Lights, Camera, Consistency: A Multistage Pipeline for Character-Stable AI Video Stories

Anonymous ACL submission

Abstract

Generating long, cohesive video stories with consistent characters is a significant challenge for current text-to-video AI. We introduce a method that approaches video generation in a filmmaker-like manner. Instead of creating a video in one step, our proposed pipeline first uses a large language model to generate a detailed production script. This script guides a text-to-image model in creating consistent visuals for each character, which then serve as anchors for a video generation model to synthesize each scene individually. Our baseline comparisons validate the necessity of this multistage decomposition; specifically, we observe that removing the visual anchoring mechanism results in a catastrophic drop in character consistency scores (from 7.99 to 0.55), confirming that visual priors are essential for identity preservation. Furthermore, we analyze cultural disparities in current models, revealing distinct biases in subject consistency and dynamic degree between Indian vs Western-themed generations.

1 Introduction

The domain of text-to-video generation represents a frontier at the intersection of natural language processing (NLP) and computer vision, where realistic and coherent video sequences are synthesized directly from textual descriptions. The domain has rapidly advanced with the advent of diffusion models (Ho et al., 2020). These models excel at generating short, high-quality video clips from textual prompts, demonstrating exceptional capabilities in generating dynamic scenes and motions (Singer et al., 2022; Ho et al., 2022).

Recent research in text-to-video (T2V) generation has primarily focused on improving visual quality and temporal coherence within short, continuous clips, leading to powerful models such as Google's Veo (DeepMind, 2025) and OpenAI's Sora (OpenAI, 2025). Despite these advances, the prevailing T2V paradigm remains largely single-shot, where a single prompt produces one uninterrupted clip. While effective for short sequences, this approach is inherently limited in its ability to generate structured narratives involving multiple scenes, recurring characters, and coherent plot progression.

Existing literature largely reflects this clip-centric focus, emphasizing foundational model training (Ho et al., 2022), efficient fine-tuning methods (Wu et al., 2023), and benchmark development for evaluating per-clip fidelity and text–video alignment (Chen et al., 2023). As a result, a significant gap remains in the development of holistic frameworks capable of producing long-form, multi-scene stories with consistent characters and narrative structure. To address this limitation, we propose a system that automates the entire filmmaking pipeline—from script generation to final video composition—enabling the creation of coherent, story-driven videos from a single prompt.

To bridge this gap, we introduce a novel inference-only pipeline for generating narrative-driven videos from a single text prompt. Unlike frameworks relying on probabilistic agent simulations, our approach employs structured deterministic control via JSON blueprints to enforce precise adherence to a script. This methodology mimics a traditional filmmaking workflow but is optimized for computational efficiency. First, we utilize a Large Language Model (LLM) to generate a script with strict scene constraints. Second, we introduce a zero-shot "Asset-First" mechanism that decouples character design from scene generation; this ensures visual consistency through direct conditioning, eliminating the need for expensive fine-tuning or LoRA training. Third, a scene synthesis module utilizes a "Temporal Bridge" to explicitly model the flow of time, using the final frame of a previous scene to visually condition the next. Finally, a merging module produces the final output with

synchronized audio.

Our evaluation highlights three critical insights regarding this architecture. First, we demonstrate that the generated image-to-image (I2I) seed frame is the architectural linchpin for consistency; removing this visual anchor results in a catastrophic drop in character consistency scores (from 7.99 to 0.55), proving that a character model alone is insufficient for identity retention. Second, we identify a distinct performance gap between visual and narrative fidelity: while the model excels at Prompt Adherence (Score: 4.47), it struggles with strict Script Adherence (Score: 3.49), indicating that current architectures prioritize style transfer over complex narrative sequencing. Finally, our bias analysis uncovers a "Subject-World Decoupling" phenomenon. We find that while the model maintains non-Western (e.g., Indian) character identities effectively, the physical coherence of the surrounding environment degrades significantly under high-motion stress compared to Western contexts, highlighting a specific robustness failure in diverse physical simulations.

2 Related Work

2.1 Multi-Agent Frameworks

Current multi-agent frameworks often decompose video generation into sub-tasks mimicking film production. However, our methodology distinguishes itself from these approaches by prioritizing structured, deterministic inference over probabilistic simulation or computationally intensive fine-tuning.

Unlike DreamFactory (Xie et al., 2024), which relies on probabilistic social negotiation between roleplaying agents (e.g., Director vs. Screenwriter) to derive scene parameters, our approach utilizes deterministic JSON blueprints. This shifts the paradigm from simulating a conversation to engineering a control flow. Furthermore, while DreamFactory relies on text-based keyframe iteration, we decouple character design from scene generation. By employing an asset-first strategy—generating explicit character sheets before scene synthesis—we ensure visual consistency through direct visual conditioning rather than semantic descriptions alone.

In contrast to StoryAgent (Hu et al., 2024), which necessitates computationally expensive fine-tuning (e.g., LoRA-BE) to embed character identity into model weights, our framework offers a scalable, inference-only solution. We achieve consistency through a zero-shot "temporal bridge," where the final frame of a preceding scene visually conditions the initial frame of the subsequent shot. This explicitly models the linear flow of time without the overhead of retraining models for every new narrative.

Finally, compared to AniMaker (Shi et al., 2025), which employs computationally heavy Monte Carlo Tree Search (MCTS) for quality control, our pipeline emphasizes efficiency and low latency. By leveraging a linear, structured workflow grounded by LLM-generated anchor frames, we avoid the latency of iterative search algorithms while maintaining multimodal completeness, including the integration of audio and scoring absent in many visual-only pipelines like Mora (Yuan et al., 2024).

2.2 Video Generation

Driven by the success of diffusion models (Ho et al., 2020), video generation has evolved from simple text-to-video (T2V) tasks to complex image-to-video (I2V) synthesis. Pioneering zero-shot frameworks like Text2Video-Zero (Khachatryan et al., 2023) demonstrated that pretrained text-to-image models could generate video without additional training. Subsequent works have introduced temporal self-attention and LLM-guided frame prompts to improve coherence (Wu et al., 2022; Qi et al., 2023). Notably, DirecT2V (Hong et al., 2024) leverages LLMs to generate time-varying prompts for better controllability. However, these methods primarily operate on a per-clip basis, often struggling to maintain narrative and character consistency when extending beyond a single short sequence.

In the domain of Image-to-Video (I2V), architectures like Dynamic-I2V (Liu et al., 2025) represent the state-of-the-art, utilizing multimodal large language models (MLLMs) and conditional adapters to synthesize video from single images. While these models excel at animating static inputs with high motion fidelity, they are typically designed for single-shot animation rather than sequential storytelling. Our work repurposes these I2V capabilities, utilizing them not just for motion generation, but as a "temporal bridge" mechanism where the final frame of one scene serves as the visual anchor for the next.

Recent benchmarks such as VBench (Huang et al., 2024) and DEVIL (Liao et al., 2024) have shifted evaluation focus toward balancing motion

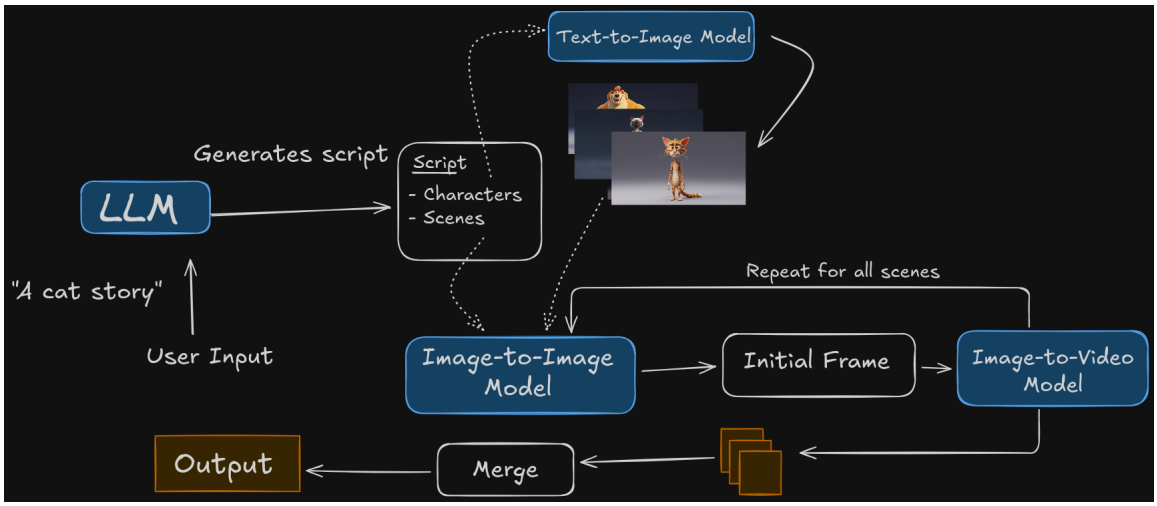

Figure 1: The Proposed Pipeline. A high-level user prompt initiates an LLM to generate a detailed script and character descriptions. A Text-to-Image model creates visual assets for characters. For each scene, an Image-to-Image model generates an initial frame, which then guides an Image-to-Video model to synthesize the scene clip. All clips are then merged into the final output video.

realism with frame-wise quality. Despite these advances, a gap remains in holistic frameworks capable of producing long-form narratives. Unlike existing approaches that rely on probabilistic generation or complex fine-tuning, our pipeline introduces a deterministic, blueprint-driven workflow to enforce character and narrative consistency across multiple generated clips.

3 Methodology

To address the challenge of generating long-form structured videos, we propose a multistage pipeline that breaks down the generation task into a sequence of sub-problems. Our approach mimics a traditional filmmaking workflow, proceeding from a concept to scriptwriting, character design, scene rendering, and final composition.

3.1 Stage 1: Script and Blueprint Generation

The pipeline (Figure 1) begins with a user prompt that outlines the desired story (e.g., "A cat story"). The prompt is fed into a scriptwriting module consisting of a LLM such as Gemini 2.5 suite (Comanici et al., 2025). We carefully engineer a prompt that instructs the LLM to act as a screenwriter and director which outputs a structured JSON object. This blueprint contains three main components:

- **Story Analysis & Art Style**: Details defining the core themes, and a detailed `art_style` object specifying the visual aesthetic, color palette, and references (e.g., "Gritty Epic CGI" inspired by "Game of Thrones").
- **Character Sheets**: A list of all characters in the story. Each character is defined by their name, role, and personality, detailed `physical_description`, `visual_notes`, and a complete character arc (e.g., "a ginger tabby cat wearing a tiny trench coat and fedora").
- **Scene Breakdown**: A sequential list of scenes. Each scene is defined by its duration, setting, a description of the `action`, a list of characters present, visual notes for cinematography, sound effects, and the emotional tone. This structured output serves as the blueprint for the next stages, which ensures a clear narrative and well-described features.

3.2 Stage 2: Character Visualisation

For each character identified in the production blueprint, this module uses a Text-to-Image (T2I) Model to generate a set of consistent images. The prompt for the T2I model is meticulously constructed from the character's `physical_description` and `visual_notes` fields within the blueprint, combined with the overall `art_style` directives. To further ensure consistency across various expressions, advanced prompting techniques with the T2I model can be employed to create a character sheet, which will contain the reference image of the character with multiple views.

3.3 Stage 3: Iterative Scene Synthesis

This stage generates each video clip corresponding to a scene in the script. It operates in an iterative loop for all scene outlines in the production blueprint.

- **Initial Frame Generation:** For each scene, an LLM determines whether to incorporate the final frame from the preceding scene. An Image-to-Image model then generates the initial frame, conditioned on the scene's textual description—encompassing setting, action, and visual parameters from the blueprint, the character references established in Stage 2 and the prior frame when applicable. This approach ensures visual coherence between successive clips while maintaining consistent character representation throughout the narrative.

- **Video Clip Generation:** The generated "Initial Frame" then serves as the primary input for an Image-to-Video (I2V) Model (e.g Google Veo). This I2V model, guided by the scene's action description and the established art style, generates a short video clip of about 8 seconds and 24 fps that extends the initial frame into a clip. The use of an initial frame helps maintain visual and temporal coherence in the generated clips.

3.4 Stage 4: Final Composition

The final stage of the pipeline involved composing the individual video clips into a single narrative. An editing module takes the sequence of clips generated in the previous stage and concatenates them in the correct order as defined by the script. Furthermore, the `sound_effects` and `emotional_tone` cues guide the generation or selection of appropriate audio, which is then merged with the visual compilation. The output is the final, complete video, ready for viewing.

4 Evaluation

To assess the effectiveness of our proposed pipeline, we employ an evaluation strategy that covers different features of long-form T2V generation. Our goal is to measure the pipeline's ability to generate coherent videos that strictly follow the user prompts while maintaining consistency between the scenes. The evaluation combines objective automated metrics with subjective assessments from advanced AI models. Additionally, an ablation study is conducted to validate the contribution of each key component of the pipeline.

4.1 Dataset

Dataset: To evaluate our system, we collect story prompts from diverse social media sources to characterize the landscape of user-generated narrative requests. Given our focus on Indian-themed content, we prioritize prompts that incorporate Indian cultural elements, mythology, and narrative traditions. We then employ an LLM to generate variations and extensions of these prompts, expanding the candidate pool. From this augmented set, we select the 39 most representative prompts for evaluation. These prompts cover a diverse range of genres (e.g., sci-fi, fantasy, mystery, etc.), all grounded in Indian contexts and storytelling conventions. For each prompt, we generate a video using our proposed pipeline.

4.2 Baseline Study

To validate the architectural design of our pipeline, we compare our full model against a series of internal baselines. By systematically excluding key components, we evaluate the resulting impact on video quality and character coherence. The primary baseline configurations are:

- **Baseline 1 (No Character Visualization):** This configuration excludes Stage 2 (Character Generation) from the pipeline. Consequently, the Image-to-Image (I2I) model generates the scene using only the textual descriptions from the blueprint and, if available, the last frame of the previous scene, without a dedicated visual reference for the character.

- **Baseline 2 (No Initial Frame):** This configuration excludes Stage 3 (I2I Seed Generation). As a result, the final Image-to-Video (I2V) model (Stage 4) receives only the textual scene descriptions from the blueprint, forcing it to operate without the visual anchor provided by the starting frame.

4.3 Metrics

We evaluate our system using two comprehensive metrics that assess the fidelity of the most critical components: character consistency and script adherence.

4.3.1 Character Consistency

We measure character consistency through a dual-pronged approach that combines both computational and semantic evaluation. Following (Atzmon et al., 2024), we first establish a computational

baseline via cross-scene frame similarity. We uniformly sample f frames from each scene, yielding $nf + 1$ frames in total (where n denotes the number of scenes), combined with the reference character image from Stage 2. To isolate character features from confounding visual factors, we employ CLIPSeg (Lüddecke and Ecker, 2021), a CLIP-based segmentation model, to extract the character region. We further refine this by masking the background following (Fu et al., 2023), ensuring that only the character contributes to similarity computations. We then extract deep visual features using the DINO model (Caron et al., 2021), conditioning the feature extraction on the character's physical description specified in the blueprint. We compute pairwise cosine similarity across all frame pairs, excluding intra-frame comparisons, to accurately measure cross-scene character consistency, and average the results to obtain a quantitative consistency core. This metric is computed independently for each character.

To complement the computational metric, we additionally employ MLLM-as-a-judge evaluation, where an MLLM assesses the visual fidelity of each frame's character relative to the character's image generated in stage 2, on a scale of 0 to 10 (10 = perfect consistency). This semantic evaluation captures nuances in character preservation that may not be fully captured by feature-level similarity alone.

4.3.2 Script and Prompt Adherence

We assess the semantic quality of the generated videos by measuring their alignment with both the user's intent and the system's internal narrative. Following (Zheng et al., 2023) we employ a Multimodal Large Language Model (MLLM) to directly analyze the visual content of the generated video. The MLLM acts as a judge, assigning scores on a Likert scale from 1 to 5 for two distinct dimensions:

- **Script Adherence:** This metric evaluates the fidelity of the generated video to the detailed intermediate script produced by our pipeline. It ensures that specific scene details, character actions, and dialogue specified in the script are accurately visualized.

- **Prompt Adherence:** This metric evaluates the alignment between the final video and the original user prompt. It assesses whether the high-level intent, core concepts, and thematic elements requested by the user are preserved in the final output.

By conducting this dual-factor evaluation, we ensure that the system not only follows its own generated roadmap (the script) but also satisfies the user's original request.

4.3.3 Bias Evaluation

We employ specific metrics to detect potential biases in identity retention and physical stability across the generated content.

Subject Consistency (Identity Stability): To evaluate whether the model preserves the protagonist's identity across frames, we utilize the **DINOv2** architecture (Oquab et al., 2023). Unlike traditional pixel-based metrics (PSNR/SSIM) which are sensitive to pose changes, DINO extracts semantic, high-level features that represent the "essence" of the subject. Following the protocol in (Ruiz et al., 2023), we calculate the average cosine similarity between the feature embeddings of the initial anchor frame (I_0) and all subsequent frames (I_t). A score closer to 1.0 indicates robust identity retention.

$$S_{subject} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\phi(I_t) \cdot \phi(I_0)}{||\phi(I_t)|| \cdot ||\phi(I_0)||} \quad (1)$$

World Consistency (Physical Stability): To assess the stability of the background and inanimate objects—critical for immersive storytelling—we employ **CoTracker** (Karaev et al., 2023), a transformer-based dense point-tracking model. We initialize a grid of tracking points across the video duration and measure their "visibility duration." Generative hallucinations (e.g., objects morphing into the background) cause tracking points to vanish. We aggregate the normalized visibility of all points to derive a World Consistency score (S_{world}), where a lower score indicates a "jittery" or structurally unstable environment.

Dynamic Degree (Motion Intensity): To ensure that lower stability scores are not simply artifacts of higher motion, we calculate the **Dynamic Degree** using the Farneback Optical Flow algorithm (Farnebäck, 2003). This metric quantifies the average magnitude of pixel displacement between frames. This serves as a control variable; by stratifying our analysis based on motion intensity (Low, Medium, High), we can determine if the model is

"brittle" under specific dynamic conditions, independent of the visual domain.

5 Results and Analysis

We present the results of our evaluation, comparing the proposed pipeline to our baselines across automated and LLM-based metrics. The findings demonstrate the significant advantages of our structured, multi-stage approach.

5.1 Baseline Study

	Finn	Felica	Brutus	Chad	Average
Baseline 1 (No Stage 2)	6.67	2.83	6.61	7.00	5.78
Baseline 2 (No Seed)	0.34	0.08	1.05	0.67	0.55
Ours (Full Model)	7.42	9.00	6.73	10.00	7.99

Table 1: Character consistency comparison between our proposed method and internal baselines.

We evaluated our proposed full pipeline against two baseline configurations to isolate the contributions of specific modules. The results are summarized in Table 1. All models used the identical blueprint (from Stage 1) to ensure a fair comparison.
Our analysis of the results is as follows:

- **Proposed Method (Full Pipeline):** Our complete approach achieved the highest character consistency score (7.99). This demonstrates the effectiveness of the multi-stage architecture, where a character image (Stage 2) provides a strong visual prior for the I2I seed frame, which in turn acts as a robust anchor for the final I2V animation.

- **Baseline 1 (No Character Image):** This baseline operates without the dedicated character image (Stage 2). Consequently, the I2I model generates both the scene and the character's appearance relying solely on textual descriptions from the blueprint. This resulted in a considerable performance gap compared to our full model (7.99 $\rightarrow$ 5.78). While the I2I seed frame still provided a stable composition, the character's appearance was re-imagined from text for each scene, leading to inconsistencies.

- **Baseline 2 (No I2I Seed Frame):** This baseline excludes the I2I seed frame generation entirely. The final I2V model is forced to operate in a standard "Text-to-Video" mode, receiving only the textual blueprint. This led to a catastrophic drop in consistency (7.99 $\rightarrow$ 0.55). This finding is critical: it indicates that the I2I seed frame is the most vital component for locking in character identity. Without this strong visual anchor, the I2V model struggles to maintain consistency, even when provided with a character image.

In conclusion, the comparison against these baselines validates our architectural design. The results confirm that both the dedicated character model (Stage 2) and the I2I seed frame are essential for high-fidelity character consistency.

5.2 Character Consistency Analysis

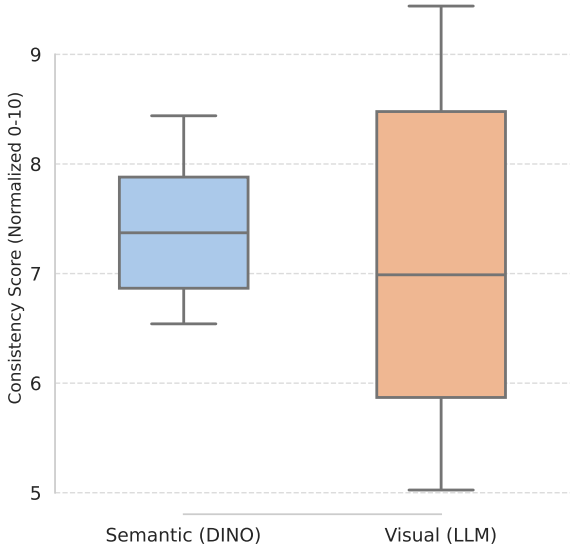

Figure 2: Distribution of Evaluation Scores.Box plots summarizing the performance across all test samples.

Visual Consistency: Our model demonstrates high visual stability,score of **7.14** ($\sigma = 1.43$). The interquartile range observed in Figure 2 (red box) indicates that the model was not able to consistently maintain character features—such as clothing, facial structure, and color palette—across diverse scenes and camera angles with minimal variance.

Semantic Consistency: In terms of narrative fidelity, the model achieved a mean DINO score of **7.41** ($\sigma = 0.59$). The compact spread in the semantic distribution (blue box) demonstrates strong consistency in maintaining character identity across different scenes with most samples successfully maintain semantic coherence throughout the video.

Overall, the results validate that our pipeline effectively balances the trade-off between rigid visual

stability and the flexibility required for semantic storytelling.

5.3 Script and Prompt Adherence

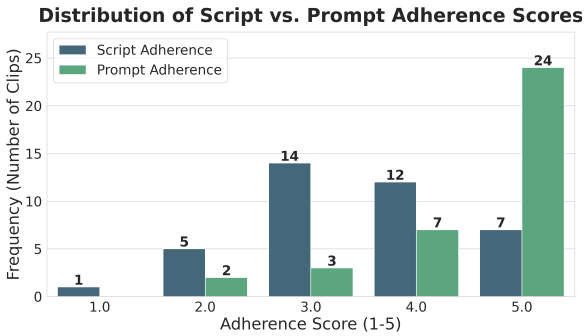

Figure 3: The distribution of scores for Script Adherence versus Prompt Adherence across all generated clips.

Figure 3 illustrates the distribution of scores for both metrics. The model demonstrated exceptional performance in Prompt Adherence, achieving a mean score of 4.47 ($SD = 0.88$). A significant majority of clips (75%) received a perfect score of 5, indicating high reliability in translating textual visual descriptions into video content.

In contrast, Script Adherence showed greater variability, with a lower mean score of 3.49 ($SD = 1.02$). The distribution for script adherence is centered around scores of 3 and 4, suggesting that while the model generally follows the narrative arc, it occasionally misses specific scripted actions or timing details.

Notably, while 31 clips (approximately 86% of the valid dataset) achieved a Prompt Adherence score of 4 or higher, only 14 clips (36%) achieved the same threshold for Script Adherence. This disparity suggests that the current architecture prioritizes visual style transfer and prompt fidelity over complex narrative sequencing.

5.4 Bias Analysis

To assess the potential demographic and contextual biases in our pipeline, we conducted a comparative study between two distinct datasets: **Western/Fantasy (Control)** and **Indian/Realistic (Test)**. We evaluated performance across two dimensions: Identity Stability and Environmental Persistence.

A critical finding of our study is the divergence between subject and world stability. As visualized in Figure 4, the model exhibits a **Subject-World Decoupling** in non-Western contexts.

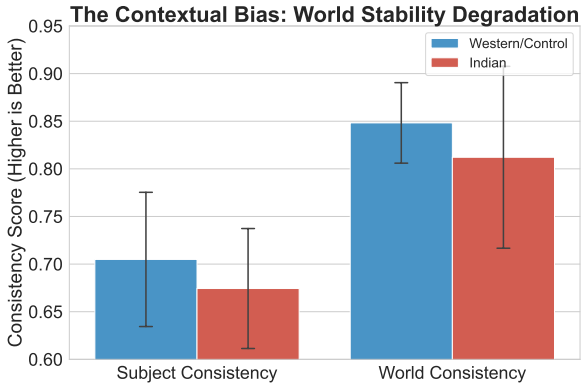

Figure 4: While Subject Consistency remains comparable across demographics, World Consistency drops significantly for Indian videos, indicating contextual fragility.

The **Subject Consistency gap** is considerable ($\Delta - 0.03$), indicating that the generative model maintains Indian identities (faces, skin tone, clothing) with relatively less fidelity as Western identities. However, the **World Consistency gap** is significant ($\Delta - 0.04$). This suggests that while the model can "anchor" a diverse character, it struggles to maintain the physical coherence of the surrounding environment (backgrounds, objects, crowds) in diverse settings.

5.5 Stratified Motion Stress Test

To address potential concerns regarding unequal sample sizes, we conducted a **Stratified Analysis**. We categorized all videos into three "Difficulty Tiers" based on their Dynamic Degree scores (Low, Medium, High) and compared performance within each tier.

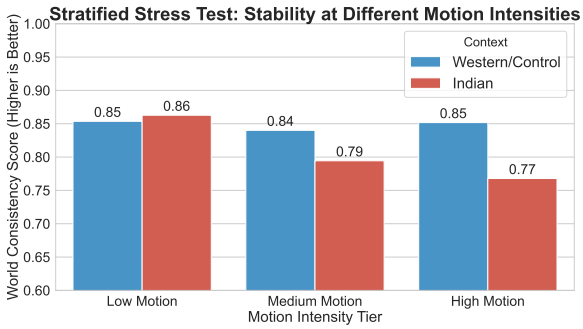

Figure 5: Stratified Analysis. The stability gap is minimal in static scenes (Low Motion) but widens drastically in dynamic scenes (High Motion), confirming that the bias is motion-dependent.

The stratified results (Figure 5) confirm that the instability is not an artifact of dataset size, but a fun-

damental limitation of the model's physics engine in diverse contexts:

- **Low Motion:** The stability gap is minimal. In static scenes, the model renders Indian environments with high fidelity.
- **High Motion:** The gap widens significantly. Under dynamic stress, the Western control group maintains stability, while the Indian test group degrades.

This confirms that the model possesses the *capacity* to render Indian contexts but lacks the *robustness* to maintain them during complex physical interactions.

6 Conclusion

In this work, we presented a novel "filmmaker" pipeline designed to address the challenge of creating long, cohesive video narratives. Rather than relying on end-to-end generation, our approach utilized a large language model to author detailed production scripts, followed by a character-anchoring process that guided the final scene synthesis.

Experimental results demonstrated that this structured decomposition significantly improved narrative coherence and character consistency compared to standard baselines. Our ablation study confirmed that each stage of the pipeline was essential for achieving these results. Additionally, our evaluation highlighted performance disparities in subject consistency and dynamic degree biases between Indian and Western-themed videos, underscoring the need for more culturally diverse training data in future text-to-video systems.

7 Limitations

While our proposed pipeline demonstrates a robust method for generating consistent and coherent short-form videos from a single prompt, we acknowledge several limitations that present opportunities for future research.

The final quality of the generated video is fundamentally capped by the performance of the underlying generative models at each stage. Failures in any stage-the script LLM, character generator, Image-to-Image model, or Image-to-Video model- will inevitably propagate and compound through the pipeline. The system's ability to render complex scenes, specific art styles or characters is therefore constrained by the training data and capabilities of these foundation models.

The blueprint specifies an action (e.g., "a character walks across the room"), but the I2V model determines the final motion path, speed, and character gait. The pipeline lacks a mechanism to explicitly direct precise cinematography (e.g., "dolly zoom," "rack focus") or complex character choreography within the video generation stage.

8 Acknowledgement

The authors acknowledge the use of Large Language Models (LLMs), including Gemini, ChatGPT and Claude, to assist in refining the language, improving grammatical accuracy, and enhancing the overall presentation of this paper. All outputs generated by these tools were carefully reviewed, edited, and verified by the authors. The final content remains an accurate representation of the authors' contributions.

References

Yuval Atzmon, Rinon Gal, Yoad Tewel, Yoni Kasten, and Gal Chechik. 2024. Multi-Shot Character Consistency for Text-to-Video Generation. *Preprint*, arXiv:2412.07750.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Shuaiwei Chen, Yixang Wang, Yixiao Ge, Ying Shan, Wenhai Wang, and Tong Lu. 2023. EvalCrafter: Benchmarking and evaluating large video generation models. *Preprint*, arXiv:2310.11440.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, and 1 others. 2025. Gemini 2.5: Advanced reasoning, multimodality, long context, and agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google DeepMind. 2025. Veo 3.1. `https://deepmind.google/models/veo/`. Text-to-video generative model.

Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis (SCIA)*, pages 363–370. Springer.

Wei Fu, Kai Han, Yan Pang, Yuqi Liu, Yu Wang, Shuai Liu, and Shun-Ren Pu. 2023. Object-Aware Masked Autoencoders for Self-Supervised Learning. In *Advances in Neural Information Processing Systems*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High definition video generation with diffusion models. *Preprint*, arXiv:2210.02303.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. 2024. DirecT2V: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. 2024. StoryAgent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*.

Chang Huang, Zhaoyang Chen, Runzhou Xu, Fan Zhang, Yong Wang, and 1 others. 2024. Vbench: Comprehensive benchmark suite for text-to-video generation evaluation. In *arXiv preprint arXiv:2408.xxxxx*.

Nikita Karaev, Ignacio Rocco, Osama Makansi, Andrea Vedaldi, Natalia Neverova, and Iasonas Kokkinos. 2023. Cotracker: It is better to track together. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page To appear.

Aram Khachatryan, Yinghao Guo, Hao Zhang, Xintao Li, Sai Wang, Zhan Li, Chen Wang, Zhongang Xu, Yan Huang, and Chen Change Loy. 2023. Text2Video-Zero: Text-to-video generation without text-video data. *arXiv preprint arXiv:2301.14897*.

Rui Liao, Ze Gao, Zhen Yu, Rui Zhang, and Yu Liu. 2024. DEVIL: Dynamic evaluation of video with interactive learning. *arXiv preprint arXiv:2402.07864*.

Peng Liu, Xiaoming Ren, Fengkai Liu, Qingsong Xie, Quanlong Zheng, Yanhao Zhang, Haonan Lu, and Yujiu Yang. 2025. Dynamic-I2V: Exploring image-to-video generation models via multimodal large language models. *arXiv preprint arXiv:2505.19901*.

Timo Lüddecke and Alexander S. Ecker. 2021. Image segmentation using text and image prompts. In *arXiv preprint arXiv:2112.10003*.

OpenAI. 2025. Sora 2. https://openai.com/index/sora/. Text-to-video generative model.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Daniel, Babak Adigital, Julien Mairal, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. In *arXiv preprint arXiv:2304.07193*.

Zhuyun Qi, Chenlin Xu, Ying Xu, Li Yuan, and Chen Change Loy. 2023. Text-driven Video Generation via Diffusion Models with Temporal Self-Attention. *arXiv preprint arXiv:2302.04010*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. AniMaker: Automated multi-agent animated storytelling with mcts-driven clip generation. *arXiv preprint arXiv:2506.10540*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-video generation without text-video data. *Preprint*, arXiv:2209.14792.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. *Preprint*, arXiv:2212.11565.

Rui Wu, Zhehao Liu, Nan Song, Lin Wang, and Xiaokang Yang. 2022. Tune-A-Video: One-Shot Tune for Text-to-Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16544–16554.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawendé F. Bissyandé, and Saad Ezzini. 2024. DreamFactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*.

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Preprint*, arXiv:2306.05685.