The Evaluation Imperative for Video Generative Models: A Survey on Metrics, Benchmarks, and Trustworthiness

Anonymous CVPR submission

Paper ID 16

Abstract

The rapid evolution of video generative models has shifted the research focus from basic synthesis to high-fidelity, physically grounded content. However, traditional evaluation frameworks remain largely insufficient, often failing to capture the spatiotemporal consistency and semantic nuances required for verifiable quality. This survey provides a critical analysis of the AI-generated video evaluation (AIGVE) field, categorizing methodologies into metric-based, human-involved, and model-centered paradigms. We examine how architectural shifts primarily toward diffusion transformers and autoregressive frameworks have redefined evaluation requirements. We trace the evolution of metrics from frame-level heuristics to perceptually grounded spatiotemporal measures and analyze benchmarks spanning text-to-video, specialized conditional generation, and long-form storytelling. The review also addresses the critical dimensions of human preference alignment, physical world simulation, and safety-aware assessment, while exploring the impact of generative priors on downstream applications like video editing and compression. By identifying current limitations, we propose a path toward unified and trustworthy evaluation frameworks for the next generation of video foundation models.

1. Introduction

Establishing rigorous evaluation is the cornerstone of reliable video generation.

1.1. From Visual Synthesis to Verifiable Quality

The rapid evolution of video generative models has shifted the research frontier from basic visual synthesis to the pursuit of high-fidelity, temporally consistent, and physically grounded content. However, existing evaluation frameworks such as VBench and EvalCrafter often fall short in assessing dynamic scenarios, primarily due to their reliance on subject-centric, static prompts and frame-level metrics [1]. These benchmarks frequently prioritize aesthetic fidelity over cinematic camera motion and temporal causality, leading to technical failures in consistency metrics when faced with complex occlusions or disocclusions [1].

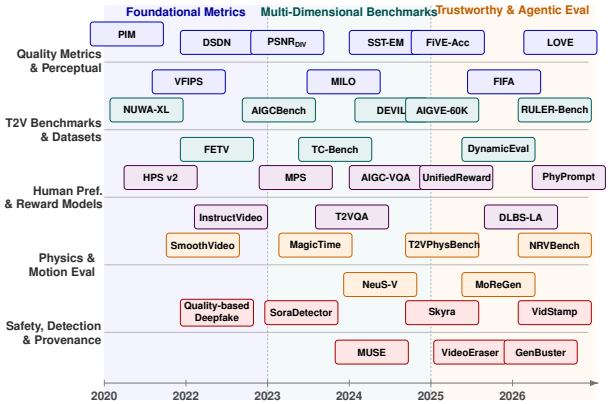

Figure 1. Chronological development timeline of the video generative model evaluation landscape (2020–2026), organized across five research lanes: quality metrics, benchmarks, human preference alignment, physics-aware evaluation, and safety/provenance. Three evolutionary phases are identified: foundational per-frame metrics, multi-dimensional benchmark expansion, and the current shift toward trustworthy, agentic evaluation frameworks.

Furthermore, traditional automated metrics exhibit a significant disconnect from human perception. Standard measures like Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD) are loosely coupled with subjective quality, while CLIP-based scores often fail to capture temporal nuances [16, 22, 32]. This misalignment is particularly evident in temporal compositionality tasks; for instance, state-of-the-art models achieve less than 20% success on the TC-Bench benchmark [13]. The imperative for verifiable quality thus necessitates a move beyond simple image-based heuristics toward metrics that can rigorously evaluate physical consistency, complex motion dynamics, and long-term temporal coherence [56, 64].

Table 1. Key papers advancing evaluation across metrics, benchmarks, alignment, and trustworthiness.

Category	Paper	Main Contribution
Evaluation Metrics	$PSNR_{DIV}$ [9]	Enhances PSNR with motion-field divergence weighting, achieving +0.09 PLCC over FloLPIPS at 2.5× speed.
	VFIPS [18]	Swin Transformer-based spatiotemporal perceptual metric for video frame interpolation quality.
	NeuS-V [41]	Neuro-symbolic text-video alignment via Temporal Logic specifications; 5× higher human correlation.
	FIFA [21]	Spatio-Temporal Semantic Dependency Graph for unified faithfulness evaluation of T2V and V2T.
	DEVIL [26]	Multi-granularity dynamics protocol achieving >90% Pearson correlation with human ratings.
T2V Benchmarks	AIGVE-60K [47]	Largest AIGV evaluation dataset: 58,500 videos, 2.6M human ratings, 30 T2V models, 20 task dims.
	DynamicEval [1]	45K human annotations for dynamic camera motion; proposes MS-Debias and Tracker-FG metrics.
	TC-Bench [13]	Temporal compositionality benchmark; reveals <20% success rate for state-of-the-art generators.
	RULER-Bench [17]	40 cognitive reasoning tasks across 6 categories; rule coherence scores often below 50%.
	T2VPhysBench [14]	First-principles physics benchmark evaluating 12 core physical laws via human judgment.
	LoCoT2V-Bench [65]	Long-form generation benchmark with hierarchical metadata for character and camera consistency.
Human Alignment	HPS v2 [53]	Fine-tunes CLIP on 798K human preference choices; superior generalization across distributions.
	UnifiedReward [49]	First unified reward model supporting pairwise ranking & pointwise scoring for image and video.
	DLBS-LA [36]	Inference-time beam search with VLM-calibrated rewards; training-free alignment improvement.
	PhyPrompt [52]	RL-based prompt refinement with GRPO; +8.6pp joint success on VideoPhy2 with zero-shot transfer.
Safety & Trust	Skyra [24]	MLLM for grounded artifact reasoning with ViF-CoT-4K dataset and two-stage SFT+RL training.
	GenBuster-200K [50]	200K+ video forensics dataset with three-track benchmark and MLLM-as-Judge protocol.
	VideoEraser [55]	Training-free concept erasure achieving 46% average reduction in undesirable content via SPEA+ARNG.
	SafeVid [48]	DPO-based safety alignment for VLMMs; up to 42.39% safety improvement on SafeVidBench.
	VidStamp [45]	Temporally-aware watermarking: 48 bits/frame, 0.96 tamper localization accuracy, robust to distortions.

1.2. Scope, Taxonomy, and Organization

This review provides a structured overview of the AI-generated video evaluation (AIGVE) landscape, categorized into three primary paradigms: metric-based, human-involved, and model-centered evaluations [31]. Metric-based approaches utilize quantitative formulas to assess objective dimensions such as visual fidelity and temporal consistency [20]. Human-involved protocols validate these metrics through user studies and preference tests, while emerging model-centered paradigms leverage Multimodal Large Language Models (MLLMs) and agent-based systems to automate perceptually grounded assessments [56, 67]. A summary of key papers across these dimensions is provided in Tab. 1.

The unique challenges of video evaluation stem from the dual consideration of spatial and temporal dynamics. Unlike static image assessment, video evaluation must account for motion direction, event ordering, and temporal compositionality, including attribute transitions and object relationship shifts over time [13, 32]. We categorize these challenges across four primary quality dimensions: visual fidelity, temporal consistency, semantic alignment, and physical consistency [11, 64]. Furthermore, we examine the taxonomy of applications based on input modalities, including text-to-video (T2V), image-to-video (I2V), video-to-video (V2V), and audio-conditioned generation [35]. The remainder of this paper is organized as follows: Section 2 discusses technical foundations; Section 3 details evaluation metrics; Sections 4 and 5 explore general-purpose and specialized benchmarks; Section 6 examines human preference alignment; Section 7 addresses safety and trustworthiness; and Section 8 concludes with open challenges and future directions.

2. Foundations: Video Gen. Model Taxonomy

Architectural innovations drive the spatiotemporal complexity of modern video models.

2.1. Diffusion Transformer Architectures

Diffusion Transformers (DiT) have emerged as the dominant backbone for high-fidelity video synthesis, leveraging scalable attention mechanisms to model complex spatiotemporal dependencies. CogVideoX [57] exemplifies this trend by utilizing a 3D causal VAE and expert transformer blocks to facilitate deep text-video fusion. To optimize the tradeoff between semantic fidelity and computational cost, FlashVideo [63] employs a two-stage framework that decouples motion alignment from high-resolution detail refinement, strategically allocating model capacity across stages. Efficiency remains a critical focus in DiT design, with methods like ADAPTOR [38] introducing adaptive token reduction to exploit temporal redundancy. Similarly, SNED [25] utilizes neural architecture search to address the cost-resolution tradeoff. For temporal stability, the Dual-Stream Diffusion Net (DSDN) [27] separates content and motion into distinct streams, mitigating flicker through cross-transformer synchronization. These architectural advancements necessitate evaluation frameworks that can specifically target motion-detail decoupling and adaptive compute efficiency.

2.2. Autoregressive Architectures

Autoregressive (AR) models leverage the predictive power of Large Language Models (LLMs) to synthesize video as a sequence of discrete tokens. Lumos-1 [60] adapts this paradigm by introducing spatiotemporal position embeddings and discrete diffusion forcing to address frame-wise loss imbalance. This approach achieves performance comparable to state-of-the-art diffusion models on benchmarks like VBench and GenEval. Further innovations in AR frameworks focus on long-range consistency and multimodal integration. ART•V [51] decomposes video synthesis into a sequential process, utilizing anchored conditioning to handle extended sequences. For interactive

CVPR #16

applications, Midas [7] integrates audio, pose, and text encodings into an LLM backbone, utilizing high-ratio spatial compression to maintain real-time efficiency. These models highlight the importance of evaluating long-term coherence and cross-modal synchronization in token-based synthesis.

3. Evaluation Metrics for Video Generation

Quantitative measures bridge the gap between pixel-level fidelity and human perception.

3.1. Perceptual and Full-Reference Quality Metrics

Traditional full-reference metrics such as PSNR and SSIM often fail to capture video-specific artifacts like ghosting and temporal flickering, as they operate on a per-frame basis without considering motion coherence. To address these limitations, PSNRDIV [9] enhances standard PSNR by weighting the mean squared error with motion-field divergence, effectively identifying temporal singularities. Similarly, VFIPS [18] leverages Swin Transformer blocks to extract spatiotemporal features, significantly outperforming image-based perceptual metrics like LPIPS in correlating with human judgment on video frame interpolation tasks.

Efficiency and unsupervised learning are also key research directions. MILO [69] introduces a lightweight multiscale CNN architecture that utilizes pseudo-MOS supervision from an ensemble of existing metrics, enabling training without large-scale human-labeled datasets. From an information-theoretic perspective, the Perceptual Information Metric (PIM) [3] employs unsupervised contrastive learning on video frame pairs. Grounded in human visual system physiology, PIM extracts temporally persistent features that demonstrate competitive performance with supervised metrics while exhibiting superior robustness to geometric distortions.

3.2. Temporal Consistency and Motion Metrics

Assessing temporal consistency requires metrics that can distinguish between intentional motion and technical artifacts. DynamicEval [1] identifies critical failures in standard motion smoothness metrics, which often yield large errors near occlusions caused by camera movement. To correct this, they propose MS-Debias, which utilizes object segmentation masks to isolate problematic regions, and Tracker-FG, which tracks points within object instances to assess fidelity under dynamic camera motion. These methods significantly improve correlation with human preferences by focusing on pixel-level temporal smoothness rather than global features.

Comprehensive dynamics evaluation is further addressed by the DEVIL protocol [26], which measures video dynamics across multiple temporal granularities, inter-frame, inter-segment, and video-level. DEVIL defines metrics for dynamics range, controllability, and dynamics-based quality, achieving a Pearson correlation exceeding 90% with human ratings. For adjacent frame transitions, SmoothVideo [37] introduces the Video Latent Score (VL score), which processes frames through a fine-grained autoencoder to ensure temporal order and capture granular details. Additionally, the SST-EM framework [4] integrates semantic extraction, object tracking, and temporal consistency for video editing assessment, with component weights derived from regression analysis against human evaluations.

3.3. Semantic Alignment and Text-Video Metrics

Achieving high semantic alignment is a core challenge in text-to-video generation, yet traditional metrics like CLIPScore and FVD often correlate poorly with human judgment due to their lack of deep vision-language understanding and sensitivity to sample size [32]. To address this, the FETV benchmark introduces UMTScore and FVD-UMT, which leverage the Unbiased Multimodal Transformer (UMT) to provide more consistent rankings and stronger correlation with human perception. For formal verification, NeuS-V [41] converts text prompts into Temporal Logic (TL) specifications and translates videos into automaton representations, demonstrating a correlation with human evaluations that is over five times higher than existing heuristics.

Emerging metrics also utilize Large Multimodal Models (LMMs) for diagnostic assessment. FiVE-Acc [19] leverages VLMs to perform automated semantic querying, using Yes/No and multiple-choice questions to verify fine-grained object-level editing success. Similarly, the FIFA framework [21] models semantic dependencies via a Spatio-Temporal Semantic Dependency Graph (STSDG). By structuring descriptive facts as a directed acyclic graph, FIFA can systematically assess hallucinations through VideoQA, ensuring that prerequisite entities are verified before their dependent attributes, thereby providing a more faithful evaluation of text-video correspondence.

3.4. Human Preference and Multidimensional Assessment

Aligning automated metrics with human subjective preferences is essential for evaluating the real-world impact of generative models. The Human Preference Score v2 (HPS v2) [53] fine-tunes CLIP on a large-scale dataset of human choices, demonstrating superior generalization across diverse image distributions compared to prior metrics. Extending this to multiple dimensions, the Multi-dimensional Preference Score (MPS) [62] incorporates a preference condition module to capture aesthetics, semantic alignment, and detail quality separately, ensuring that the model attends to visual features relevant to specific human criteria.

For holistic video quality assessment, T2VQA [22]

CVPR #16

utilizes a transformer-based architecture that combines a BLIP-based alignment encoder with a Video Swin Transformer backbone. By leveraging cross-attention fusion and a Large Language Model (Vicuna v1.5) for score regression, T2VQA achieves significant improvements in subjective alignment on the T2VQA-DB dataset. Similarly, AIGC-VQA [33] integrates three functional branches for technical quality, aesthetics, and alignment. Utilizing a Spatial-Temporal Adapter (ST-Adapter) for image-to-video knowledge transfer and a divide-and-conquer training strategy, AIGC-VQA achieves state-of-the-art performance, providing predictions that are more closely aligned with human Mean Opinion Scores (MOS) than single-branch baselines.

4. Text-to-Video Benchmarks

Standardized benchmarks provide the testing grounds for general-purpose video synthesis.

4.1. General-Purpose T2V Evaluation Benchmarks

The evaluation of text-to-video (T2V) models has transitioned from simple visual quality metrics to comprehensive, multi-dimensional frameworks, as illustrated in the unified evaluation pipeline (Fig. 2). AIGVE-60K [47] represents the largest effort to date, comprising 58,500 videos and 2.6 million human ratings. It introduces the LOVE metric, which leverages a Large Multimodal Model (LMM) architecture to disentangle perceptual quality, text-video correspondence, and task-specific accuracy. Similarly, AIGCBench [11] employs 11 metrics across four dimensions, control-video alignment, motion effects, temporal consistency, and video quality, to provide a unified assessment of image-to-video and text-to-video generation.

To improve alignment with human perception, recent benchmarks have integrated advanced reasoning capabilities. Video-Bench [16] utilizes MLLMs with a Chain-of-Query (CoQ) technique and few-shot scoring to achieve superior correlation with human preferences. Addressing the limitations of model-level aggregated scores, DynamicEval [1] focuses on video-level evaluation through 45,000 human annotations, specifically targeting background and foreground consistency in dynamic scenes. Furthermore, agent-based systems like VideoGen-Eval [56] integrate LLM-based content structuring with MLLM judgment and specialized patch tools to evaluate temporal-dense dimensions across thousands of videos from both commercial and open-source models.

4.2. Dynamic Motion and Camera Benchmarks

Assessing motion dynamics and camera control is a critical challenge, as many existing benchmarks exhibit a bias toward static or low-dynamic content. DynamicEval [1] addresses this by curating 100 prompts focused on dynamic camera motion and operationalizing background and foreground consistency through debiased optical flow and point tracking. Similarly, the DIVE benchmark [30] mitigates the low-dynamic bias in image-to-video (I2V) frameworks by utilizing GPT-4o to quantify dynamic range and controllability, penalizing static outputs that might otherwise achieve high quality scores.

Temporal granularity is further explored in the DEVIL protocol [26], which defines dynamics scores across inter-frame, inter-segment, and video-level scales. By establishing a benchmark of 50,000 prompts across five dynamics grades, DEVIL enables a more nuanced evaluation of motion complexity versus visual quality. For multi-shot scenarios, the TalkCuts dataset [6] provides a comprehensive taxonomy of camera shot categories (e.g., close-up, full-body) for human speech video generation. The associated Orator framework leverages LLM-guided orchestration to ensure cinematographic coherence, evaluating shot transitions and emotional flow across extended narratives.

4.3. Physics, Reasoning, and Semantic Benchmarks

As video generative models advance, evaluating their adherence to physical laws and logical reasoning becomes paramount. The MoReSet benchmark [2] spans multiple classes of Newtonian phenomena to assess physical validity. Similarly, T2VPhysBench [14] shifts focus toward first-principles physics, utilizing human judgment to evaluate model adherence to core physical laws such as conservation principles. Cognitive reasoning and semantic compositionality are further probed by RULER-Bench [17] and TC-Bench [13], which reveal significant reasoning limitations and gaps in temporal transition success. Additionally, T2VTextBench [15] identifies critical deficiencies in on-screen text fidelity, highlighting that even state-of-the-art systems struggle with textual accuracy and temporal consistency across complex strings.

4.4. Long-Form and Storytelling Benchmarks

The synthesis of long-form videos and structured narratives introduces unique challenges in identity preservation and temporal coherence. LoCoT2V-Bench [65] provides hierarchical metadata for character settings and camera behaviors, revealing that while current models excel in global stability, they struggle with long-term character consistency and fine-grained alignment. To evaluate narrative fidelity, the T2Vid2T framework [44] proposes a cyclical evaluation approach, text-to-video-to-text, identifying significant performance drops when models are tasked with generating complex, multi-scene stories compared to simple factual captions.

Architectural innovations such as NUWA-XL [58] address the computational demands of long video generation. By employing a Diffusion over Diffusion (DoD) hierarchy,

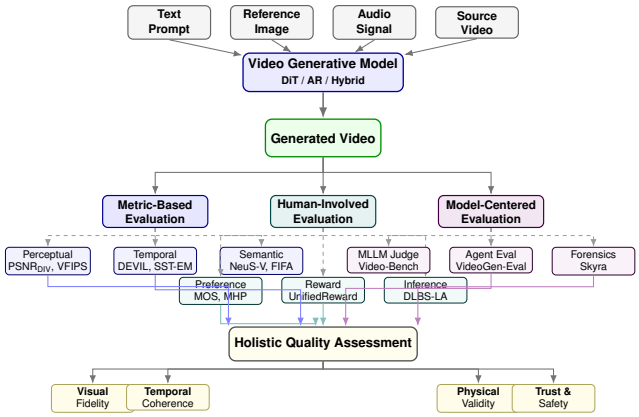

Figure 2. Unified evaluation pipeline for AI-generated video. Multi-modal inputs are processed by generative architectures and assessed via metric-based, human-involved, and model-centered paradigms. These signals converge into a holistic quality score covering visual fidelity, temporal coherence, physical validity, and trustworthiness.

NUWA-XL reduces inference time for 1024 frames by over 90% and introduces the FlintstonesHD benchmark, a densely annotated dataset for validating long-term coherence. Furthermore, VideoThinkBench [46] frames video generation as a reasoning paradigm, distinguishing between vision-centric tasks (e.g., spatial puzzles) and text-centric tasks (e.g., mathematical logic). Evaluations on this benchmark show that advanced models like Sora-2 can achieve high accuracy on complex subsets like MATH and MMMU, suggesting a path toward more intelligent, reasoning-aware video synthesis.

5. Specialized and Conditional Benchmarks

Conditional generation requires fine-grained assessment of structural and modal alignment.

5.1. Image-to-Video and Editing Benchmarks

Evaluation of conditional video generation and editing requires fine-grained assessment of instruction alignment and structural preservation. The FiVE benchmark [19] addresses this by providing six editing types, ranging from object substitution to material modification, across 420 object-level prompt pairs. It utilizes 15 metrics, including the VLM-based FiVE-Acc, which reveals that rectified-flow methods like Wan-Edit significantly outperform traditional diffusion-based approaches in accuracy and background preservation. For non-rigid dynamics, NRVBench [39] introduces 180 physics-based videos across six categories (e.g., fluid flow, fabric dynamics), employing the NRVE-Acc metric to evaluate physical compliance and temporal smoothness via VLM judges.

Creative and controllable editing is further benchmarked by BalanceCC [12], which structures 100 videos with diverse prompts (style, object, background changes) for extensive user studies against eight state-of-the-art methods. Specialized tasks such as object removal are evaluated by WIPER-Bench [23], which introduces the TokSim metric. Leveraging DINOv3 embeddings, TokSim quantifies temporal consistency, foreground-background coherence, and removal effectiveness, providing a robust measure for inpainting quality without requiring ground truth.

5.2. Audio-Video and Multi-Modal Benchmarks

The integration of auditory and visual modalities introduces complex synchronization challenges. JavisBench [29] provides a large-scale dataset of text-captioned sounding videos, introducing JavisScore, a robust synchrony metric that utilizes temporal segmentation to detect desynchronization patterns. For conversational dynamics, MTAVG-Bench [66] evaluates multi-talker dialogue across levels ranging from signal fidelity to cinematic expression.

Specialized multi-modal tasks such as narrated presentations are assessed via PresentEval [43], which measures content fidelity and audience comprehension. Furthermore, the JavisGPT framework [28] demonstrates state-of-the-art performance in joint audio-video comprehension and generation. By maintaining high semantic consistency and temporal synchrony, these benchmarks facilitate the development of more cohesive multi-modal generative systems.

6. Human Preference Alignment

Aligning models with human intent ensures the subjective quality of generated content.

6.1. Human Preference Annotation Datasets

The alignment of video generative models with human expectations requires large-scale, high-quality preference datasets. AIGVE-60K [47] enables bidirectional benchmarking for both text-to-video (T2V) generation and video-to-text (V2T) interpretation, comprising 58,500 videos from 30 models and 2.6 million human annotations. This dataset disentangles Mean Opinion Scores (MOS) into perceptual quality and text-video correspondence, providing a granular evaluation across 20 task dimensions. Similarly, the Multi-dimensional Human Preference (MHP) Dataset [62] offers 918,315 pairwise choices across aesthetics, semantic alignment, and detail quality, facilitating robust preference modeling for visual synthesis.

Beyond general quality, specialized datasets address motion and safety. For text-to-motion generation, researchers have explored preference learning using 3,528 pairs from MotionGPT, identifying Direct Preference Optimization (DPO) with Identity Preference Optimization (IPO) loss as superior to traditional RLHF for alleviating overfitting [42].

To address safety concerns, SafeVid-350K [48] provides a video-specific safety preference dataset constructed through a systematic pipeline of adversarial question generation and preference synthesis. Utilizing this dataset within a DPO framework significantly enhances the safety alignment of Video Large Multimodal Models (VLMMs) without compromising general utility.

6.2. Reward Models and Preference Optimization

The development of dedicated reward models is critical for automating the alignment process. UnifiedReward [49] represents a unified framework supporting both pairwise ranking and pointwise scoring across image and video tasks. By leveraging joint multi-task training, it enhances video quality assessment and mitigates modality-specific overfitting, utilizing a two-stage filtering strategy to construct high-quality preference data for optimization. To address the computational cost of reward fine-tuning, InstructVideo [59] recasts the process as an editing task, utilizing partial inference steps and segmental sparse sampling to repurpose established image reward models. Furthermore, temporally attenuated reward mechanisms are employed to mitigate modeling degradation. By prioritizing central frames and ensuring motion continuity, these methods prevent visual artifacts such as flickering during the fine-tuning process, facilitating more robust human-aligned synthesis.

6.3. Inference-Time Alignment

Inference-time strategies offer a training-free alternative for aligning generated content with user intent. Diffusion Latent Beam Search (DLBS) [36] optimizes text-to-video alignment by exploring multiple latent paths, leveraging calibrated rewards from VLMs to select high-quality trajectories. Similarly, the 3R framework [40] combines RAG-based modifier extraction with preference optimization to enhance fidelity and coherence without model retraining.

Prompt optimization further bridges the gap between user instructions and model capabilities. PhyPrompt [52] utilizes reinforcement learning to refine prompts for physical plausibility, integrating a dynamic reward curriculum that balances semantic adherence with physical commonsense. These advancements are supported by reward calibration methodologies that utilize weighted linear combinations of perceptual metrics to align automated scores with human and VLM evaluations [36].

7. Safety and Trustworthiness

Ensuring safety and provenance is critical for responsible AI deployment.

7.1. AI-Generated Video Detection and Forensics

The proliferation of high-fidelity video generative models necessitates robust detection and forensic tools to mitigate the risks of misinformation. Skyra [24] addresses this by utilizing fine-grained human annotations and Chain-of-Thought explanations to enhance artifact perception across diverse generators. This is complemented by comprehensive forensics datasets like GenBuster-200K [50] and standardized evaluation platforms like ViF-Bench.

Fine-grained artifact recognition is further formalized by the Dense Video Artifact Recognition (DVAR) framework [67], which taxonomizes defects across appearance, motion, and camera axes. To detect hallucinations, the SoraDetector framework [8] identifies semantic and temporal inconsistencies via knowledge graphs. Furthermore, recent deepfake detection approaches employ quality-based matrix representations to enable frame-by-frame analysis, identifying minute spatio-temporal anomalies that standard playback might overlook.

7.2. Adversarial Robustness of T2V Models

The vulnerability of text-to-video (T2V) models to adversarial perturbations poses significant security risks. T2VAttack [5] demonstrates that minor word-level modifications, such as single-word substitution or insertion, can cause substantial degradation in both semantic fidelity and temporal dynamics across diverse architectures like CogVideoX and HunyuanVideo. By targeting semantic alignment and motion coherence, these attacks reveal a lack of robustness in current generative pipelines. Furthermore, the MUSE platform [10] introduces Inter-Turn Modality Switching (ITMS) to probe whether safety alignments generalize across modality boundaries. MUSE reports that multi-turn strategies can achieve near-100% attack success rates against models that otherwise exhibit perfect single-turn refusal, highlighting the fragility of cross-modal defenses.

Safety alignment is further complicated by ”mismatched generalization,” where safety competencies learned from static text and images fail to transfer to dynamic video contexts. SafeVid [48] identifies this failure mode, noting that models may reject harmful queries in isolation but comply when they are paired with relevant video content. To address this, the SafeVid-350K dataset and Direct Preference Optimization (DPO) are utilized to transfer robust textual safety reasoning to the video domain, significantly improving safety rates on the SafeVidBench benchmark while maintaining model utility.

7.3. Video Watermarking and Provenance

Establishing the provenance of AI-generated videos is essential for accountability and copyright protection. VidStamp [45] introduces a temporally-aware watermarking framework that utilizes a two-stage fine-tuning process to ensure spatial separation and temporal consistency. By embedding unique frame-level bit sequences, VidStamp en-

ables precise temporal tamper localization, detecting frame swaps and drops while maintaining high visual quality.

Furthermore, VidStamp supports dynamic watermarking, allowing for adaptive provenance via control signals during inference. The framework demonstrates robust performance across diverse distortions, including cropping and MPEG4 compression. Comparative evaluations show that VidStamp achieves stronger detectability than existing methods such as VideoSeal and RivaGAN, providing a reliable solution for long-term video integrity and authentication.

7.4. Concept Erasure and Safety Generation

The ability to selectively remove undesirable concepts or align generation with safety guidelines is paramount for responsible AI deployment. VideoEraser [55] provides a training-free framework for concept erasure, covering objects, styles, and explicit content, by adjusting prompt embeddings and noise guidance. This approach ensures broad generalizability across diverse T2V models without costly fine-tuning. Safety alignment is further formalized through closed-loop systems like SafeVid [48], which transfers robust textual safety reasoning to the video domain. By aligning models on safety preference datasets, researchers have demonstrated substantial improvements in safety rates on standardized benchmarks. For commercial applications, frameworks like BrandFusion [68] utilize multi-agent systems to ensure brand recognizability and contextually natural integration, facilitating the safe use of generative video in advertising.

8. Open Challenges and Future Directions

The path toward unified world models requires overcoming fundamental spatiotemporal gaps.

8.1. Long-Form Coherence, Temporal Reasoning, and World Models

A critical frontier in video generation is the transition from short-clip synthesis to long-form, coherent storytelling. Evaluations reveal a significant performance gap: while current models demonstrate strong perceptual quality, they exhibit deficiencies in fine-grained alignment and long-term character consistency. Achieving temporal coherence over extended durations remains an open challenge, necessitating architectures that can bridge the training-inference gap through hierarchical strategies and direct training on long sequences. Beyond visual consistency, the integration of temporal reasoning is essential for developing true world models. Framing video generation as a reasoning chain leverages dynamic visualization to solve complex spatial and textual puzzles, offering advantages over static image-based reasoning. However, existing models still struggle with basic temporal order relations. Future research must focus on instilling relational consistency and leveraging test-time scaling to transform video generators into unified multimodal understanding and generation systems.

8.2. Physical World Simulation

A fundamental limitation of current neural video generators is the divergence between high visual realism and physical plausibility. Evidence from benchmarks like MoReSet [2] and T2VPhysBench [14] indicates that state-of-the-art models frequently violate basic physical laws, such as momentum conservation and rigid-body collision dynamics, relying instead on surface pattern memorization. To address this, research is moving toward "metamorphic simulators" like MagicTime [61], which encode physical world knowledge through the simulation of complete metamorphic processes. Principled directions for physically coherent synthesis include the integration of simulation-aware architectures and reinforcement learning. MoReGen [2] unifies multi-agent LLMs with physics simulators and renderers to ensure dynamic coherence via code-domain generation. Similarly, PhyPrompt [52] utilizes Group Relative Policy Optimization (GRPO) to inject physical reasoning into instructions, overcoming prompt underspecification. Advancing this field requires standardized evaluation platforms like NRVBench [39], which specifically targets non-rigid dynamics, such as fluidity and elasticity, using VLM-based metrics to assess deformation plausibility and temporal consistency.

8.3. Computational Efficiency, Scalability, and Sustainable Generation

The escalating computational demands of video generative models necessitate a focus on efficiency and sustainability. Analytical models reveal that energy consumption scales quadratically with spatial and temporal dimensions, requiring prioritization of compute-bound efficiency and hardware-aware optimizations. To this end, H3AE [54] achieves real-time decoding on mobile devices by utilizing high spatial-temporal compression ratios, significantly reducing latent token counts without compromising reconstruction fidelity.

Inference-time compute strategies further optimize the efficiency-quality tradeoff. ADAPTOR [38] exploits temporal redundancy through adaptive token reduction, while FlashVideo [63] employs a two-stage framework that decouples prompt fidelity from visual detail generation. These methods, combined with search-based latent selection, enable high-quality video synthesis at a fraction of the traditional computational cost, paving the way for more accessible and sustainable generative AI.

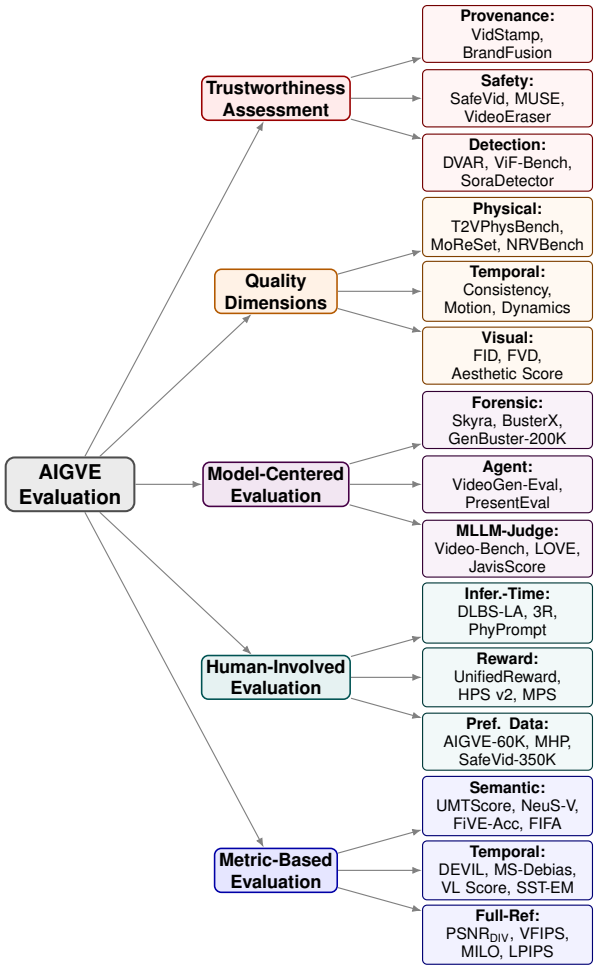

Figure 3. Hierarchical taxonomy of the AI-Generated Video Evaluation (AIGVE) landscape across five branches: metric-based, human-involved, model-centered, quality dimensions, and trustworthiness assessment.

8.4. Toward Unified, Adaptive, and Trustworthy Evaluation Frameworks

The current landscape of video evaluation is characterized by significant fragmentation, with individual benchmarks often prioritizing specific dimensions while neglecting others. A hierarchical taxonomy of this landscape, spanning metrics, human alignment, and model-centered paradigms, is provided in Fig. 3. To overcome the limitations of traditional metrics like FID and FVD, which fail to align with human perception, research is shifting toward model-centered evaluation frameworks [31]. Multimodal Large Language Models (MLLMs) are emerging as the cornerstone of next-generation assessment, providing unified and scalable frameworks that leverage few-shot scoring and chain-of-query processes to simulate human evaluation logic [16]. Agent-based systems like VideoGen-Eval [56] further enhance this by integrating dynamic content structuring with objective judgment, achieving high alignment with professional human preferences across diverse quality axes.

The future roadmap for trustworthy video evaluation emphasizes the convergence of quality, safety, and provenance into a cohesive framework. Detection benchmarks such as Skyra [24] and GenBuster-200K [50] are transitioning from binary classification toward grounded artifact reasoning, utilizing MLLMs to identify spatio-temporal inconsistencies. Simultaneously, safety alignment frameworks like SafeVid [48] address the dynamic complexities that static alignments fail to capture. A truly unified protocol must simultaneously evaluate visual quality, temporal coherence, physical validity, and semantic reasoning, while incorporating cryptographic provenance and content credentials to mitigate the risks of deceptive deepfakes [20]. This holistic approach is essential for ensuring that AI-generated content remains aligned with human intentions, ethical standards, and real-world physical constraints.

9. Conclusion

This survey has analyzed the AI-generated video evaluation (AIGVE) field, highlighting the transition from basic visual synthesis to the pursuit of verifiable spatiotemporal quality. We have traced the evolution of evaluation methodologies, from traditional frame-level heuristics to model-centered paradigms that leverage Multimodal Large Language Models (MLLMs) for perceptually grounded assessment. Our analysis indicates that developing true video world models requires addressing fundamental gaps in physical simulation, long-form coherence, and temporal reasoning. As video generative models are increasingly integrated into downstream applications, the convergence of quality, safety, and provenance into unified evaluation frameworks has become a critical necessity. By establishing these rigorous standards, the field can ensure that the next generation of video foundation models remains aligned with human intent, physical reality, and ethical principles.

Acknowledgements

The authors developed all intellectual organization and critical synthesis in this review. PaperQA [34], a RAG-based literature tool, assisted in factual retrieval from surveyed papers; all retrieved information was verified by the authors.

References

[1] Nithin C Babu, Aniruddha Mahapatra, Harsh Rangwani, R. Soundararajan, and Kuldeep Kulkarni. Dynamiceval: Rethinking evaluation for dynamic text-to-video synthesis. In *arXiv.org*, 2025. 1, 2, 3, 4

[2] Xiangyu Bai, He Liang, Bishoy Galoaa, Utsav Nandi, Shayda Moezzi, Yuhang He, and Sarah Ostadabbas. Moregen: Multi-agent motion-reasoning engine for code-based text-to-video synthesis. In *arXiv.org*, 2025. 4, 7

[3] Sangnie Bhardwaj, Ian Fischer, Johannes Ball'e, and Troy T. Chinen. An unsupervised information-theoretic perceptual quality metric. In *Neural Information Processing Systems*, 2020. 3

[4] Varun Biyyala, Bharat Chanderprakash Kathuria, Jialu Li, and Youshan Zhang. Sst-em: Advanced metrics for evaluating semantic, spatial and temporal aspects in video editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2025. 3

[5] Chang bo Li, Yuecong Min, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. T2vattack: Adversarial attack on text-to-video diffusion models. In *arXiv.org*, 2025. 6

[6] Jiaben Chen, Zixin Wang, Ailing Zeng, Yang Fu, Xueyang Yu, Siyuan Cen, Julian Tanke, Yihang Chen, Koichi Saito, Yuki Mitsufuji, and Chuang Gan. Talkcuts: A large-scale dataset for multi-shot human speech video generation. In *arXiv.org*, 2025. 4

[7] Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, Xiaoqiang Liu, and Pengfei Wan. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. In *arXiv.org*, 2025. 3

[8] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. Sora detector: A unified hallucination detection for large text-to-video models. In *arXiv.org*, 2024. 6

[9] Conall Daly, D. Ramsook, and Anil C. Kokaram. An efficient quality metric for video frame interpolation based on motion-field divergence. In *International Workshop on Quality of Multimedia Experience*, 2025. 2, 3

[10] Xin Dong, Sen Jia, Ming Wang, Yan Li, Zhenheng Yang, Bingfeng Deng, and Hongyu Xiong. Coef-vq: Cost-efficient video quality understanding through a cascaded multimodal llm framework. In *Knowledge Discovery and Data Mining*, 2024. 6

[11] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2024. 2, 4

[12] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *Computer Vision and Pattern Recognition*, 2023. 5

[13] Weixi Feng, Jiachen Li, Michael Stephen Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. In *arXiv.org*, 2024. 1, 2, 4

[14] Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. In *arXiv.org*, 2025. 2, 4, 7

[15] Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vtextbench: A human evaluation benchmark for textual control in video generation models. In *arXiv.org*, 2025. 4

[16] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *Computer Vision and Pattern Recognition*, 2025. 1, 4, 8

[17] Xuming He, Zehao Fan, Hengjia Li, Fan Zhuo, Hankun Xu, Senlin Cheng, Diwang Weng, Haifeng Liu, Can Ye, and Boxi Wu. Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. In *arXiv.org*, 2025. 2, 4

[18] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *European Conference on Computer Vision*, 2022. 2, 3

[19] Ming hui Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *arXiv.org*, 2025. 3, 5

[20] Muhammad Tanveer Jan, Mohammed G. Al-Jassani, Martinraj Nadar, Emmanuel Melchizedek Vunnava, Vangmai Chakrapani, Hayat Ullah, Abbas Khan, Sardar Ali Abbas, and B. Furht. Text-to-video generators: a comprehensive survey. *Journal of Big Data*, 2025. 2, 8

[21] Liqiang Jing, Viet Dac Lai, Seunghyun Yoon, Trung Bui, and Xinya Du. Fifa: Unified faithfulness evaluation framework for text-to-video and video-to-text generation. In *arXiv.org*, 2025. 2, 3

[22] 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. In *ACM Multimedia*, 2024. 1, 3

[23] Saksham Singh Kushwaha, Sayan Nag, Yapeng Tian, and Kuldeep Kulkarni. Object-wiper : Training-free object and associated effect removal in videos. In *arXiv.org*, 2026. 5

[24] Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. In *arXiv.org*, 2025. 2, 6, 8

[25] Zhengang Li, Yan Kang, Yuchen Liu, Difan Liu, Tobias Hinz, Feng Liu, and Yanzhi Wang. Sned: Superposition network architecture search for efficient video diffusion model. In *Computer Vision and Pattern Recognition*, 2024. 2

[26] 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. In *Neural Information Processing Systems*, 2024. 2, 3, 4

[27] Binhui Liu, Xin Liu, Anbo Dai, Zhiyong Zeng, Zhen Cui, and Jian Yang. Dual-stream diffusion net for text-to-video generation. In *arXiv.org*, 2023. 2

[28] Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Gen Zhan, Jiayi Ji, Fan Zhou, Liang Zheng,

Shuicheng Yan, Hao Fei, and Tat-Seng Chua. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. In *arXiv.org*, 2025. 5

[29] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *arXiv.org*, 2025. 5

[30] Peng Liu, Xiaoming Ren, Feng Liu, Qingsong Xie, Quanlong Zheng, Yanhao Zhang, Haonan Lu, and Yujiu Yang. Dynamic-i2v: Exploring image-to-video generation models via multimodal llm. In *arXiv.org*, 2025. 4

[31] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Wei Ye, and Jiawei Zhang. A survey of ai-generated video evaluation. In *arXiv.org*, 2024. 2, 8

[32] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Neural Information Processing Systems*, 2023. 1, 2, 3

[33] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 4

[34] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023. 8

[35] A. Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge J. Ritter. Video diffusion models: A survey. *Trans. Mach. Learn. Res.*, 2024. 2

[36] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *arXiv.org*, 2025. 2, 6

[37] Liang Peng, Haoran Cheng, Zheng Yang, Ruisi Zhao, Linxuan Xia, Chaotian Song, Qinglin Lu, Boxi Wu, and Wei Liu. Smoothvideo: Smooth video synthesis with noise constraints on diffusion models for one-shot video tuning, 2023. 3

[38] Elia Peruzzo, Adil Karjauv, N. Sebe, Amir Ghodrati, and A. Habibian. Adaptor: Adaptive token reduction for video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2, 7

[39] Bingzheng Qu, Kehai Chen, Xuefeng Bai, Jun Yu, and Min Zhang. Beyond rigid: Benchmarking non-rigid video editing. In *arXiv.org*, 2026. 5, 7

[40] Zillur Rahman, Alex Sheng, and Cristian Meo. Retrieval, refinement, and ranking for text-to-video generation via prompt optimization and test-time scaling, 2026. 6

[41] S. Sharan, Minkyu Choi, Sahil Shah, Harsh Goel, Mohammad Omama, and Sandeep P. Chinchali. Neuro-symbolic evaluation of text-to-video models using formal verification. In *Computer Vision and Pattern Recognition*, 2024. 2, 3

[42] Jenny Sheng, Matthieu Lin, Andrew Zhao, Kevin Pruvost, Yu-Hui Wen, Yangguang Li, Gao Huang, and Yong-Jin Liu. Exploring text-to-motion generation with human preference. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 5

[43] Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. Presentagent: Multimodal agent for presentation video generation. In *Conference on Empirical Methods in Natural Language Processing*, 2025. 5

[44] Andrew Shin, Yusuke Mori, and Kunitake Kaneko. The lost melody: Empirical observations on text-to-video generation from a storytelling perspective. In *arXiv.org*, 2024. 4

[45] Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. Vidstamp: A temporally-aware watermark for ownership and integrity in video diffusion models. In *arXiv.org*, 2025. 2, 6

[46] Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yong Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Y. Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *arXiv.org*, 2025. 5

[47] Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. Love: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. In *arXiv.org*, 2025. 2, 4, 5

[48] Yixu Wang, Jiaxin Song, Yifeng Gao, Xin Wang, Yang Yao, Yan Teng, Xingjun Ma, Yingchun Wang, and Yu-Gang Jiang. Safevid: Toward safety aligned video large multimodal models. In *arXiv.org*, 2025. 2, 6, 7, 8

[49] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. In *arXiv.org*, 2025. 2, 6

[50] Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. In *arXiv.org*, 2025. 2, 6, 8

[51] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art•v: Auto-regressive text-to-video generation with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 2

[52] Shang Wu, Chenwei Xu, Zhuofan Xia, Weijian Li, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation, 2026. 2, 6, 7

[53] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. In *arXiv.org*, 2023. 2, 3

[54] Yushu Wu, Yanyu Li, Ivan Skorokhodov, Anil Kag, Willi Menapace, Sharath Girish, Aliaksandr Siarohin, Yanzhi Wang, and Sergey Tulyakov. H3ae: High compression, high speed, and high quality autoencoder for video diffusion models. In *arXiv.org*, 2025. 7

[55] Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. Videoeraser: Concept erasure in text-to-video diffusion models. In *Conference on Empirical Methods in Natural Language Processing*, 2025. 2, 7

[56] Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, Feilin Han, Wei hao Zhai, Wei Liu, Yang Cao, and Zhengjun Zha. Videogen-eval: Agent-based system for video generation evaluation. In *arXiv.org*, 2025. 1, 2, 4, 8

[57] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2024. 2

[58] Sheng-Siang Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. Nuwa-xl: Diffusion over diffusion for extremely long video generation. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 4

[59] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Computer Vision and Pattern Recognition*, 2023. 6

[60] Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, Hao Luo, Jiasheng Tang, Fan Wang, and Yi Yang. Lumos-1: On autoregressive video generation from a unified model perspective. In *arXiv.org*, 2025. 2

[61] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

[62] Sixian Zhang, Bo Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Computer Vision and Pattern Recognition*, 2024. 3, 5

[63] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. In *arXiv.org*, 2025. 2, 7

[64] Zhichao Zhang, Wei Sun, and Guangtao Zhai. A perspective on quality evaluation for ai-generated videos. In *Italian National Conference on Sensors*, 2025. 1, 2

[65] Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. Locot2v-bench: Benchmarking long-form and complex text-to-video generation, 2025. 2, 4

[66] Yang-Hao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqi Zhou, Yudong Li, Jiajun Xu, Jing Liao, Yi Cheng, Xuefeng Chen, Xian-Ling Mao, and Yousheng Feng. Mtavg-bench: A comprehensive benchmark for evaluating multi-talker dialogue-centric audio-video generation, 2026. 5

[67] Chen Zhu, Jiashu Zhu, Yanxun Li, Meiqi Wu, Bingze Song, Chubin Chen, Jiahong Wu, Xiangxiang Chu, and Yangang Wang. Artifact-aware evaluation for high-quality video generation. In *arXiv.org*, 2026. 2, 6

[68] Zihao Zhu, Ruotong Wang, Siwei Lyu, Min Zhang, and Baoyuan Wu. Brandfusion: A multi-agent framework for seamless brand integration in text-to-video generation, 2026. 7

[69] Ugur Çogalan, Mojtaba Bemana, K. Myszkowski, Hans-Peter Seidel, and Colin Groth. Milo: A lightweight perceptual quality metric for image and latent-space optimization. *ACM Transactions on Graphics*, 2025. 3