
- [216] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL <https://arxiv.org/abs/1503.03585>.
- [217] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.
- [218] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [219] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- [220] Robert C. Streijl, Stefan Winkler, and David S. Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016. doi: 10.1007/s00530-014-0446-1.
- [221] Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *ArXiv*, abs/2311.09730, 2023. doi: 10.48550/arXiv.2311.09730.
- [222] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL <https://arxiv.org/abs/2406.06525>.
- [223] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos, 2021. URL <https://arxiv.org/abs/2106.01111>.
- [224] 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*, MM ’22. ACM, October 2022. doi: 10.1145/3503161.3548329. URL <http://dx.doi.org/10.1145/3503161.3548329>.
- [225] Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhichao Zhang, Linhan Cao, Qiubo Chen, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Enhancing blind video quality assessment with rich quality-aware features, 2024. URL <https://arxiv.org/abs/2405.08745>.
- [226] Qu Tang, Xiangyu Zhu, Zhen Lei, and Zhaoxiang Zhang. Intrinsic physical concepts discovery with object-centric predictive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23252–23261, 2023.
- [227] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL <https://arxiv.org/abs/2405.09818>.
- [228] The ModelScope Team. Modelscope: bring the notion of model-as-a-service to life. <https://github.com/modelscope/modelscope>, 2023.
- [229] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. URL <https://arxiv.org/abs/2003.12039>.
- [230] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis, 2021. URL <https://arxiv.org/abs/2104.15069>.
- [231] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL <http://arxiv.org/abs/2302.13971>. arXiv:2302.13971 [cs].

- [232] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillerm Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madiam Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mi-haylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- [233] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. A comparative evaluation of temporal pooling methods for blind video quality assessment, 2020. URL <https://arxiv.org/abs/2002.10651>.
- [234] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. A comparative evaluation of temporal pooling methods for blind video quality assessment, 2020. URL <https://arxiv.org/abs/2002.10651>.
- [235] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. ISSN 1941-0042. doi: 10.1109/tip.2021.3072221. URL <http://dx.doi.org/10.1109/TIP.2021.3072221>.
- [236] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. ISSN 2644-1322. doi: 10.1109/ojsip.2021.3090333. URL <http://dx.doi.org/10.1109/OJSP.2021.3090333>.
- [237] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation, 2017. URL <https://arxiv.org/abs/1707.04993>.
- [238] 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 preprint arXiv:1812.01717*, 2018.
- [239] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. URL <https://openreview.net/forum?id=rylgEULtdN>.
- [240] Giuseppe Valenzise, Stefano Magni, Marco Tagliasacchi, and Stefano Tubaro. No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(4):605–618, 2012. doi: 10.1109/TCSVT.2011.2171211.
- [241] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL <https://arxiv.org/abs/1711.00937>.
- [242] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [243] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL <https://arxiv.org/abs/1411.5726>.
- [244] Maria Torres Vega, Decebal Constantin Mocanu, and Antonio Liotta. Predictive no-reference assessment of video quality, 2016. URL <https://arxiv.org/abs/1604.07322>.
- [245] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022. URL <https://arxiv.org/abs/2210.02399>.
- [246] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics, 2016. URL <https://arxiv.org/abs/1609.02612>.

- [247] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.
- [248] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022.
- [249] Ruochen Wang, Si Si, Felix Yu, Dorothea Wiesmann, Cho-Jui Hsieh, and Inderjit Dhillon. Large language models are interpretable learners, 2024. URL <https://arxiv.org/abs/2406.17224>.
- [250] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *arXiv preprint arXiv:2403.06098*, 2024.
- [251] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019.
- [252] Xinyi Wang, Angeliki Katsenou, and David Bull. Relax-vqa: Residual fragment and layer stack extraction for enhancing video quality assessment, 2024. URL <https://arxiv.org/abs/2407.11496>.
- [253] 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.
- [254] 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. *arXiv preprint arXiv:2307.06942*, 2023.
- [255] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019.
- [256] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13430–13439, 2021. doi: 10.1109/CVPR46437.2021.01323.
- [257] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.
- [258] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.
- [259] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.
- [260] Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. Crossscore: Towards multi-view image evaluation and scoring, 2024. URL <https://arxiv.org/abs/2404.14409>.
- [261] Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2763–2772, June 2024.
- [262] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions, 2021. URL <https://arxiv.org/abs/2104.14806>.
- [263] 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, 2022. URL <https://arxiv.org/abs/2210.05357>.

- [264] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023. URL <https://arxiv.org/abs/2304.14672>.
- [265] 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, 2023. URL <https://arxiv.org/abs/2211.04894>.
- [266] 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, March 2023. URL <http://arxiv.org/abs/2211.04894>. arXiv:2211.04894 [cs, eess].
- [267] 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*, pages 7623–7633, 2023.
- [268] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [269] Fengchuang Xing, Yuan-Gen Wang, Hanpin Wang, Leida Li, and Guopu Zhu. Starvqa: Space-time attention for video quality assessment, 2021. URL <https://arxiv.org/abs/2108.09635>.
- [270] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [271] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2935–2944, 2023.
- [272] 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 conference on computer vision and pattern recognition*, pages 5288–5296, 2016.
- [273] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks, 2017. URL <https://arxiv.org/abs/1711.10485>.
- [274] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023.
- [275] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. URL <https://arxiv.org/abs/2104.10157>.
- [276] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment, 2022. URL <https://arxiv.org/abs/2204.08958>.
- [277] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding, 2023. URL <https://arxiv.org/abs/2304.06906>.
- [278] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What You See is What You Read? Improving Text-Image Alignment Evaluation.
- [279] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [280] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: ‘patching up’ the video quality problem. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14014–14024, 2021. doi: 10.1109/CVPR46437.2021.01380.

- [281] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: ‘patching up’ the video quality problem. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.01380. URL <http://dx.doi.org/10.1109/CVPR46437.2021.01380>.
- [282] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild, 2024. URL <https://arxiv.org/abs/2405.18842>.
- [283] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks, 2022. URL <https://arxiv.org/abs/2202.10571>.
- [284] Kun Yuan, Zishang Kong, Chuanchuan Zheng, Ming Sun, and Xing Wen. Capturing co-existing distortions in user-generated content for no-reference video quality assessment, 2023. URL <https://arxiv.org/abs/2307.16813>.
- [285] Kun Yuan, Hongbo Liu, Mading Li, Muyi Sun, Ming Sun, Jiachao Gong, Jinhua Hao, Chao Zhou, and Yansong Tang. Ptm-vqa: Efficient video quality assessment leveraging diverse pretrained models from the wild, 2024. URL <https://arxiv.org/abs/2405.17765>.
- [286] Ao-Xiang Zhang, Yuan-Gen Wang, Weixuan Tang, Leida Li, and Sam Kwong. Hvs revisited: A comprehensive video quality assessment framework, 2022. URL <https://arxiv.org/abs/2210.04158>.
- [287] 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, 2023. URL <https://arxiv.org/abs/2309.15818>.
- [288] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. URL <https://arxiv.org/abs/1612.03242>.
- [289] 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.
- [290] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL <https://arxiv.org/abs/2302.05543>.
- [291] 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.
- [292] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>.
- [293] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.
- [294] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023. URL <https://arxiv.org/abs/2311.01361>.
- [295] Zhichao Zhang, Xinyue Li, Wei Sun, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Puyi Wang, Zhongpeng Ji, Fengyu Sun, Shangling Jui, and Guangtao Zhai. Benchmarking aigc video quality assessment: A dataset and unified model, 2024. URL <https://arxiv.org/abs/2407.21408>.
- [296] Zicheng Zhang, Wei Wu, Wei Sun, Dangyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos, 2023. URL <https://arxiv.org/abs/2303.14933>.
- [297] Kai Zhao, Kun Yuan, Ming Sun, and Xing Wen. Zoom-vqa: Patches, frames and clips integration for video quality assessment, 2023. URL <https://arxiv.org/abs/2304.06440>.
- [298] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL <https://github.com/hpcatech/Open-Sora>.

- [299] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [300] Xunchu Zhou, Xiaohong Liu, Yunlong Dong, Tengchuan Kou, Yixuan Gao, Zicheng Zhang, Chunyi Li, Haoning Wu, and Guangtao Zhai. Light-vqa+: A video quality assessment model for exposure correction with vision-language guidance, 2024. URL <https://arxiv.org/abs/2405.03333>.
- [301] 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.
- [302] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqqa: Deep meta-learning for no-reference image quality assessment, 2020. URL <https://arxiv.org/abs/2004.05508>.
- [303] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. Measuring github copilot’s impact on productivity. *Commun. ACM*, 67(3):54–63, February 2024. ISSN 0001-0782. doi: 10.1145/3633453. URL <https://doi.org/10.1145/3633453>.
- [304] Indre Zliobaite. Fairness-aware machine learning: a perspective, 2017. URL <https://arxiv.org/abs/1708.00754>.