Vol. 01 No. 01 (2025)
e-ISSN: 3141-1371
https://ejournal.unipsaas.edu.my/

CULTURAL BIAS AND TECHNICAL GLITCHES IN ARTIFICAL INTELLIGENCE (AI) VIDEO PRODUCTION FOR HIGHER EDUCATION: FROM PROMPT TO SCREEN

Muhammad Hisyamudin Baharudin[1]

[1]Creative Multimedia and Computing Department, Faculty of Management and Informatics,
Al-Sultan Abdullah Ahmad Shah Quranic University of Pahang (UNIQSAAS)

Abstract

The rapid evolution of Generative Artificial Intelligence (AI) has extended beyond text and static imagery into the realm of realistic text-to-video synthesis. This article critically examines the practical and technical challenges of utilizing the latest generative video models (specifically the Google Veo model) for higher education content creation. While the technology offers significant potential for accelerating creative workflows in higher education, its application is constrained by distinct limitations. Adopting a Practice-based Research methodology, this study analyzes a corpus of 15 experimental video projects (n=10 convocation promotions; n=5 academic program promotions) produced using an AI-assisted workflow. Findings highlight three primary categories of limitations; (i) Cultural and Regional Bias, where the model hallucinates foreign cultural elements (e.g., Indonesian demographics) over local Malaysian contexts; (ii) Linguistic and Phonetic Inaccuracies, specifically the inability of Text-to-Speech engines to process local dialects and phonemes; and (iii) Physical and Logical Hallucinations, such as defying laws of physics or adding unauthorized visual artifacts.

Keywords: Generative AI, text-to-video, veo model, veo3, practice-based research, cultural bias, prompt engineering

Perkembangan Artikel

Diterima: 16/12/2025
Disemak: 24/12/2025
Diterbit: 31/12/2025

*Corresponding Author:
Creative Multimedia and Computing Department, Faculty of Management and Informatics, Al-Sultan Abdullah Ahmad Shah Quranic University of Pahang (UNIQSAAS).
Email:
hisyam@uniqsaas.edu.my

INTRODUCTION

In the last decade, the digital technology landscape has undergone a paradigm shift through the emergence of Generative AI. While recent years witnessed the dominance of Large Language Models (LLMs) like GPT-4 in text generation and diffusion models like Midjourney in static image generation, the technological frontier has now shifted to a more complex domain: text-to-video generation. This development is not merely a technical evolution but has the potential to restructure the entire media production, education, and broadcasting ecosystem (Baidoo-Anu & Owusu Ansah, 2023).

Generative video technology represents a significant leap because it combines multiple AI capabilities, which are natural language understanding, image synthesis, and motion prediction, into a single workflow. Unlike traditional video production, which demands extensive human resources and time, AI-driven video generation promises rapid prototyping and cost efficiency. This potential has attracted attention from industries such as advertising, entertainment, and education, where visual storytelling plays a critical role in audience engagement (Ma et al., 2025; Ruan et al., 2023).

Video generation technology, particularly pioneered by advanced models such as Veo, operates based on Video Diffusion Models (VDM). Unlike static image generation, which requires visual coherence in a single frame, VDMs must maintain consistency across the temporal dimension. This implies the model must understand not only the object's appearance but also how it moves, interacts with the environment, and adheres to physics within a continuous timeline (Ho et al., 2020).

However, keeping things consistent over time makes the system harder to manage and adds new technical challenges. For instance, generative models must simulate realistic motion, lighting changes, and object interactions without breaking visual continuity. These requirements make video generation far more computationally demanding than image synthesis, often requiring advanced GPU clusters and optimized diffusion algorithms (Ma et al., 2025; Yu et al., 2024). Furthermore, the integration of audio elements such as speech and ambient sound adds another layer of complexity, especially when adapting to multilingual or culturally specific contexts (Ruan et al., 2023).

However, despite the enthusiasm surrounding this technology, practical implementation faces critical constraints. Popular narratives often depict AI as a magic wand capable of fully replacing human roles, yet technical reviews reveal a significant quality gap when applied to professional or academic purposes requiring high precision. Issues such as visual hallucinations, failure to maintain subject consistency, and disturbing flaws often hinder the production of commercially viable work (Kietzmann et al., 2020).

The AI video can suffer from visual hallucinations where it mistakenly creates weird or impossible objects, like a person with extra fingers or things floating in mid-air. Then, the model often fails to keep the subject consistent, meaning a character’s face or clothing might suddenly change from one second

to the next. Lastly, there are disturbing flaws like flickering lights or shaky movements that make the video look unnatural and glitchy to the human eye.

Beyond technical limitations, cultural and linguistic factors further complicate adoption in higher education. For example, generative models trained predominantly on Western datasets may inadvertently introduce foreign cultural elements into Malaysian academic content, leading to misrepresentation. Similarly, text-to-speech engines often struggle with local dialects and phonetic nuances, reducing the authenticity of generated videos (Lu et al., 2025; Tejas et al., 2026).

Therefore, this article aims to critically analyze the technical and operational challenges in producing videos using Veo-format generative models. Through a series of controlled production experiments, this article discusses phenomena such as temporal inconsistency, which refers to the lack of visual smoothness or objects changing shape between frames; directability challenges, meaning the difficulty in getting the AI to follow specific creative instructions accurately; and operational latency, which is the time delay or slow processing speed encountered during video generation. Understanding these terms is crucial, as they provide a guide for educators and media practitioners in managing expectations for future AI workflows.

LITERATURE REVIEW

This section traces the development of generative models, focusing on the shift from static images to video synthesis, technical challenges in temporal consistency, and cultural bias in algorithms. In addition, this review explicitly sets the work in the context of institutional video content creation within Malaysian higher education (e.g., convocation and programme promotions). This localisation matters because most evidence and benchmarks are Western-centric, while educational policy guidance emphasises cultural/linguistic suitability and human-centred practices in local implementations.

Evolution: From Image to Generative Video

Modern synthetic media began with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). While revolutionary, GANs often faced training instability. A shift occurred with Denoising Diffusion Probabilistic Models (DDPM) introduced by Ho et al. (2020), offering higher stability. This foundation is now adapted into text-to-video models such as Veo, developed by Google DeepMind. Veo is designed to generate high-quality 1080p cinematic video and represents Google's most capable generative video model to date, focusing on advanced consistency and creative control. Singer et al. (2023) emphasize that the challenge in the video domain is far more complex as models must understand real-world physics priors, such as gravity and lighting, across a timeline, not just static pixels.

Subsequent large-scale advances demonstrate how image-diffusion foundations can be extended to video. Meta's Make-A-Video leverages text-image pairs to learn appearance and unlabeled videos to learn motion, enabling text-to-video without paired text-video data (Singer et al., 2023). Google's Lumiere introduces a space-time U-Net that generates an entire video in a single pass, improving global temporal consistency compared with cascaded keyframe + TSR pipelines (Bar-Tal et al., 2024). Recent surveys (Ma et al., 2025; Wang et al., 2025) confirm that video diffusion models require significantly higher computational resources and advanced temporal modeling compared to image generation, making them less accessible for resource-constrained institutions.

Temporal Consistency and the Semantic Gap

The primary hurdle in professional AI video production is maintaining temporal consistency. Wu et al. (2023) found that diffusion models often fail to maintain subject identity when frames move, causing "flickering" or illogical morphing. This is worsened by the "Semantic Gap," the difficulty in translating abstract user intent into precise text prompts. Liu and Chilton (2022) argue that users' failure to provide sufficiently detailed descriptions often causes AI models to "hallucinate" or ignore parts of the instruction, justifying the need for iterative processes.

Evaluation practices have evolved beyond per-frame metrics. The Fréchet Video Distance (FVD) captures perceptual realism and temporal coherence, and correlates with human judgements in large-scale studies (Unterthiner et al., 2019). To diagnose model weaknesses more meticulously, VBench proposes 16 disentangled dimensions (e.g., subject consistency, motion smoothness, temporal flicker, spatial relations) and couples automated metrics with human preference annotations, better aligning with human perception (Huang et al., 2024).

Multimodal pipelines compound the semantic gap: joint audio-video generation can improve semantic fidelity yet raises alignment and compute demands; MM-Diffusion demonstrates coupled denoising across modalities with superior FVD/FAD scores but increased complexity (Ruan et al., 2023). On the systems side, acceleration efforts (e.g., reducing diffusion steps, frame skipping/interpolation) achieve ~10× speed-ups but entail quality–efficiency trade-offs that producers must balance for professional outputs (Yu et al., 2024).

This challenge is particularly relevant in educational contexts where institutional branding requires strict adherence to visual identity standards. Studies by Ruan et al. (2023) highlight that multimodal integration (text, image, audio) further complicates prompt engineering, reinforcing the need for human oversight.

Data Bias and Cultural Representation

In the context of local institutional content, training data bias is highly relevant. Qadri et al. (2023) note that generative AI models trained on large-scale datasets often exhibit Western-centric bias or broad regional generalizations. This results in inaccuracies when dpotraying local cultural nuances, such as traditional attire or specific etiquette. These failures demand aggressive Human-in-the-Loop (HITL) intervention (Wang et al., 2023) to ensure visual representation remains authentic to the local context. Lu et al. (2024) further argue that cultural misalignment in AI-generated content can preserve stereotypes and destroy trust in educational media, making cultural sensitivity a critical design consideration.

Empirical evidence shows default cultural tendencies in generative models: outputs frequently align with English-speaking/Protestant European values; cultural prompting improves alignment across 71–81% of countries tested (Tao et al., 2024). For Malaysia, the MyCulture benchmark (Bahasa Melayu) evaluates cultural understanding (food, attire, customs, religion, etc.) and documents model disparities, underscoring the need for locally grounded datasets and evaluations (Hew Jia Xin et al., 2025).

Meanwhile, audio layers add complexity where SeamlessM4T and Massively Multilingual Speech (MMS) extend ASR/TTS to ~100 – 1100+ languages, but practical fidelity for dialectal phonemes remains uneven, an element critical for authenticity in institutional videos (Barrault et al., 2023). Thus, without human oversight, generative video systems risk cultural misrepresentation and inauthentic speech – directly impacting audience trust for higher-education communications. (Tao et al., 2024; Hew Jia Xin et al., 2025).

Synthetic Media: Educational Implications and Ethics

The integration of AI in education and institutional promotion carries inherent risks. While Baidoo-Anu and Owusu Ansah (2023) highlight the potential of AI as an effective visual aid, Kietzmann et al. (2020) raise critical concerns regarding authenticity. The ease of creating manipulated videos (deepfakes) presents significant ethical dilemmas, particularly when involving institutional reputation. Consequently, Wang et al. (2023) advocate for a Human-in-the-Loop (HITL) approach, where AI functions as a creative copilot that is strictly monitored and edited by human oversight, ensuring that the final output observes to both quality standards and academic ethics. Operationally, a HITL production pipeline for institutional videos should include: culturally informed prompt development; technical checks using FVD/VBench; post-production fixes for artefacts/physics; and ethical/policy review before publication. (Unterthiner et al., 2019; Huang et al., 2024)

Recent literature also emphasizes policy development for AI adoption in education (MIT Sloan, 2025), suggesting that institutions must establish clear guidelines to balance innovation with ethical responsibility. Human-Centred AI (HCAI) frameworks call for high levels of automation and high levels of human control to achieve reliable, safe, and trustworthy systems in education (Shneiderman,

2020). In practice, the Guidelines for Human-AI Interaction recommend making system capabilities/limits explicit, notifying users of state changes, and enabling correction—principles directly applicable to AI-assisted video workflows (Amershi et al., 2019).

Conclusion

In summary, the literature charts a rapid transition from GANs to diffusion-based video models (e.g., Make-A-Video, Lumiere, Sora), but persistent challenges include temporal coherence, prompt directability, and computational cost—now assessed via video-specific metrics and benchmarks (FVD, VBench). Cultural and linguistic biases remain salient, motivating cultural prompting, local benchmarks (MyCulture), and robust multilingual TTS/ASR. Educational policy and HCAI/HAI design principles converge on a Human-in-the-Loop approach for higher-education institutions, ensuring professional quality and cultural integrity. This study addresses the gap by focusing on Malaysian higher education as the study area and operationalising a HITL workflow for AI-assisted institutional video production.

METHODOLOGY

This study adapts a Practice-based Research approach, where the creation of a creative artifact serves as the primary instrument of investigation (Candy, 2006). The research also applies the Reflective Practitioner framework by Schön (1983), where the researcher engages in "reflection-in-action" to critically evaluate technical decisions during the production process.

Practice-based research is particularly suited for this study because the primary objective is to explore the creative and technical affordances of generative video models within real-world production workflows. Rather than relying solely on theoretical analysis, this approach positions the researcher as both creator and evaluator, enabling iterative experimentation and contextual interpretation of outcomes.

The study area is Malaysian higher education, focusing on institutional video content such as convocation promotions and academic programme advertisements. This context was chosen because universities increasingly seek innovative tools for media production, yet face cultural, linguistic, and ethical constraints that differ from Western-centric benchmarks.

A corpus of 15 experimental video projects was produced using an AI-assisted workflow: 10 convocation-themed videos and 5 programme promotion videos. These videos were produced by the Promotion and Publicity Committee for the 22nd UnIPSAS Convocation Ceremony 2025 at Universiti Islam Pahang Sultan Ahmad Shah, Kuantan, Pahang, Malaysia. The development of the corpus followed established professional video production conventions, starting with scripting to define the

narrative, followed by storyboarding to visualize the AI-generated scenes, and applying instructional design principles to ensure the promotional content was pedagogically sound and effective for its target audience.

Each video project followed a structured pipeline comprising three stages: (i) prompt engineering and text-to-video generation using Google Veo; (ii) integration of audio elements via text-to-speech engines; and (iii) post-production refinement using industry-standard editing tools. The methodological choice reflects the study's dual aim: to generate actionable insights for practitioners and to contribute empirical evidence to the academic discourse on AI-assisted video production in non-Western educational contexts.

Video Production Processes

The technical experiments utilized an AI-assisted workflow:

1. Text Generation : ChatGPT was used as an intermediary agent to refine visual descriptions. This step ensured that prompts were semantically rich and contextually aligned with Malaysian cultural elements. The iterative refinement process involved generating multiple prompt variations, incorporating institutional branding signs (e.g., university colors, attire, and ceremonial objects) and linguistic adjustments for clarity. This approach mitigates the "semantic gap" problem identified in prior studies (Liu & Chilton, 2022), where vague prompts often lead to hallucinated outputs.

2. Video Generation : The Google Veo model (accessed via the Google Flow/VideoFX interface) was used to synthesize video clips.
    a. Output Settings : Average duration of 16 seconds per clip.
    b. Aspect Ratio : 9:16 (Vertical) for mobile platforms and 16:9 (Horizontal) for cinematic displays.

    The choice of aspect ratios reflects dual distribution channels, which is vertical formats for social media engagement and horizontal formats for institutional archives and event screenings. Each clip was generated using high-resolution settings to preserve visual fidelity, with motion coherence and lighting consistency monitored during generation.

Sources of Data

A total of 15 experimental videos were produced using Purposive Sampling (Creswell, 2014). The sample was divided into two clusters:

1. Cluster A : Convocation Ceremony Promotion (n=10):
    a. Status : 5 videos were published on official university social media; 5 were rejected.
    b. Justification : Rejected videos are analyzed as "Negative Cases" (Yin, 2018) to identify technical limits and operational failures (e.g., missed deadlines due to generation latency).

    This cluster was selected because convocation ceremonies represent high-stakes institutional events requiring strict adherence to cultural norms, visual identity, and ceremonial protocols. The inclusion of rejected outputs provides critical insight into failure modes such as prompt misalignment, artifact generation, and latency bottlenecks, which are often underreported in generative AI literature. Negative case analysis strengthens methodological difficulty by exposing conditions where AI workflows break down, offering lessons for risk mitigation in future deployments.

2. Cluster B : Academic Program Promotion (n=5):
    a. Status : Published on the researcher's personal TikTok platform to test public audience acceptance in an uncontrolled environment.

    This cluster was chosen to explore generative video performance in informal, high-engagement channels where aesthetic appeal and cultural resonance drive audience interaction. TikTok was selected as the testbed due to its algorithmic amplification of short-form video content and its popularity among prospective students. Publishing in an uncontrolled environment allows observation of organic engagement metrics (views, likes, comments) and qualitative feedback, providing a proxy for audience perception of AI-generated institutional media.

Sampling Rationale

Purposive sampling was employed to ensure representation of two distinct institutional communication contexts, formal ceremonial promotion and informal program marketing. This design enables comparative analysis of technical and cultural challenges across different production goals. The sample size (n = 15) was determined based on resource constraints and the exploratory nature of practice-based research, prioritizing depth of analysis over statistical generalization. Each video was treated as a discrete case, with metadata (prompt structure, generation parameters, evaluation scores) systematically recorded for cross-case synthesis.

Data Analysis Procedure

Videos were analyzed using a Real-world Contextual Analysis approach based on the concept of Satisficing (Simon, 1996), accepting a good enough solution rather than perfection given time constraints. To provide a clearer description of the analysis procedures, the data underwent a structured

thematic process. Initially, the researcher documented specific individual defects observed in the video corpus. These isolated patterns (such as floating objects, foreign architectural styles, or mispronounced phonemes) were subsequently synthesized and categorized to form three overarching themes: Cultural Bias, Linguistic Limitations, and Physical Hallucinations. Analysis focused specifically on:

1. Visual Artifact Analysis – Detecting hallucinations and physical inconsistencies (Amershi et al., 2019).
2. Operational Efficiency – Evaluating the "friction" and turnaround time required to rectify errors versus the strict promotional deadlines (Inie et al., 2023).

FINDINGS

This section presents the analysis of the 15 video projects, categorized by the technical and operational failures identified.

Cultural and Regional Bias (The "Nusantara" Effect)

A finding was the model's inability to distinguish specific "Malaysian" cultural nuances from broader "Indonesian" demographics.

- Observation – In Video 11 (IT Diploma Promotion), despite the location being specified as the local campus mosque (Masjid TMTHIAS), the generated characters wore traditional Baju Melayu with distinct Indonesian-style sarong tying methods and fabric textures, differing from the local Malaysian standard.

- Observation – Video 1 (Convocation) was rejected because the campus architecture and student attire strongly resembled Indonesian universities, failing to capture the unique corporate identity of UnIPSAS.

Linguistic and Phonetic Limitations

The integrated Text-to-Speech (TTS) models showed significant weaknesses in handling local dialects and loanwords.

- Vowel Sounds – In Video 3 and Video 4, the model failed to distinguish between the Malay Schwa sound (e-pepet) and the Close-mid front unrounded vowel (e-taling). The word "Ye" was pronounced phonetically incorrectly for the context.

- Loanword Pronunciation – In Video 14 and Video 15 (Multimedia Degree Promotion), the model

read the word "Multimedia" literally according to Malay standard spelling, rather than the expected English pronunciation (/mʌltiˈmiːdiə/), requiring phonetic respelling in the prompt.

- Intonation Drift – Video 6 and Video 7 were rejected because the dialogue intonation drifted into an Indonesian accent midway through the clip.

Physical and Logical Hallucinations

- Physics Failure – In Video 1, graduates were depicted throwing mortarboards, but the hats "levitated" upwards without being held or thrown by hands.

- Object Permanence – In Video 7, a character was generated holding a flagpole, but the flag itself (requested as the Palestinian flag) was missing.

- Unwanted Artifacts – Video 9 was rejected because the AI hallucinated an unauthorized logo next to the official university signage, which means the system autonomously generated a distorted or incorrect version of the university's official emblem. This output was deemed unusable and rejected because publishing a video with an inaccurate institutional logo would not only violate branding guidelines but also confuse the audience and damage the credibility of the university's official communication.

Operational Challenges – Latency vs. Quality

The study observed a clear link between technical defects and operational failure. The 5 rejected videos in Cluster A were not only visually flawed but were deemed operationally failed because the time required to regenerate and fix these errors caused the production to miss the event's promotional window. Conversely, the published videos (n=10) contained minor defects, such as slight attire inaccuracies, but were released under the principle of satisficing.

This approach means choosing a 'good enough' outcome rather than chasing a difficult-to-reach state of perfection. The production team decided to prioritize meeting deadlines over 'pixel perfection' to ensure the content reached the audience during the important promotional window. Instead of wasting time and resources on endless AI regenerations that showed little improvement, the team accepted results that met professional standards while remaining practical for real-world use.

Therefore, all the findings can be summarized into following table:

Table 1 : Summary of AI-Generated Video Projects – Status, Observations, and Error

Categories

Video ID	Project Description	Status	Primary Defect / Observation	Error Category
V1	Convocation General Promotion	**Failed**	Mortarboards levitated without hands; campus architecture resembled Indonesian universities.	Physics Hallucination & Cultural Bias
V3, V4	Robe Collection & Carnival Promo	**Published**	The word "Ye" was mispronounced with the wrong phoneme (*e-taling* instead of *e-pepet*).	Linguistic Inaccuracy
V7	Nasyid Concert Teaser	**Failed**	Character holding a flagpole without the Palestinian flag; dialogue drifted into Indonesian accent.	Object Logic Failure & Cultural Bias
V9	Congratulatory Message	**Failed**	AI hallucinated an unauthorized logo next to the official university signage.	Visual Hallucination
V11	IT Diploma Promotion	**Published**	Character generated wearing *Baju Melayu* with Indonesian-style *sarong* tying method.	Cultural Bias (Regional Data Dominance)
V14, V15	Multimedia Degree Promotion	**Published**	The word "Multimedia" was pronounced literally according to Malay spelling rules, not the English academic term.	Linguistic Inaccuracy

RECOMMENDATIONS

The study reveals that while models like Google Veo are powerful, they are not autonomous creators. In this context, autonomy refers to the model's limited ability to make independent creative and contextual decisions without human guidance. The findings from the 15 video projects demonstrate that the AI lacks contextual autonomy, as it frequently hallucinates foreign cultural elements (e.g.,

Indonesian-style attire) when specific local nuances are not provided.

Furthermore, the model lacks decision-making autonomy in professional workflows; it cannot self-correct physical glitches or linguistic errors without iterative human intervention (prompt engineering). Therefore, this conclusion is derived from the observation that the AI functions as a raw material generator that still requires a Human-in-the-Loop to ensure cultural accuracy and professional standards. To mitigate the identified failures, the following strategies are proposed for AI practitioners.

Descriptive vs. Declarative Prompting

To overcome cultural bias, directors must shift from Declarative Prompting (e.g., "Malaysian vibes") to Descriptive Prompting. Instead of relying on the AI's generalized cultural knowledge, prompts should explicitly describe visual components (e.g., "Male student wearing a stiff collar shirt with a songket fabric wrapped around the waist").

This recommendation aligns with Liu and Chilton (2022), who highlight the "semantic gap" as a major source of generative errors, where vague or abstract prompts lead to hallucinated outputs. By specifying attire, setting, and ceremonial objects, descriptive prompting reduces ambiguity and improves alignment with institutional branding. This approach also resonates with Tao et al. (2024), who argue that cultural prompting techniques significantly enhance cultural representation in generative outputs.

Phonetic Manipulation

To address linguistic limitations, such as the AI's tendency to mispronounce local Malaysian terms or use an unnatural robotic accent during video generation, prompt engineers should utilize Phonetic Spelling. This strategy involves spelling words based on how they sound rather than their correct grammar. For example, to force the AI's Text-to-Speech (TTS) engine to produce the correct local dialect for the UnIPSAS promotional videos, words were intentionally misspelled, such as writing 'Multimedia' as 'Maltimedia' or 'Ye' as 'Yuh'. This manual adjustment ensures the audio matches the natural flow of local speech, overcoming the model's default setting which often struggles with regional linguistic nuances.

The Hybrid Workflow

AI should be viewed as a "Raw Material Generator" rather than a final product creator. The most efficient workflow involves generating base clips via AI and rectifying artifacts (such as lip-sync issues or unwanted logos) using external post-production software, rather than relying on endless regeneration cycles which lead to operational delays.

This approach is consistent with findings in recent studies on AI-assisted media production, which emphasize that generative AI is most effective when used for rapid prototyping rather than as a standalone solution (Smith & Anderson, 2023; Zhao et al., 2024). Literature on hybrid workflows highlights that human intervention remains critical for ensuring quality, creativity, and contextual accuracy (Brown et al., 2022). Continuous regeneration within AI systems often leads to diminishing returns and increased operational costs, as noted by Lee and Kim (2023), making external post-production tools a more efficient strategy.

The implications of this study are significant for educational institutions and media organizations. By adopting a hybrid approach, institutions can reduce production costs while maintaining creative control, but they must also invest in training personnel to develop prompt engineering skills and ethical oversight mechanisms.

CONCLUSION

In conclusion, producing institutional videos using Generative AI is not a 'one-click' solution, but rather a complex collaboration between human and machine. Our study shows that the primary challenge is no longer about the sharpness or resolution of the image, but about getting the small details right. In simple terms, this means that while AI can easily create a beautiful picture, it still struggles to follow specific, subtle instructions, such as getting a local dialect exactly right or keeping a university logo consistent across different scenes.

Consequently, the role of the human director has evolved into that of a curator. This shift requires the director to act as a bridge, manually fixing the AI’s mistakes, bypassing its built-in biases, and deciding when a video is good enough to meet a deadline. The insights from this research highlight that while AI speeds up the initial creation process, the final quality still depends on human judgment. The most important lesson learned is that for AI to be useful in professional workflows, practitioners must move away from seeking perfection and instead focus on mastering prompt manipulation to manage the balance between technical quality and the urgent need for timely delivery.

LIMITATIONS OF THE STUDY

However, this study has certain limitations. It primarily focuses on the creative and operational aspects of Generative AI in video production without conducting empirical measurements of efficiency or audience perception. Future research should incorporate quantitative evaluations, such as cost-benefit analysis and user engagement metrics, to validate the effectiveness of hybrid workflows.

For policymakers, the findings highlight the need to establish clear guidelines on intellectual property, data privacy, and transparency in AI-generated content. Regulatory frameworks should ensure that AI-

assisted productions adhere to institutional branding standards and avoid algorithmic biases that could misrepresent cultural or organizational values.

Finally, future studies should explore the scalability of hybrid workflows across different content types, such as interactive learning modules or multilingual video campaigns. Comparative research between fully automated and hybrid models could provide deeper insights into efficiency, quality, and ethical considerations in AI-driven media production.

CONFLICT OF INTEREST

The author declares no conflicts of interest.

ACKNOWLEDGEMENTS

Thanks to the Faculty of Management and Informatics (FMI) and the Committee Members of Promotion and Public Relations for the 22nd UnIPSAS Convocation Ceremony for their support of this study**.**

AUTHOR CONTRIBUTIONS STATEMENT

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The published video samples analyzed during the current study are publicly available on the UnIPSAS official social media channels and the author's public TikTok account. The unpublished/rejected video samples and specific prompt logs are available from the corresponding author on reasonable request.

ETHICS STATEMENT

Not applicable.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1-13.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52-62.

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., Li, Y., Rubinstein, M., Michaeli, T., Wang, O., Sun, D., Dekel, T., & Mosseri, I. (2024). Lumiere: A space-time diffusion model for video generation (arXiv:2401.12945). arXiv.

Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., & Wang, S. (2023). SeamlessM4T: Massively multilingual & multimodal machine translation (arXiv:2308.11596). arXiv.

Candy, L. (2006). Practice based research: A guide. CCS Report, 1, 1-19. University of Technology, Sydney.

Creswell, J. W. (2014). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (4th ed.). Sage Publications.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Hew Jia Xin, Z. K., Low, J. X., Yang, S. J., & Chan, C. S. (2025). MyCulture: Exploring Malaysia's diverse culture under low-resource language constraints (arXiv:2508.05429). arXiv.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., & Liu, Z. (2024). VBench: Comprehensive benchmark suite for video generative models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Inie, N., & Dalsgaard, P. (2023). Designing with AI: Design practitioners' perspectives on generative AI tools. Proceedings of the 2023 ACM Designing Interactive Systems Conference.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135-146.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. Proceedings of the CHI Conference on Human Factors in Computing Systems

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large

language models. PNAS Nexus, 3(9), page 346.

Ma, W., Wang, Y., & Zhang, X. (2025). Video diffusion generation: Comprehensive review and open problems. Artificial Intelligence Review.

Qadri, R., et al. (2023). AI's regimes of representation: A community-centered study of text-to-image models in South Asia. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Ruan, L., Zhang, Y., & Wang, J. (2023). MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12345–12354.

Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Basic Books.

Shneiderman, B. (2020). Human-centred artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction. Volume 36, 2020 - Issue 6

Simon, H. A. (1996). The Sciences of the Artificial (3rd ed.). MIT Press.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., ... & Taigman, Y. (2023). Make-A-Video: Text-to-Video Generation without Text-Video Data. International Conference on Learning Representations (ICLR).

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2019). Towards accurate generative models of video: A new metric & challenges. (arXiv:1812.01717).

Wang, J., et al. (2023). Empowering novice users in creative tasks with generative AI: A human-in-the-loop perspective. International Journal of Human-Computer Interaction.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., ... & Shou, M. Z. (2023). Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Yin, R. K. (2018). Case Study Research and Applications: Design and Methods (6th ed.). Sage Publications.

Yu, P., Luo, D., Rupprecht, T., Lu, L., Kong, Z., Zhao, P., Li, Y., Camps, O., Lin, X., & Wang, Y. (2024). FasterVD: On acceleration of video diffusion models. *Proceedings of IJCAI 2024*