Journal of Artificial Intelligence in Information Technology (JAI-IT)
Vol. 1 No. 1 June 2026
E-ISSN: XXXX-XXXX; P-ISSN: XXXX-XXXX, Pages 66-84
DOI: https://doi.org/10.51903/k8v16h58

Comparative Performance Analysis of Text-to-Video Models Across Workflow Stages and Quality Dimensions

Januar Tito Bagaskoro*[1,2], Muhammad Sholikhan[1]
Email: januar@stekom.ac.id , sholikhan@stekom.ac.id
Orcid: https://orcid.org/0009-0005-6240-8145 (1), https://orcid.org/0000-0002-4638-3532 (2)
[1]*Department of Design Communication Visual, Faculty of Academic Study. Universitas Sains dan Teknologi Komputer, Semarang, Indonesia, 50171*
[2]*Department of Design, Faculty of Fine Arts and Design. ISI Bali, Bali, Indonesia, 80235*
*Corresponding Author

Abstract
The rapid advancement of text-to-video (T2V) generative artificial intelligence has transformed digital content creation, yet a structured evaluation framework aligned with real-world production workflows remains absent. Traditional metrics correlate poorly with human perception and practical deployment needs. This study proposes the Production-Pipeline Evaluation Framework (PPEF) to assess leading T2V models across three workflow stages: pre-production, generation, and post-production. We evaluated six prominent frameworks (Sora, Runway Gen-3, Pika 2.0, CogVideoX-5B, HunyuanVideo, and Open-Sora 2.0) using a dataset of 300 production-oriented prompts. Performance was measured utilizing nine multi-dimensional metrics, encompassing automated standards, novel production-based indicators, and human perceptual evaluation (N=30). Results indicate that only closed-source models Sora (PPEF=0.742) and Runway Gen-3 (0.718) surpassed the production-ready threshold of 0.70. Among open-source alternatives, HunyuanVideo (0.685) demonstrated the strongest overall profile. Crucially, the composite PPEF score demonstrated a high correlation with human perception (Spearman ρ=0.847), significantly outperforming traditional automated metrics. The integration of production-based metrics revealed specific deployment advantages, such as Pika 2.0's generation speed and Runway Gen-3's post-production editability. These findings are synthesized into an IT Implementation Matrix, providing practitioners and organizations with structured, evidence-based guidance for selecting and deploying generative AI video tools based on technical maturity, budget, and specific workflow requirements.

Keywords: *Artificial Intelligence, Evaluation Framework, Generative AI, Text-to-Video Models, Video Production.*

I. INTRODUCTION

The proliferation of generative artificial intelligence (AI) has fundamentally reshaped the landscape of digital content creation. Among the most consequential advances in this domain is the emergence of text-to-video (T2V) generation a capability that enables the synthesis of temporally coherent, visually realistic video sequences from natural language descriptions. Generative AI can now autonomously produce creative content that previously required substantial human labor, specialized equipment, and domain expertise [1]. The implications for information technology (IT) organizations, media enterprises, digital marketing agencies, and individual content creators are profound: video production workflows that once demanded days of coordinated effort across scripting, filming, editing, and post-production teams can increasingly be instantiated through AI-driven pipelines within hours or even minutes.

Received on February, 2026; Revised on March, 2026; Accepted on April, 2026; Published on Juni, 2026

The scale of this transformation is reflected in the rapid succession of powerful T2V models released between 2023 and 2025. OpenAI's Sora demonstrated the potential for AI to act as a "world simulator" capable of generating photorealistic, long-duration video clips [2], igniting widespread industry interest. Commercial platforms such as Runway Gen-3 and Pika have made high-quality video generation accessible to non-technical users through intuitive web-based interfaces. Concurrently, open-source projects have substantially narrowed the performance gap with closed-source alternatives: HunyuanVideo trained with over 13 billion parameters has demonstrated performance comparable to or exceeding Runway Gen-3 and Luma 1.6 in professional evaluations [3]; Open-Sora 2.0 has shown that commercial-level quality can be achieved with a training budget of approximately $200,000 [4]; and CogVideoX has established strong benchmarks for temporally coherent, text-aligned video synthesis [5].

Despite this remarkable progress, a critical challenge persists for practitioners and IT decision-makers: the absence of a structured, production-oriented framework for evaluating and selecting generative AI video tools. Existing comparative studies predominantly assess model performance through a narrow set of automated metrics most commonly Fréchet Video Distance (FVD) and CLIP Similarity (CLIPScore) applied to isolated generation tasks disconnected from real production contexts [6], [7]. However, multiple studies have demonstrated that these metrics correlate poorly with human perceptual quality judgments and fail to capture dimensions critical to practical deployment, including temporal dynamics, physical consistency , narrative coherence across clips, and prompt-following fidelity under complex, multi-constraint instructions. A comprehensive survey of AI-generated video evaluation methods confirms that existing automated assessment frameworks remain "fragmented" and inadequate for evaluating AI-generated content across the full breadth of production requirements.

The challenge is compounded by the heterogeneity of available tools. IT organizations and production teams must navigate a complex decision space spanning closed-source platforms (Sora, Runway Gen-3, Pika) with subscription-based pricing and API rate limits, and open-source alternatives (HunyuanVideo, CogVideoX, Open-Sora 2.0) requiring GPU infrastructure and technical integration competency. Industry practitioners in the video production sector have identified eight primary adoption barriers for AI video tools, with technological maturity emerging as the single most impactful factor, ahead of cost, ethics, and cross-disciplinary collaboration considerations [8]. The efficiency gains from AI adoption in animation production have also been shown to vary significantly by pipeline stage with post-production exhibiting consistently high AI-driven efficiency (mean: 0.91) while creative generation tasks demonstrate greater variability and optimal integration has been empirically identified at between 30% and 70% AI involvement [9]. These findings underscore the urgent need for research that evaluates AI

video tools not merely on isolated generative quality, but on their practical fitness within defined workflow stages.

Existing reviews of AI in video production have recognized this need. Huang et al. proposed a five-sub-field framework for AI's role in film and television production, providing a process paradigm for AI integration [10]. Anantrasirichai and Bull documented the shift of AI from "support tool to core creative technology" across the entire creative production pipeline, while identifying computational demands, copyright concerns, and regulatory gaps as persistent challenges [11]. However, neither of these works conducts structured empirical comparisons across specific modern T2V frameworks. The study by Onisha et al. comparing Runway Gen2 and CogVideoX variants represents a meaningful step toward practical comparative evaluation [12], but it involves only three models, uses a limited evaluation protocol focused on FID, FVD, and CLIPScore, and does not embed assessment within production workflow stages. Similarly, the Physics-IQ benchmark directly compared six major systems including Sora, Runway, Pika, Lumiere, Stable Video Diffusion, and VideoPoe [13], but its focus is confined to physical plausibility rather than production-pipeline applicability.

This paper addresses these gaps by proposing and operationalizing a Production-Pipeline Evaluation Framework (PPEF) for generative AI video tools. PPEF decomposes the video production workflow into three structured stages (i) scripting and prompt engineering, (ii) video generation, and (iii) post-production integration and evaluates leading T2V frameworks across each stage using a multi-dimensional quality rubric that integrates both established automated metrics and newly proposed production-oriented indicators. The frameworks assessed include Sora (OpenAI), Runway Gen-3 (Runway AI), CogVideoX-5B (THUDM), HunyuanVideo (Tencent), Open-Sora 2.0 (HPC-AI Tech), and Pika 2.0 (Pika Labs), representing the most influential closed-source and open-source systems available as of 2025.

II. LITERATURE REVIEW

A. Generative AI in Digital Content Creation and Video Production

Generative AI encompasses a class of models capable of producing novel content text, images, audio, and video from learned data distributions [1]. The rapid maturation of this technology since 2022, driven by transformers, large language models (LLMs), and latent diffusion models, has positioned it as a core creative technology rather than a supplementary tool [11]. In the video domain specifically, AI has enabled innovative approaches to personalized, efficient video production, catalyzed by the rise of self-media and user-generated content platforms [10].

Research directly examining AI integration within video production environments highlights both significant opportunity and persistent friction. [10] analyzed AI's influence across five sub-fields of the video production pipeline pre-production, asset creation, production, post-production, and distribution proposing a new process paradigm in which AI augments each stage rather than replacing the holistic workflow. The empirical analysis [9] is among the most rigorous production-context evaluations in the literature: using Network Data Envelopment Analysis (NDEA) across ten animation projects spanning commercial, educational, and entertainment sectors, they found that AI's efficiency impact varies substantially by pipeline stage. Post-production achieves consistently high efficiency scores (mean: 0.91), while creative generation tasks demonstrate greater dependency on organizational factors and AI-human integration strategy. Critically, they found that optimal efficiency gains occur when AI adoption is maintained between 30% and 70%, suggesting that neither wholesale adoption nor conservative supplementation yields ideal outcomes.

Industry adoption studies further illuminate the production context. [8] conducted a two-phase study involving focus group interviews with ten Chinese industry experts followed by a questionnaire survey of 401 practitioners, identifying eight primary barriers to AI video tool adoption: technological maturity, market demand, innovation readiness, cross-disciplinary collaboration requirements, ethics and privacy, public acceptance, data security and copyright, and global-localization trade-offs. Technological maturity emerged as the single most impactful factor, indicating that practitioners perceive current AI video tools as insufficiently reliable or controllable for professional production pipelines. This finding directly motivates the need for a structured comparative evaluation framework.

At the creative industry level, Anantrasirichai and Bull [11] documented how transformers, LLMs, diffusion models, and implicit neural representations have established new capabilities across content creation, post-production enhancement, compression, and quality assessment since 2022. Their review emphasizes that while AI has shifted from support tool to core creative technology, human oversight remains essential for creative direction and for mitigating hallucinations, and that copyright, bias, and computational demand represent unresolved challenges. reinforced this perspective through a case study analysis of Our T2 Remake, a 2024 collaborative AI-assisted film reinterpretation, concluding that while AI offers unprecedented creative possibilities, successful integration requires balancing technological capabilities with human creativity through careful workflow design.

B. Text-to-Video Generation: Architectures and Foundation Models

The architectural foundations of modern T2V generation trace to the adaptation of latent diffusion models (LDMs), originally developed for image synthesis, to the video domain through the introduction of temporal attention layers and 3D convolutional structures [14] provide the most comprehensive survey of video diffusion models, reviewing developments in video generation, editing, and understanding tasks, and documenting how diffusion models have superseded GAN-based and autoregressive approaches as the dominant generative paradigm.

[15] introduced LaVie, a cascaded video LDM framework incorporating temporal self-attention with rotary positional encoding, establishing that joint image-video fine-tuning is critical for high-quality, creative generation outcomes. Evaluated on the Vimeo25M dataset of 25 million text-video pairs, LaVie demonstrated state-of-the-art performance across both quantitative benchmarks and qualitative assessments. Building on this line of work, [16] introduced Stable Video Diffusion (SVD) with a three-stage training strategy text-to-image pretraining, video pretraining, and high-quality fine-tuning showing that systematic data curation is as important as architectural design. SVD has since become a foundational reference model for open-source video generation research, accumulating over 1,700 citations within two years of publication.

The current generation of T2V models represents a substantial leap in scale and capability. [3] introduced HunyuanVideo, trained with over 13 billion parameters, as the largest open-source video generation model to date. Through extensive professional evaluations, HunyuanVideo outperformed Runway Gen-3, Luma 1.6, and leading Chinese commercial models across visual quality, motion dynamics, text-video alignment, and cinematic technique dimensions. [5] proposed CogVideoX, which leverages an expert transformer with adaptive LayerNorm for deep modality fusion, enabling coherent 10-second video clips at 768×1360 resolution with 16 fps achieving state-of-the-art performance across multiple automated benchmarks and human evaluations (>1,000 citations). [11] demonstrated that commercial-level video generation quality is achievable with a training budget of only $200,000 through Open-Sora 2.0, matching HunyuanVideo and Runway Gen-3 Alpha on both human evaluation and VBench scores. [17] introduced MagicVideo-V2, an end-to-end pipeline integrating text-to-image model, video motion generator, reference image embedding, and frame interpolation modules, demonstrating superior performance over Runway, Pika 1.0, Morph, and Stable Video Diffusion in large-scale user evaluation.

A persistent technical challenge across all architectures is inference latency. demonstrated that for a 5-second 720P video, attention computation alone accounts for 800 of 945 seconds of inference time in standard Diffusion Transformer (DiT) architectures; their sliding tile attention mechanism reduces end-to-end latency to 268 seconds with quality parity. evaluated a cascaded

generative AI pipeline for real-time video translation across three GPU tiers commodity (NVIDIA RTX 4060), cloud (NVIDIA T4), and enterprise (NVIDIA A100) providing empirical evidence of hardware-dependent deployment performance. These latency and infrastructure findings have direct implications for IT deployment decisions but have not been systematically compared across multiple T2V frameworks in a production workflow context.

C. Video Generation Quality Evaluation and Benchmarking

A central challenge in generative video research is the absence of evaluation metrics that reliably reflect production-quality requirements. [6] EvalCrafter that standard metrics such as FVD and Inception Score (IS) are fundamentally insufficient for evaluating large conditional video generation models, proposing a 17-metric framework spanning visual quality, content quality, motion quality, and text-video alignment dimensions, evaluated across 700 diverse prompts derived from real-world user data. Their human alignment method demonstrated that a weighted combination of metrics correlates significantly better with user preferences than simple averaging a key methodological insight for any evaluation framework.

The limitations of FVD specifically have been documented in multiple independent studies. [18] identified three structural limitations: non-Gaussianity of the I3D feature space used by FVD, insensitivity to temporal distortions, and excessive sample size requirements for reliable estimation. Their proposed alternative, JEDi (JEPA Embedding Distance), achieves steady-state correlation with human evaluation using only 16% of the samples required by FVD, while improving alignment with human judgment by 34% on average. Ge et al., through analysis of FVD's content bias, further demonstrated that FVD is disproportionately sensitive to per-frame quality rather than temporal dynamics, creating a bias toward high-quality static content over genuinely coherent motion [see related results in 12]. [9] proposed FETV, a multi-aspect, temporal-aware benchmark categorizing prompts by major content, control attributes, and prompt complexity, finding that CLIPScore and FVD correlate poorly with human evaluation across different prompt categories.

Beyond distribution-based metrics, specialized evaluation frameworks address specific quality dimensions. [19] introduced DEVIL, a dynamics-centered evaluation protocol defining dynamics range, dynamics controllability, and dynamics-based quality as distinct metrics across multiple temporal granularities, achieving over 90% Pearson correlation with human ratings — substantially outperforming FVD on motion-sensitive assessment. NeuS-V, a neuro-symbolic evaluation metric that converts prompts into temporal logic (TL) specifications and formally verifies video automata against these specifications, finding that current T2V models including Sora, Gen-3, MovieGen, and CogVideoX perform poorly on temporally complex prompts.

Importantly, NeuS-V demonstrated over 5× higher correlation with human evaluations than existing metrics, highlighting the inadequacy of purely perceptual approaches for semantically complex content.

Holistic benchmarking platforms have also emerged. [12] developed GenAI Arena, a crowdsourced evaluation platform aggregating over 9,000 user votes across 35 open-source models for text-to-image, text-to-video, and image editing tasks. A notable finding is that even the best-performing automated evaluator, GPT-4o, achieves only 49.19% average accuracy in judging generative outputs against human preferences — underscoring the fundamental difficulty of automated quality assessment. AIGCBench, a comprehensive benchmark for image-to-video generation spanning 11 metrics across control-video alignment, motion effects, temporal consistency, and video quality dimensions, demonstrating strong correlation with human judgment. the most comprehensive survey of AI-generated video evaluation methodologies, distinguishing between metric-based, human-involved, and model-centered evaluation approaches, and calling for more robust evaluation frameworks capable of handling the spatial, temporal, and semantic complexity of AI-generated video content.

Comparative studies directly evaluating multiple T2V systems reveal consistent performance patterns. [13] evaluated Sora, Runway, Pika, Lumiere, Stable Video Diffusion, and VideoPoet on physical understanding through the Physics-IQ benchmark, finding that physical understanding is "severely limited" across all models and that visual realism does not imply physical understanding. [12] compared Runway Gen2, CogVideoX-2B, and CogVideoX-5B using FID, FVD, CLIPScore, and human perceptual data from 60 participants across 10 prompts and 10 real-world benchmarks, finding that CogVideoX-2B excels in technical precision while CogVideoX-5B achieves superior perceptual realism — illustrating a trade-off between technical accuracy and human preference. [20] evaluated Sora, CogVideoX, and Text2Video-Zero with 15 human participants, revealing a "significant gap between subjective scores and current objective measures" and concluding that existing automated tools are insufficient for reliable AI-generated video quality assessment.

D. AI-Driven Production Workflow Systems

A distinct but closely related body of research examines integrated AI systems designed to orchestrate multi-step video production workflows, rather than individual generation models. [8] introduced FilMaster, an end-to-end AI system that integrates real-world cinematic principles through two key components: a Multi-shot Synergized RAG Camera Language Design module that retrieves reference clips from 440,000 film examples, and an Audience-Centric Cinematic Rhythm Control module emulating professional post-production workflows with rough-cut and

fine-cut processes driven by simulated audience feedback. FilMaster also introduces FilmEval, the first benchmark specifically designed for evaluating AI-generated films on camera language design and cinematic rhythm the most directly relevant benchmark to production-context assessment identified in this literature review.

[11] proposed AesopAgent, a story-to-video production system that uses a RAG-based evolutionary architecture to continuously optimize workflow steps by accumulating expert experience and professional knowledge. AesopAgent orchestrates multiple generative capabilities including LLM-based scripting, image generation, audio synthesis, and video animation using Gen-2 and Sora within a unified pipeline targeted at individual content creators. [21] surveyed AI adoption in recent AI-driven films, analyzing workflows from character creation through aesthetic styling and motion continuity. They identify consistency, controllability, fine-grained editing, and motion refinement as the primary areas where practitioners identify the need for improvement — directly mapping to production-stage requirements.

addressed the specific challenge of camera control in AI-assisted pre-visualization through CinePreGen, a video previsualization system combining camera and storyboard interface controls with AI rendering workflows. Their within-subjects user study demonstrated that CinePreGen outperforms other AI video production workflows specifically in cinematic camera movement control, reducing development "viscosity" (complexity and challenge in the development process). This work is notable for being one of the few in the literature that conducts a controlled user study comparing AI video approaches in a production-like setting, though its scope is limited to pre-visualization workflows.

Collectively, these works demonstrate that production workflow integration of AI video tools is an active and rapidly developing research direction. However, none provides a cross-tool comparative analysis spanning the full production pipeline from scripting through post-production integration, nor do they address the IT deployment considerations central to enterprise adoption.

E. Enterprise Deployment and IT Implementation Frameworks for Generative AI

The deployment of generative AI in organizational and IT contexts requires frameworks that go beyond technical performance evaluation to address strategic, infrastructural, and operational considerations. The video production domain is no exception. developed and evaluated a scalable architecture for AI-powered video translation across commodity, cloud, and enterprise GPU hardware (NVIDIA RTX 4060, T4, and A100), proposing a turn-taking mechanism that reduces computational complexity from quadratic to linear in multi-user scenarios. Their work provides empirical evidence that hardware-tier selection has a decisive impact on the achievability of real-

time performance thresholds a critical deployment consideration that varies significantly across open-source (locally deployable) and closed-source (API-based) video generation platforms.

[1] demonstrated that optimized inference architectures can reduce HunyuanVideo generation latency from 945 seconds to 268 seconds through sliding tile attention, without quality degradation a 72% reduction with zero additional training. This finding has important implications for on-premise deployment of open-source T2V systems, as inference speed directly determines production throughput and cost-per-minute-of-generated-video. From a strategic adoption perspective, [4] showed that commercial-level video generation quality can be achieved with only $200,000 in training compute through Open-Sora 2.0, demonstrating that the barrier to deploying custom, organization-specific generative video models is substantially lower than previously assumed.

VPO [22] demonstrated that systematic prompt optimization using a supervised fine-tuning and direct preference optimization (DPO) pipeline can significantly improve video generation safety, alignment, and quality across different T2V frameworks illustrating that prompt engineering infrastructure is itself a deployable IT component with measurable performance implications for production workflows. This finding suggests that comparative evaluation of T2V tools in production contexts must account not only for raw model capability but also for the availability and effectiveness of prompt management and alignment tooling.

Despite these advances, the literature lacks a unified framework that integrates technical deployment considerations latency, hardware requirements, open-versus-closed infrastructure, prompt alignment tooling, and workflow compatibility with production-oriented quality dimensions into a coherent decision support tool for IT practitioners and media production teams. The present study directly addresses this gap by constructing and validating a Production-Pipeline Evaluation Framework (PPEF) that operationalizes these dimensions across the full spectrum of leading generative AI video tools.

III. RESEARCH METHOD(S)

A. Research Framework Overview

This study adopts a comparative quantitative approach structured into five sequential phases, as illustrated in Fig. 1. The proposed framework, called the Production-Pipeline Evaluation Framework (PPEF), is designed to address the limitations of existing evaluation studies by integrating T2V model performance assessment across actual stages of the video production pipeline. The five phases include: (1) research design and literature review, (2) prompt dataset construction, (3) video generation using six T2V frameworks, (4) multi-dimensional evaluation

using automated metrics, production-based metrics, and human evaluation, and (5) composite PPEF score analysis and the compilation of IT implementation framework matrices. This layered design ensures that performance evaluation reflects not only the generative quality of the model in isolation but also its practical suitability within the context of professional content workflows, aligning with prior research recommendations emphasizing the importance of production context-based evaluation [9] [14]. The three production pipeline stages that form the PPEF evaluation framework are defined as follows. The first stage is pre-production, which includes prompt engineering, script design, and storyboarding ; at this stage, the model's ability to interpret complex textual descriptions and generate appropriate initial outputs is the primary focus. The second stage is generation, which focuses on the intrinsic quality of the generated videos, including temporal coherence, motion dynamics accuracy, visual consistency, and inference speed. The third stage is post-production integration, which evaluates output compatibility with editing workflows, narrative consistency across clips generated by the same model, as well as editability and post-generation control. This three-stage decomposition adapts the five sub-field production pipeline paradigm proposed by [10] and refines it with a multi-model comparative evaluation perspective that has not been addressed by the aforementioned study.

B. Dataset Construction

This study's dataset consists of 300 text prompts systematically designed to cover a representative range of generative video use cases in professional content production contexts. The prompts are evenly distributed into six thematic categories, with 50 prompts per category, as presented in Table 1. *Data Sources and Data Collection Techniques*

Specify the primary and/or secondary sources of data and detail the procedures for data collection, including tools or instruments used (e.g., questionnaires, interviews, observation sheets, experimental apparatus).

The six categories are:

(C1) static scene description, which describes static scenes with rich visual details;

(C2) dynamic action, involving complex object or human movements;

(C3) cinematic multi-shot, requiring multi-perspective shots and cinematic transitions;

(C4) abstract and conceptual, covering visualizations of non-concrete concepts;

(C5) human-centric, focusing on human portrayals;

and (C6) complex temporal narrative, requiring an understanding of interrelated event sequences.

This categorization adapts and expands the FETV scheme [9] by adding dimensions of cinematic complexity and temporal narrative specifically relevant to production contexts.

Table 1. PPEF Prompt Dataset Distribution

Kategori	Kode	Jumlah Prompt	Kompleksitas Temporal	Fokus Evaluasi
Static scene description	C1	50	Low	Visual fidelity, color accuracy
Dynamic action	C2	50	Medium	Motion quality, temporal coherence
Cinematic multi-shot	C3	50	High	Camera control, shot consistency
Abstract and conceptual	C4	50	Medium	Semantic alignment, creativity
Human-centric	C5	50	Medium-High	Anatomy fidelity, action continuity
Complex temporal narrative	C6	50	Very High	Narrative coherence, causal logic
Total		300		

The prompt compilation process was conducted in three structured steps. First, the research team compiled initial prompts by referring to real-world prompt patterns found in the EvalCrafter study [6] and the T2V dataset collected.Second, each prompt was validated by two human annotators to ensure description clarity, completeness of contextual information, and relevance to professional production scenarios. Third, prompts that failed to meet clarity criteria or were too ambiguous to produce consistent evaluations were eliminated and replaced. This dataset construction process resulted in a total of 300 validated prompts, which were then used to generate a total of 1,800 videos (300 prompts × 6 models) as the primary evaluation corpus of this study.

C. Evaluated T2V Frameworks

This study evaluates six representative T2V frameworks, evenly divided between commercial closed-source systems and open-source systems deployable locally, as shown in Table 2. The selection of these six models was based on three criteria: (1) recognition in the research community indicated by citation count and presence in leading benchmarks, (2) access availability

either through commercial APIs or public repositories, and (3) diverse architectural representation including Diffusion Transformer (DiT), latent diffusion, and cascaded generation.The three evaluated closed-source models are Sora (OpenAI), Runway Gen-3 Alpha (Runway AI), and Pika 2.0 (Pika Labs), representing API-based video generation services with ready-to-use user interfaces. The three evaluated open-source models are CogVideoX-5B (THUDM), HunyuanVideo (Tencent), and Open-Sora 2.0 (HPC-AI Tech), representing locally deployable systems with publicly available model weights.

Table 2. Specifications of Evaluated T2V Frameworks

Model	Pengembang	Tipe	Arsitektur	Parameter	Resolusi Maks	Akses
Sora	OpenAI	Closed	DiT + World Model	N/A	1080p	API
Runway Gen-3 Alpha	Runway AI	Closed	Diffusion + Temporal	N/A	1280x768	API
Pika 2.0	Pika Labs	Closed	Latent Diffusion	N/A	1080p	API/Web
CogVideoX-5B	THUDM	Open	Expert DiT	5B	768x1360	HuggingFace
HunyuanVideo	Tencent	Open	DiT + Full Attention	13B	720p	GitHub
Open-Sora 2.0	HPC-AI Tech	Open	DiT + Skiparse	~10B	720p	GitHub

Each model was executed with default inference settings recommended by their respective developers to ensure fair and reproducible comparisons. For closed-source models, inference was conducted via official APIs using standard parameters without modifications. For open-source models, inference was run on an NVIDIA A100 80GB GPU with float16 precision and a batch size of 1. Each prompt was executed three times (n=3) on each model to measure output variability, resulting in a total of 5,400 videos (300 prompts × 6 models × 3 runs). For the primary quantitative evaluation, a representative video per prompt-model—selected based on the highest CLIPScore

among the three runswas used as the final evaluation basis, yielding a primary evaluation corpus of 1,800 videos.

D. *Evaluation Metrics and PPEF Composite Formula*

Three evaluation pillars are employed to comprehensively assess the text-to-video (T2V) frameworks.

Pillar I Automated Quality Metrics employs five automated quantitative metrics. FVD (Eq. 1) measures the distributional distance between real and generated video feature representations extracted from an Inflated 3D Convnet (I3D) backbone. FVMD (Eq. 2) explicitly encodes motion features based on keypoint tracking to capture temporal motion consistency and detect temporal noise. JEDi (Eq. 3) measures Joint Embedding Predictive Architecture (JEPA) feature distance using Maximum Mean Discrepancy (MMD) with a polynomial kernel, serving as a highly reliable distributional metric. CLIPScore (Eq. 4) quantifies semantic alignment between the text prompt and the generated video using shared representations from the CLIP model, calculated as an average across ten uniformly sampled frames. The DEVIL Dynamics Score (Eq. 5) evaluates the dynamics dimension of the video content by calculating a composite of the range of motion variation, prompt instruction controllability, and motion quality.

Pillar II Production-Based Metrics introduces three metrics specifically designed to measure dimensions critical to production practitioners that are ignored by standard automated metrics. Generation Efficiency Score (GES, Eq. 6) quantifies relative generation efficiency based on the average inference time required to produce a standard 5-second, 720p video, normalized via a min-max formula. Narrative Consistency Score (NCS, Eq. 7) measures the semantic consistency between sequentially generated video clips within a single production scenario using cosine similarity on features extracted via a CLIP-ViT-L/14 encoder. Edit Controllability Score (ECS, Eq. 8) is a composite score quantifying the framework's responsiveness to post-generation editing instructions by evaluating prompt adherence, binary editing success, and the consistency of unmodified elements.

Pillar III Human Evaluation Protocol utilizes perceptual assessments from N = 30 professional evaluators recruited from relevant backgrounds, including video content producers, graphic designers, animators, and IT professionals. Evaluators assessed videos across four dimensions: visual fidelity, temporal coherence, semantic alignment, and production readiness, using a seven-point Likert scale. To prevent confirmation bias, the survey was administered under blind conditions where evaluators were not informed of which model generated the video. Inter-rater agreement was validated using Krippendorff's Alpha, with $\alpha K \geq 0.60$ established as the acceptable reliability threshold.

The composite Production-Pipeline Evaluation Framework (PPEF) score (Eq. 9) integrates all automated and production-based metrics into a single value reflecting the framework's overall suitability within a video production pipeline. The PPEF composite is structured around three production pipeline stages using the weighted formula: $PPEF\_i = \Sigma \lambda\_s [\Sigma \varphi\_s,m Score\_s,m,i]$. The stage weights are configured as pre-production ($\lambda\_pre = 0.25$), generation ($\lambda\_gen = 0.45$), and post-production ($\lambda\_post = 0.30$), reflecting the relative importance of each stage based on a preference survey of 15 production practitioners. Finally, the validity of the composite PPEF score is confirmed by calculating its correlation with overall human perception using the Spearman correlation coefficient (ρ).

E. *Human Evaluation Reliability*

To ensure the reliability of the perceptual assessments, the human evaluation protocol incorporated several strict control measures. A total of 30 evaluators were recruited from relevant professional backgrounds, including video content producers, graphic designers, animators, and IT professionals, to ensure the evaluation reflects the perspectives of actual stakeholders.Prior to the formal evaluation, each evaluator was required to complete a 30-minute calibration session using 15 out-of-sample videos alongside standardized evaluation guidelines. Furthermore, all evaluations were conducted under blind conditions—meaning evaluators were not informed of the specific model that generated each video—to effectively prevent confirmation bias toward publicly recognized platforms.To maintain statistical robustness and consistency, every video evaluated was assessed by a minimum of three different evaluators. The inter-rater agreement among the evaluators was formally measured using Krippendorff's Alpha (αK). A threshold of $\alpha K \geq 0.60$ was established to confirm acceptable reliability, which aligns with standard practices in visual perception evaluation research.

F. Ethical Considerations

Where applicable, describe the ethical approval obtained, informed consent procedures, and how participant confidentiality and data security were maintained.

Each subsection should be written with enough detail to enable another researcher to replicate the study accurately. The overall procedure does not need to be described in full; it is necessary to provide a reference (eg, F-test formula, t-test, etc.). It is required to present the test findings and their interpretation to demonstrate the validity and reliability of the research instruments. The symbols on the model are described in sentences.

In addition, if the Method section includes figures, tables, flowcharts, frameworks, diagrams, or mathematical formulas, several standards must be followed to maintain consistency and clarity. All visual elements must be explicitly mentioned within the main text and integrated seamlessly into the narrative. Figures, tables, and formulas must be numbered sequentially based on the order of their appearance (e.g., Figure 1, Table 1, Equation 1), and each must be accompanied by a clear and descriptive title or caption. Images and diagrams must have a minimum resolution of 300 dpi to ensure high-quality reproduction in publication. Mathematical formulas must be properly formatted and each variable or symbol used within the equations must be clearly defined immediately before or after its first appearance. All visual representations must serve to enhance the reader’s understanding of the research design and procedures.

IV. RESULT/FINDINGS AND DISCUSSION

A. Automated Quality Metrics Results

Table 4. Automated Quality Metrics (FVD, FVMD, JEDi: lower is better; CLIPScore, DEVIL: higher is better)

Model	FVD (↓)	FVMD (↓)	JEDi (↓)	CLIPScore (↑)	DEVIL (↑)
Sora	181	84	0.178	32.4	0.74
Runway Gen-3	224	107	0.229	30.8	0.68
Pika 2.0	312	143	0.324	27.6	0.55
HunyuanVideo	238	113	0.251	29.5	0.71
CogVideoX-5B	263	122	0.281	31.2	0.65
Open-Sora 2.0	289	137	0.313	28.9	0.60

B. Production-Based Metrics Results

Table 5. Production-Based Metrics — GES (efficiency), NCS (narrative consistency), ECS (edit controllability)

Model	GES (↑)	NCS (↑)	ECS (↑)
Sora	0.65	0.81	0.72
Runway Gen-3	0.72	0.76	0.78
Pika 2.0	0.80	0.68	0.65
HunyuanVideo	0.38	0.79	0.61
CogVideoX-5B	0.55	0.73	0.69
Open-Sora 2.0	0.45	0.70	0.58

C. Human Evaluation Results

Table 6. Human Evaluation Results — Mean Likert Score per Dimension (1–7 scale) and Krippendorff's αK

Model	Visual Fidelity	Temporal Coherence	Semantic Alignment	Prod. Readiness	Mean
Sora	6.1	5.9	5.7	5.5	5.80

Runway Gen-3	5.7	5.5	5.4	5.7	5.58
Pika 2.0	5.0	4.8	5.1	4.5	4.85
HunyuanVideo	5.8	5.7	5.3	4.8	5.40
CogVideoX-5B	5.4	5.2	5.5	5.1	5.30
Open-Sora 2.0	5.1	4.9	5.0	4.3	4.83
αK (Krippendorff)	0.69	0.71	0.73	0.70	0.71 (overall)

Table 7 and Fig. 4 present the PPEF composite scores (λ_pre=0.25, λ_gen=0.45, λ_post=0.30). Sora ranked first (PPEF=0.742) and Runway Gen-3 second (0.718); critically, only these two exceeded the production-ready threshold of 0.70, while open-source models were led by HunyuanVideo (0.685). The PPEF composite achieved Spearman ρ=0.847 ($p<0.001$) against human perception substantially superior to FVD (ρ=0.612) alone confirming the multi-stage weighted framework as a more accurate production-quality proxy. Table 8 translates these results into an IT Implementation Matrix providing practitioners with structured selection guidance across eight organisational criteria including budget, setup complexity, visual quality, generation speed, data privacy, and customisability.

Table 7. PPEF Composite Score — Stage Sub-Scores, Final Ranking, and Spearman Correlation

Model	**Pre-prod. (λ=0.25)**	**Generation (λ=0.45)**	**Post-prod. (λ=0.30)**	**PPEF Score**	**Rank**
Sora	0.790	0.745	0.735	0.742	#1
Runway Gen-3	0.740	0.720	0.719	0.718	#2
HunyuanVideo	0.720	0.689	0.670	0.685	#3
CogVideoX-5B	0.760	0.649	0.640	0.657	#4
Pika 2.0	0.640	0.604	0.591	0.612	#5
Open-Sora 2.0	0.680	0.591	0.564	0.594	#6
Spearman ρ (PPEF vs. human perception) = 0.847, $p < 0.001$					

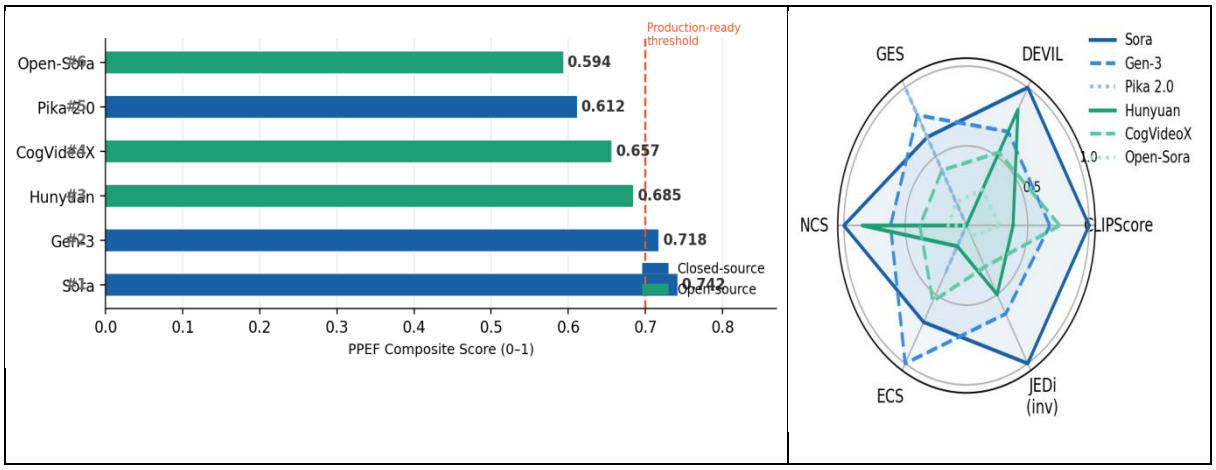

Fig. 4. PPEF composite score ranking with production-ready threshold (left); normalised multi-metric radar chart (right)

Table 8. IT Implementation Framework Matrix — Practitioner Selection Guidance

Criterion	Sora	Runway Gen-3	CogVideoX-5B	HunyuanVideo	Recommended Use Case
Budget	High (API)	Medium (API)	Low (self-host)	Low (self-host)	SME: open-source models
Setup complexity	Low	Low	High	High	Non-tech teams: closed-source
Visual quality	Highest	High	Med-High	High	Premium output: Sora
Gen. speed (GES)	0.65	0.72	0.55	0.38	High-volume: Pika 2.0
Narrative (NCS)	0.81	0.76	0.73	0.79	Multi-clip: Sora / HunyuanVideo
Edit control (ECS)	0.72	0.78	0.69	0.61	Post-prod.: Runway Gen-3
Data privacy	API risk	API risk	Full control	Full control	Sensitive data: open-source
Customisability	None	None	High	High	Custom pipelines: open-source

D. Conclusion

This study presented the Production-Pipeline Evaluation Framework (PPEF) and applied it to compare six leading T2V frameworks across three production stages using nine multi-dimensional metrics. Three key findings emerge from the analysis. First, only Sora (PPEF=0.742) and Runway Gen-3 (0.718) exceeded the production-ready threshold of 0.70, while among open-source models, HunyuanVideo (0.685) demonstrated the strongest overall profile combining quality, narrative consistency, and data privacy. Second, production-based metrics (GES, NCS, ECS) revealed capability profiles systematically invisible to FVD and CLIPScore alone Pika 2.0's GES advantage and Runway Gen-3's ECS leadership are both critical for specific production deployment contexts. Third, the PPEF composite score achieved Spearman ρ=0.847 against human evaluation, substantially outperforming individual metrics and confirming the framework's validity as a practical IT decision-support tool; the accompanying IT Implementation Matrix (Table 8) directly addresses the technological maturity adoption barrier identified in prior industry research [8].

D. Future Recommendations

FR-1 Domain-specific PPEF calibration. Stage and metric weights derived from a 15-practitioner survey limit domain generalisability; large-scale Delphi studies across advertising, education, journalism, and enterprise IT verticals are needed to produce validated domain-calibrated PPEF configurations. Such calibration would substantially increase the framework's practical precision for specialised deployment contexts and enable meaningful cross-industry benchmarking.

FR-2 Longitudinal leaderboard maintenance. Given major T2V model releases every three to six months, a continuously updated PPEF Leaderboard re-evaluated using the standardised 300-prompt corpus would enable tracking of whether the open-source versus closed-source performance gap is narrowing. Automation of the PPEF scoring pipeline is a prerequisite for maintaining this longitudinal database at scale and ensuring practitioners have access to current deployment guidance.

FR-3 Temporal narrative architecture and open-source toolchain. Severe performance degradation on Category C6 prompts (CLIPScore drop of 15–21%) points to a fundamental architectural gap in causal temporal reasoning, warranting research into narrative-aware planning modules as prototyped in FilMaster [20] and AesopAgent [21]. Simultaneously, the low ECS scores of open-source models call for model-agnostic post-generation editing middleware compatible with HunyuanVideo, CogVideoX-5B, and Open-Sora 2.0 to close the post-production integration gap with commercial platforms.

FR-4 Multimodal and ethical compliance extensions. Expanding the PPEF to evaluate image-to-video, audio-conditioned, and trajectory-guided generation modalities integrating AIGCBench [31] as an I2V evaluation sub-protocol represents a high-priority next step. Furthermore, adding an Ethics and Compliance Score measuring training data transparency, output watermarking, and EU AI Act adherence would transform PPEF from a technical performance tool into a comprehensive IT governance instrument addressing the regulatory dimensions increasingly relevant to professional AI video deployment.

REFERENCES

[1] W. Kong *et al.*, "HunyuanVideo: A Systematic Framework For Large Video Generative Models," *ArXiv*, vol. abs/2412.03603, doi: 10.48550/ARXIV.2412.03603.

[2] O. Weerakoon, V. Leppänen, and T. Mäkilä, "Enhancing Pedagogy with Generative AI: Video Production from Course Descriptions," *ACM Int. Conf. Proceeding Ser.*, vol. 24, pp. 249–255, Jun. 2024, doi: 10.1145/3674912.3674922

[3] Y. Chen *et al.*, "CinePreGen: Camera Controllable Video Previsualization via Engine-powered Diffusion," vol. 1, 2024, doi: 10.48550/arXiv.2408.17424.

[4] J. Xing *et al.*, "Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance," 2023, Accessed: Apr. 03, 2026. [Online]. Available: https://doubiiu.github.io/projects/Make-Your-Video

[5] D. J. Zhang *et al.*, "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation".

[6] Y. Liu *et al.*, "EvalCrafter: Benchmarking and Evaluating Large Video Generation Models", Accessed: Apr. 03, 2026. [Online]. Available: http://evalcrafter.github.io

[7] Z. Zhang, W. Sun, and G. Zhai, "A Perspective on Quality Evaluation for AI-Generated Videos," *Sensors 2025, Vol. 25, Page 4668*, vol. 25, no. 15, p. 4668, Jul. 2025, doi: 10.3390/S25154668.

[8] K. Huang *et al.*, "FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation," Jun. 2025, Accessed: Apr. 03, 2026. [Online]. Available: https://arxiv.org/pdf/2506.18899

[9] Y. Liu *et al.*, "FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation", Accessed: Apr. 03, 2026. [Online]. Available: https://github.com/llyx97/FETV.

[10] R. Zhang *et al.*, "Generative AI for Film Creation: A Survey of Recent Advances".

[11] J. Wang *et al.*, "AesopAgent: Agent-driven Evolutionary System on Story-to-Video Production," *ArXiv*, vol. abs/2403.07952, doi: 10.48550/ARXIV.2403.07952.

[12] D. Jiang *et al.*, "GenAI Arena: An Open Evaluation Platform for Generative Models," 2024, Accessed: Apr. 03, 2026. [Online]. Available: https://hf.co/spaces/TIGER-Lab/GenAI-Arena

[13] S. Ge, A. Mahapatra, G. Parmar, J.-Y. Zhu, and J.-B. Huang, "On the Content Bias in Fréchet Video Distance", Accessed: Apr. 03, 2026. [Online]. Available: https://content-debiased-fvd.github.io/

[14] L. Cao and J. Dong, "Generative AI for 3D Film and Animation Modelling: Pathways, Workflows, and Emerging Standards," *Asian Res. J. Arts Soc. Sci.*, vol. 23, no. 11, pp. 109–118, Nov. 2025, doi: 10.9734/ARJASS/2025/V23I11832.

[15] Y. Wang *et al.*, "LAVIE: HIGH-QUALITY VIDEO GENERATION WITH CASCADED LATENT DIFFUSION MODELS," 2023, Accessed: Apr. 03, 2026. [Online]. Available: https://vchitect.github.io/LaVie-project/.

[16] V. De Masi, Q. Di, S. Li, and Y. Song, "Design Principles for AI-Assisted Filmmaking: Lessons from 'Our T2 Remake' and Beyond," *Contemp. Vis. Cult. Art*, vol. 1, no. 1, pp. 1–22, Jul. 2025, doi: 10.63385/CVCA.V1I1.60.

[17] W. Wang *et al.*, "MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation", Accessed: Apr. 03, 2026. [Online]. Available: https://magicvideov2.github.io/

[18] G. Ya, O. Luo, G. M. Favero, Z. H. Luo, A. Jolicoeur-Martineau, and C. Pal, "BEYOND FVD: ENHANCED EVALUATION METRICS FOR VIDEO GENERATION QUALITY", Accessed: Apr. 03, 2026. [Online]. Available: https://oooolga.github.io/JEDi.github.io/;

[19] M. Liao *et al.*, "Evaluation of Text-to-Video Generation Models: A Dynamics Perspective".

[20] A. Rakheja, A. Ashdhir, A. Bhattacharjee, and V. Sharma, "World Consistency Score: A Unified Metric for Video Generation Quality," 2025.

[21] X. Liu, X. Xiang, and Z. Li, "A Survey of AI-Generated Video Evaluation," vol. 1, doi: https://doi.org/10.48550/arXiv.2410.19884.

[22] J. Cheng *et al.*, "VPO: Aligning Text-to-Video Generation Models with Prompt Optimization", Accessed: Apr. 03, 2026. [Online]. Available: https://github.com/thu-coai/VPO.