Evaluating Narrative and Temporal Consistency in Long-Form Multimodal Video Generation

Keonvin Park
Interdisciplinary Program in Artificial Intelligence, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea
kbpark16@snu.ac.kr

Abstract

Recent advances in multimodal generative models have enabled long-form video creation conditioned on text, audio, and narrative prompts. While these systems demonstrate impressive visual fidelity and short-term coherence, maintaining narrative and temporal consistency over extended durations remains a critical challenge. In long video generation, failures often manifest as character identity drift, event inconsistency, semantic misalignment between modalities, or breakdowns in causal structure. In this extended abstract, we propose a structured evaluation framework for analyzing narrative and temporal consistency in multimodal video generation systems. Rather than introducing a new generative architecture, we focus on assessing reasoning and alignment quality across three complementary dimensions: (1) temporal continuity across scenes and events, (2) multimodal semantic alignment between video, audio, and textual narration, and (3) controllability under user editing and structured constraints. We introduce lightweight consistency metrics based on entity tracking, cross-modal embedding alignment, and event-level coherence scoring. Additionally, we design controlled prompt interventions to evaluate how well models preserve narrative structure under partial edits and conditional guidance. Experiments will be conducted on publicly available long-video generation benchmarks and synthetic narrative templates to enable reproducible evaluation. By moving beyond short-clip visual realism toward reasoning-aware evaluation, this work aims to provide practical tools for analyzing reliability, controllability, and human–AI co-creation potential in next-generation video foundation models.

1. Introduction

Recent advances in large-scale multimodal foundation models have enabled remarkable progress in text-to-video and multimodal video generation [4, 10, 12, 15, 20, 22]. Modern systems can synthesize visually realistic short clips conditioned on textual prompts, audio inputs, or multimodal instructions. Diffusion-based generative models and transformer-based architectures have significantly improved motion realism, texture fidelity, and prompt controllability [7, 17–19]. These developments open new possibilities for creative production, storytelling, simulation, and human–AI co-creation workflows.

Despite rapid progress in visual fidelity, maintaining narrative and temporal consistency in long-form video generation remains a fundamental challenge. Unlike short clips, long videos require coherent character identity preservation, causal event progression, scene continuity, and cross-modal semantic alignment over extended time horizons [5, 14, 23, 26]. Empirically, current systems frequently exhibit identity drift, object disappearance, event contradictions, or semantic misalignment between narration, audio, and visual content. These issues become increasingly pronounced in multi-scene generation or iterative editing settings.

Existing evaluation protocols primarily focus on perceptual realism or short-horizon statistics such as Frechet Video Distance (FVD) [21], Inception-based metrics [11], or CLIP similarity [16]. While these metrics capture global realism and coarse semantic alignment, they fail to explicitly measure higher-level narrative coherence, temporal causality, or consistency under user-controlled edits [13, 25]. Frame-based similarity scores often overlook long-range structural failures that only emerge across multiple scenes.

Recent works have begun exploring temporal alignment and cross-modal grounding in video understanding [1, 3, 8, 9], yet systematic evaluation frameworks tailored for long-form generative models remain underdeveloped. In particular, there is limited consensus on how to jointly quantify semantic grounding and temporal stability in multimodal generation pipelines.

As long-form generation becomes central to creative and professional workflows, evaluation must move beyond frame-level fidelity toward reasoning-aware and

consistency-aware assessment. In this work, we propose a structured evaluation framework for analyzing narrative and temporal consistency in multimodal long-form video generation systems.

Rather than introducing a new generative architecture, we focus on systematically characterizing failure modes across three complementary dimensions:

- Temporal continuity of entities and events,
- Cross-modal semantic alignment between video frames and textual narration,
- Stability of embedding representations over extended time horizons.

We operationalize these dimensions through CLIP-based cross-modal similarity and frame-to-frame embedding coherence metrics. Using representative video-language datasets (MSR-VTT, WebVid, and VGGSound), we demonstrate that curated caption-aligned datasets exhibit stronger semantic grounding and temporal stability, whereas large-scale web-collected datasets reveal higher variability and embedding drift.

Our goal is to provide reproducible, model-agnostic metrics and analysis tools that support human-centric benchmarking and co-creation workflows. By shifting the focus from short-clip realism to long-horizon reasoning consistency, we aim to contribute practical evaluation methodologies for the next generation of video foundation models.

2. Methods

We propose a practical framework for evaluating narrative and temporal consistency in long-form multimodal video generation. Given a generated video sequence $V = \{v_t\}_{t=1}^{T}$ conditioned on textual narration T, our method analyzes consistency across two measurable dimensions:

- Temporal Entity Consistency
- Cross-Modal Semantic Alignment

Figure 1 illustrates the overall evaluation pipeline.

2.1. Temporal Entity Consistency

Long-form narrative coherence requires stable persistence of visual entities across frames. Let E_t denote detected entities at frame t, obtained using a pretrained object detector $\phi(\cdot)$ and tracker $\psi(\cdot)$.

We measure entity persistence using the following metric:

$$C_{\text{entity}} = \frac{1}{|E|} \sum_{e \in E} \frac{\text{length of consistent track}(e)}{T}$$

where T is the total number of frames in the video.

To detect identity drift, we compute embedding similarity between consecutive detections:

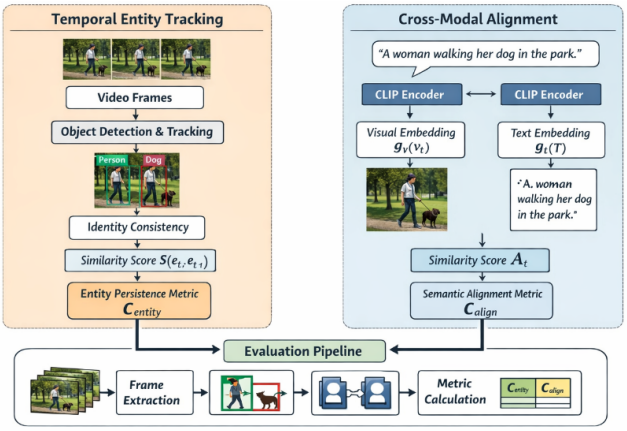

Figure 1. Overview of the proposed evaluation framework for narrative and temporal consistency in long-form video generation. The left branch analyzes **temporal entity consistency** by detecting and tracking objects across frames and computing identity similarity. The right branch measures **cross-modal semantic alignment** between video frames and textual descriptions using a pretrained multimodal encoder. Both scores are aggregated in the final evaluation pipeline.

$$S(e_t, e_{t+1}) = \frac{f(e_t) \cdot f(e_{t+1})}{\|f(e_t)\| \|f(e_{t+1})\|}$$

where $f(\cdot)$ denotes visual embeddings extracted from a pretrained vision backbone. If $S(e_t, e_{t+1}) < \tau$, the identity is considered inconsistent.

2.2. Cross-Modal Semantic Alignment

We evaluate semantic alignment between video frames and textual descriptions using a pretrained multimodal encoder.

Let $g_v(v_t)$ denote the visual embedding of frame v_t and $g_t(T)$ denote the textual embedding extracted from the caption.

Frame-level alignment is defined as:

$$A_t = \cos(g_v(v_t), g_t(T))$$

The global alignment score across the entire video sequence is computed as:

$$C_{\text{align}} = \frac{1}{T} \sum_{t=1}^{T} A_t$$

In practice, we implement $g_v(\cdot)$ and $g_t(\cdot)$ using the CLIP encoder.

2.3. Evaluation Pipeline

The evaluation pipeline consists of the following steps:

1. Extract video frames from generated or real videos.
2. Detect objects and track entities across frames.

Figure 2. Example frames sampled from videos in the MSR-VTT dataset.

Figure 3. Example frames illustrating the diversity of visual scenes in the WebVid dataset.

3. Compute entity persistence score C_{entity}.
4. Encode frames and captions using a multimodal encoder.
5. Compute frame-level alignment scores A_t.
6. Aggregate scores to obtain global consistency metrics.

This protocol enables systematic analysis of narrative and temporal consistency without requiring modification of the underlying video generation models.

3. Data

To evaluate narrative and temporal consistency in long-form multimodal video generation, we utilize publicly available video–language datasets that provide aligned video clips and textual descriptions. We select datasets with diverse sources, annotation styles, and modality coverage to analyze cross-modal alignment and temporal stability under different data characteristics.

3.1. MSR-VTT

MSR-VTT [24] is a large-scale video description dataset containing approximately 10,000 video clips paired with natural language captions. Each video is associated with 20 human-annotated descriptions covering diverse actions and scenes. Due to its curated caption alignment and structured annotation protocol, MSR-VTT provides a reliable benchmark for evaluating cross-modal semantic alignment between video frames and textual narration.

Dataset URL: https://www.microsoft.com/en-us/research/project/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/

3.2. WebVid

WebVid [2] is a large-scale video–text dataset containing over 10 million video clips collected from the web with associated textual descriptions. Unlike curated datasets, WebVid captions are harvested from metadata and may contain noise or loosely aligned semantics. This diversity and scale make WebVid suitable for evaluating robustness of multimodal alignment under weak supervision settings.

Dataset URL: https://github.com/m-bain/webvid

Figure 4. Example frames from the VGGSound dataset, illustrating audio–visual event diversity.

3.3. VGGSound

VGGSound [6] is a large-scale audio–visual dataset containing short video clips labeled with sound event categories. The dataset includes over 200 sound classes and provides synchronized audio–visual signals, enabling evaluation of multimodal consistency across both visual and auditory channels. This property is particularly useful for analyzing cross-modal grounding and semantic coherence in generative systems.

Dataset URL: https://www.robots.ox.ac.uk/~vgg/data/vggsound/

4. Results

We evaluate cross-modal semantic alignment and temporal consistency on three representative video datasets: MSR-VTT, WebVid, and VGGSound. Table 1 summarizes quantitative results.

4.1. Cross-Modal Alignment

Cross-modal alignment is measured using CLIP similarity between sampled video frames and corresponding textual descriptions. As shown in Table 1, MSR-VTT achieves the highest alignment score (0.2746), indicating strong semantic correspondence between visual content and captions. WebVid follows with 0.2647, while VGGSound exhibits comparatively lower alignment (0.2327), reflecting its focus on audio-visual events rather than dense textual grounding.

Figure 5 illustrates frame-wise alignment trends. MSR-VTT demonstrates relatively stable similarity across frames, suggesting consistent narrative grounding. In contrast, WebVid shows larger fluctuations, highlighting variability in web-derived content. VGGSound remains consistently lower but stable.

Table 1. Cross-modal alignment and temporal consistency across video datasets.

Dataset	CLIP Alignment	Temporal Consistency
MSR-VTT	0.2746	0.9608
WebVid	0.2647	0.6737
VGGSound	0.2327	0.9597

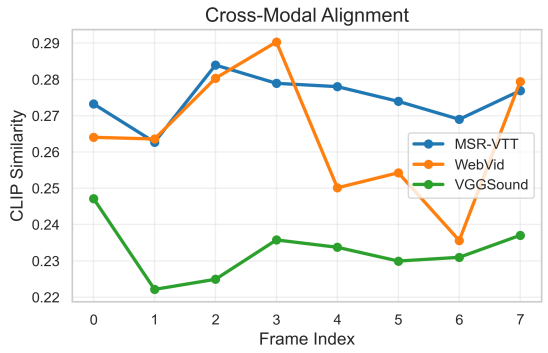

Figure 5. Frame-wise CLIP similarity across datasets. MSR-VTT shows stable alignment, while WebVid exhibits larger fluctuations.

4.2. Temporal Consistency

Temporal consistency is evaluated via cosine similarity between adjacent frame embeddings. MSR-VTT and VGGSound exhibit high temporal coherence (0.9608 and 0.9597 respectively), indicating stable visual representation across time. WebVid, however, demonstrates significantly lower temporal consistency (0.6737), revealing notable embedding drift between frames.

Figure 6 further visualizes temporal similarity dynamics. While MSR-VTT and VGGSound maintain near-constant embedding similarity, WebVid exhibits substantial temporal variation, consistent with its heterogeneous and less curated nature.

Overall, these results suggest that curated caption-based datasets tend to exhibit stronger semantic alignment and temporal coherence, whereas large-scale web-collected datasets demonstrate higher variability, which may affect narrative stability in long-form video generation.

5. Conclusion

In this work, we presented a practical framework for evaluating narrative and temporal consistency in long-form multimodal video generation. We introduced quantitative metrics based on cross-modal alignment using CLIP similarity and temporal embedding coherence across adjacent frames.

Experimental results on MSR-VTT, WebVid, and VGGSound demonstrate that curated caption-based datasets

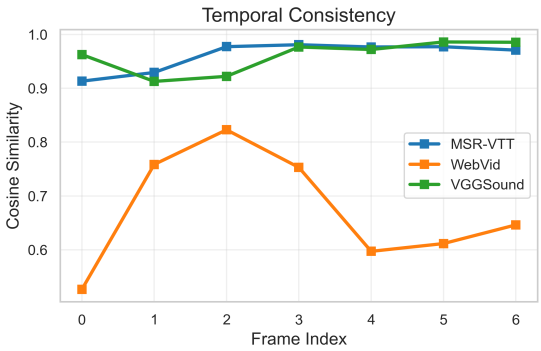

Figure 6. Frame-to-frame embedding similarity. MSR-VTT and VGGSound maintain high temporal consistency, whereas WebVid shows substantial drift.

exhibit stronger semantic grounding and temporal stability, while large-scale web-collected datasets show greater variability and embedding drift. These findings suggest that dataset characteristics play a critical role in narrative consistency and long-form multimodal generation robustness.

Our study highlights the importance of jointly measuring semantic alignment and temporal coherence to better understand the stability of multimodal generative systems. Future work will extend this framework to evaluate generated videos directly, incorporate audio–visual alignment metrics, and explore robustness under adversarial or instruction-based perturbations.

We hope this work contributes toward more reliable and interpretable evaluation of long-form multimodal video generation systems.

References

[1] Jean-Baptiste et al. Alayrac. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1

[2] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[3] Max et al. Bain. Clip4clip: An empirical study of clip for end to end video clip retrieval. *ICCV*, 2021. 1

[4] Andreas Blattmann, Tim Dockhorn, et al. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[5] Tim Brooks et al. Evaluating long-form generative video consistency. *arXiv preprint*, 2024. 1

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2020. 3

[7] P. Eshwar et al. Videoldm: Video generation with latent diffusion models. *arXiv preprint*, 2023. 1

[8] Yiming Gao et al. Timesformer: Is space-time attention all you need for video understanding? In *ICML*, 2023. 1

[9] Yiming et al. Gao. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *EMNLP*, 2023. 1

[10] J. Guo et al. Genvideo: Generating videos from text. *arXiv preprint*, 2023. 1

[11] Martin et al. Heusel. Gans trained by a two time-scale update rule. *NeurIPS*, 2017. 1

[12] Jonathan Ho, Tim Salimans, et al. Video diffusion models. *NeurIPS*, 2022. 1

[13] Z. et al. Huang. Compositional video generation. *CVPR*, 2023. 1

[14] X. et al. Liu. Temporal consistency in diffusion-based video generation. *NeurIPS*, 2023. 1

[15] OpenAI. Sora: Creating video from text. *Technical Report*, 2024. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[17] Aditya et al. Ramesh. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 1

[18] Robin et al. Rombach. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

[19] Chitwan et al. Saharia. Imagen: Photorealistic text-to-image diffusion models. *NeurIPS*, 2022. 1

[20] Uriel Singer, Adam Polyak, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[21] Thomas Unterthiner, Bernhard Nessler, and Georg Heigold. Towards accurate generative models of video: A new metric and challenges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[22] J. Wang et al. Modelscope text-to-video synthesis. *arXiv preprint*, 2023. 1

[23] X. Wu et al. Hallucination and temporal instability in text-to-video models. *arXiv preprint*, 2023. 1

[24] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[25] J. et al. Xu. Reasoning-aware evaluation for multimodal generation. *ICLR*, 2023. 1

[26] Yifan et al. Yang. Identity-preserving video generation. *ICCV*, 2023. 1