

HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives

Supplementary Material

001	Appendix	
002	A. Details on Evaluation metrics	
003	We evaluate the models across five key dimensions: aes-	
004	thetic quality, semantic consistency, intra-shot consis-	
005	tency(capturing subject and background stability), inter-	
006	shot consistency, and transition control.	
007	A.1. Transition Control Evaluation Metrics	
008	To comprehensively evaluate the model’s ability to follow	
009	explicit shot-cut instructions, we propose the Shot Cut Ac-	
010	curacy (SCA) metric. This metric holistically assesses shot	
011	control by quantifying both the accuracy of the number of	
012	cuts and the temporal precision of their placement.	
013	To compute the SCA, we first apply a state-of-the-art	
014	shot boundary detector, TransNet V2 [11], to the gener-	
015	ated video to obtain the set of predicted cut locations $P =$	
016	$\{p_j\}_{j=1}^{N_{pred}}$. These are compared against the ground truth cut	
017	locations $G = \{g_i\}_{i=1}^{N_{gt}}$ specified in the user instructions.	
018	SCA is defined as:	
019	$SCA = \exp(-NSD) \quad (S1)$	
020	where NSD is the <i>Normalized Shot Discrepancy</i> , represent-	
021	ing the total error relative to the total video frames, F_{total} :	
022	$NSD = \frac{E_{matched} + E_{penalty}}{F_{total}} \quad (S2)$	
023	Here, E_{total} is composed of two parts:	
024	1. Matched error ($E_{matched}$): This term quantifies the	
025	temporal deviation of successfully detected cuts. We em-	
026	ploy a one-to-one greedy matching strategy to identify the	
027	set of matched pairs $\mathcal{M} = \{(g, p)\}$. The error is the sum of	
028	absolute distances between these pairs:	
029	$E_{matched} = \sum_{(g,p) \in \mathcal{M}} g - p \quad (S3)$	
030	2. Penalty error ($E_{penalty}$): This term penalizes discrep-	
031	ancies in the shot count, including missed cuts (False Neg-	
032	atives, N_{FN}) and extraneous cuts (False Positives, N_{FP}).	
033	To balance the error weight, we penalize each unmatched	
034	cut by the average shot length (L_{avg}) of the target video:	
035	$E_{penalty} = (N_{FN} + N_{FP}) \cdot L_{avg}, \quad \text{where } L_{avg} = \frac{F_{total}}{N_{gt} + 1} \quad (S4)$	
036	The final SCA score ranges in $(0, 1]$, where 1 indicates a	
037	perfect match. The exponential formulation makes the met-	
038	ric particularly sensitive to large errors, heavily penalizing	
039	significant temporal deviations or incorrect shot counts.	
	A.2. Aesthetic Quality	040
	We assess the aesthetic and artistic value of each video	041
	frame using the LAION aesthetic predictor [9]. This met-	042
	ric reflects human-perceived qualities such as composition,	043
	color harmony, realism, naturalness, and overall artistic ap-	044
	peal of the generated frames.	045
	A.3. Semantic Consistency.	046
	We evaluate the alignment between the text prompt and the	047
	generated video by measuring two types of semantic con-	048
	sistency: global and shot-level. For global consistency, we	049
	extract the representations of the entire prompt and the full	050
	video using ViCLIP [12] and compute their cosine similar-	051
	ity. For shot-level consistency, the video is divided into seg-	052
	ments based on the input shot prompts, and the cosine sim-	053
	ilarity between each shot clip and its corresponding shot-	054
	level prompt features is calculated using ViCLIP.	055
	A.4. Intra-shot Consistency	056
	To compute intra-shot consistency, we first employ the pre-	057
	trained shot boundary detector TransNet V2 [11] to identify	058
	cut locations within the generated videos. We then compute	059
	subject consistency and background consistency, following	060
	the design of VBench [4].	061
	Subject consistency. For the main subject in the video, we	062
	measure the stability of its visual appearance across frames.	063
	Specifically, we extract DINO [1] features for each frame	064
	and compute the average cosine similarity between consec-	065
	utive frames and between each frame and the first frame.	066
	Background consistency. To evaluate the temporal stabil-	067
	ity of the scene background, we compute CLIP [8] feature	068
	similarities across frames. A higher similarity indicates	069
	a smoother and more coherent background transition over	070
	time.	071
	A.5. Inter-shot Consistency	072
	To assess consistency across different shots, a naive ap-	073
	proach would be to extract ViCLIP features for each shot	074
	and compute the cosine similarity between them. How-	075
	ever, since different shots may depict distinct characters or	076
	scenes, this simple comparison ignores diversity and may	077
	lead to biased results. To address this, we identify the char-	078
	acters described in the prompt and group the corresponding	079
	shots by character identity. We then compute the ViCLIP	080
	feature similarity among shots belonging to the same char-	081
	acter group to obtain the inter-shot consistency score.	082

Table S1. **Human evaluation results.** We report the percentage of “best” votes received by each model from our user study across three key perceptual criteria. The best and runner-up are in **bold** and underlined.

Method	Prompt Adherence \uparrow	Inter-Shot Consistency \uparrow	Overall Preference \uparrow
Wan2.2	2.67 %	1.11%	2.44%
StoryDiffusion+Wan2.2	6.44%	10.00%	5.11%
IC-LoRA+Wan2.2	<u>8.67%</u>	<u>15.78%</u>	<u>12.22%</u>
CineTrans	1.11%	0.44%	1.78%
HoloCine(Ours)	81.11%	72.67%	78.44%

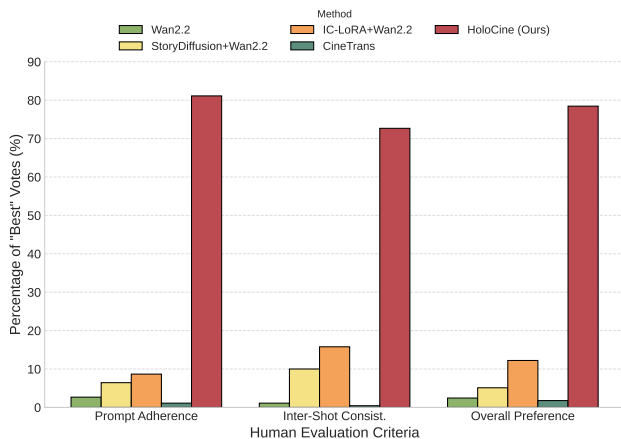


Figure S1. **Human evaluation results.** We plot the percentage of “best” votes received by each model across three key perceptual criteria: Prompt Adherence, Inter-Shot Consistency, and Overall Preference. Our method is overwhelmingly preferred by human evaluators across all categories, highlighting its superior ability to generate coherent and well-controlled multi-shot videos.

B. Human Evaluation

While our quantitative metrics quantify specific aspects, they cannot fully capture the holistic, perceptual quality of multi-shot videos. High-level concepts such as narrative coherence, logical consistency, and artistic appeal are best assessed by human evaluators.

To this end, we conduct a user study based on a forced-choice, best-of-N methodology. We recruited 30 participants from diverse backgrounds to ensure a comprehensive evaluation, comprising: (1) 10 computer vision researchers, (2) 10 professional artists and animators, and (3) 10 non-expert users. This diversity mitigates potential biases and provides a balanced assessment of technical, artistic, and general perceptual quality. We compare HoloCine (Ours) against the four baseline models. We used 25 diverse and challenging prompts from our benchmark set. Each participant was shown the text prompt and the five corresponding videos (Ours + 4 baselines), presented side-by-side in a randomized order to mitigate positional bias. Each participant was asked to evaluate 15 randomly assigned prompts, selecting the single best video for each of the three criteria

most central to this work:

- **Prompt adherence:** “Which video most accurately and faithfully follows the text prompt, including all described actions, characters, and shot transitions?”
- **Inter-shot consistency:** “Which video best maintains the consistency of characters, styles, and scene logic across the different shots?”
- **Overall preference:** “Considering all aspects (prompt following, consistency, and visual quality), which video do you prefer overall as the most coherent and high-quality multi-shot video?”

This methodology resulted in a total of 450 preference votes (30 participants \times 15 prompts) for each of the three criteria. The results of the user study are shown in Tab. S1 and Fig. S1, which clearly demonstrate that HoloCine achieves superior performance across all evaluation metrics, validating our method’s effectiveness in generating narrative-consistent and high-fidelity multi-shot videos.

C. Comparison with Closed-source Models

C.1. Comparison with Commercial Models

To further situate HoloCine’s capabilities, we conducted a qualitative comparison with leading closed-source commercial models. As illustrated in Fig. S2, while models like Vidu [10] and Kling 2.5 Turbo [5] generate visually impressive clips, they struggle with the core task of multi-shot storytelling. Given a hierarchical prompt, they produce a single, continuous shot, failing to understand or execute the specified shot transitions. In contrast, HoloCine demonstrates narrative comprehension and control on par with the latest state-of-the-art model, Sora 2 [7]. Both models successfully parse the prompt to generate a coherent sequence of distinct shots—transitioning from a medium shot to a dramatic close-up—while maintaining high character and stylistic consistency. This result validates that our framework’s ability to create complex, directed narratives is comparable to the leading proprietary solutions in the field.

C.2. Comparison with Closed-source Research Models

We further situate our method by comparing it with the most recent closed-source research work, LCT [3]. As the code



Figure S2. Qualitative comparison with state-of-the-art commercial models. While Vidu and Kling 2.5 Turbo fail to interpret multi-shot instructions and generate only a single, continuous clip, HoloCine successfully executes complex shot transitions. Our method demonstrates narrative control and consistency comparable to the leading closed-source model, Sora 2, accurately rendering the sequence from medium shots to close-ups as directed by the prompt.



Figure S3. Qualitative comparison with the closed-source model LCT [3]. Since the original prompts for LCT’s official examples are not provided, we employed a VLM to generate captions from their videos. These VLM-generated captions were then used as input for our method. The results are presented for a visual assessment of conceptual interpretation and sequence quality.

and model weights for LCT are not publicly available, a direct quantitative comparison is not feasible.

Therefore, we conduct a qualitative comparison using

the demonstration cases provided on their official project website. Since the exact prompts used to generate the LCT examples are not disclosed, we adopt a “re-captioning”



Figure S4. A failure case in causal reasoning. After an action (pouring water, [Shot 2]) is applied to an object (empty glass, [Shot 1]), the model fails to render its logical consequence. It incorrectly reverts to the initial empty state in [Shot 3], prioritizing visual consistency over the action’s outcome.

methodology. Specifically, we first use a state-of-the-art Video-Language Model (Gemini 2.5 pro [2]) to generate descriptive captions for the videos showcased by LCT, and then feed these VLM-generated captions directly as prompts into our method.

As illustrated in Fig. S3, we present a qualitative comparison of the results. This qualitative analysis shows that our method is capable of interpreting the same core concepts and narratives derived from the LCT outputs, generating coherent and high-fidelity video sequences.

D. Limitations

While our model excels at maintaining visual consistency, it exhibits limitations in causal reasoning. It can fail to comprehend how an action should alter an object’s physical state. Fig. S4 illustrates this clearly. Given an empty glass [Shot 1] and the action of water being poured into it [Shot 2], the model fails to render the logical outcome. Instead, it regenerates the glass as empty in [Shot 3], prioritizing visual consistency with the initial shot over the physical consequence of the action. This highlights a key challenge for future work: advancing from perceptual consistency to logical, cause-and-effect reasoning.

E. Discussion and Future Work

While our method has demonstrated strong performance in generating high-quality, coherent multi-shot videos (as evidenced by our supplementary website), we identify several promising avenues for future research and development.

Efficiency and deployment. A key challenge for practical deployment is the inference cost. Currently, generating a one-minute video requires approximately one hour on a single NVIDIA H800 GPU. We note recent works on acceleration, such as LightX2V [6], can distill 50-step models to 4 steps. Future work could integrate such techniques, potentially reducing inference time to 5 minutes and significantly enhancing user experience. Furthermore, leveraging FP8 quantization could drastically reduce the model’s memory requirement, lowering deployment barriers.

Architectural enhancements. From an architectural perspective, the bi-directional attention paradigm could evolve into a causal attention mechanism, enabling infinitely long

generation and reducing inference costs. Besides, for our sparse inter-shot attention, we found that simply selecting the first-frame token of each shot already achieves good results. Future work could consider further exploring different design, like introducing learnable strategies.

Towards end-to-end film generation. To move closer to genuine film production, introducing fine-grained controls for reference characters or keyframes is crucial. Another vital direction is audio-video co-generation(dialogue, music, effects) to achieve true end-to-end film generation.

Memory and world models. We observed that our model exhibits emergent memory capabilities, such as maintaining character and stylistic consistency across long sequences. This, to some extent, demonstrates the potential of large-scale video models to serve as a base for World Models. Future work could focus on explicitly scaling these models to further enhance these emergent properties, exploring their capabilities for more robust, long-range world state prediction.

Open sourcing and community. To facilitate research and progress within the community, we will open-source our code and pre-trained models. We believe this will provide a strong baseline for future multi-shot video generation and encourage further innovation.

References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pages 9630–9640. IEEE, 2021. 1
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit S. Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornrathop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilai Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Serkan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichter, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal,

- 245 Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5:
246 Pushing the frontier with advanced reasoning, multimodal-
247 ity, long context, and next generation agentic capabilities.
248 *CoRR*, abs/2507.06261, 2025. 4
- 249 [3] Yuwei Guo, Ceyuan Yang, Ziyang Yang, Zhibei Ma, Zhijie
250 Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context
251 tuning for video generation. *CoRR*, abs/2503.10589, 2025.
252 2, 3
- 253 [4] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si,
254 Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin,
255 Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin
256 Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Com-
257 prehensive benchmark suite for video generative models.
258 In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21807–
259 21818. IEEE, 2024. 1
- 260 [5] Kuaishou. Kling video model. <https://kling.kuaishou.com/en>, 2024. 2
- 261 [6] ModelTC. Lightx2v: A lightweight and high-performance
262 video diffusion model. <https://github.com/ModelTC/LightX2V>, 2024. Accessed: November 13,
263 2025. 4
- 264 [7] OpenAI. Sora 2 technical report. <https://openai.com/research/sora-2>, 2025. Accessed: 2025-10-15.
265 2
- 266 [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya
267 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,
268 Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen
269 Krueger, and Ilya Sutskever. Learning transferable visual
270 models from natural language supervision. In *Int. Conf.*
271 *Mach. Learn.*, pages 8748–8763. PMLR, 2021. 1
- 272 [9] Christoph Schuhmann. Improved aesthetic predictor.
273 <https://github.com/christophschuhmann/improved-aesthetic-predictor>, 2022. 1
- 274 [10] Shengshu Technology and Tsinghua University. Vidu: A
275 sora-level text-to-video model. <https://www.vidu.com>, 2024. Accessed: 2025-10-15. 2
- 276 [11] Tomás Souček and Jakub Lokoc. Transnet V2: an effective
277 deep network architecture for fast shot transition detection.
278 In *ACM Int. Conf. Multimedia*, pages 11218–11221. ACM,
279 2024. 1
- 280 [12] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu,
281 Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui
282 Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and
283 Yu Qiao. Internvid: A large-scale video-text dataset for mul-
284 timodal understanding and generation. In *Int. Conf. Learn.*
285 *Represent.* OpenReview.net, 2024. 1