

A SUPPLEMENTARY MATERIALS

A.1 RESULTS - README: BEFORE WATCHING COMPRESSED VIDEOS

Please refer to the supplementary video for more results.

README: BEFORE WATCHING COMPRESSED VIDEOS

Dear Reviewers,

Due to the **100MB submission limit** for ICLR 2026, **we had to significantly compress the supplementary files**. The original file size for the Main Results video was 168.6 MB, and the Comparison video was 38.8 MB. **Please understand that the compressed version of videos may introduce visible artifacts**. If permitted by the conference, **we would be happy to provide the original high-quality videos/high-resolution images as well**.

Sincerely,
The Authors

A.2 ETHICS STATEMENT

Ethics Statement

All names, characters, and events appearing in this work are fictitious. They have no connection whatsoever to any real persons, places, buildings, products, or stories. All videos presented in the results are generated by an AI model based only on a text prompt dataset. The stories depicted in the results exist solely for the qualitative and quantitative evaluation of the generative model and are not intended to reflect or deliberately distort any specific individuals, organizations, regions, or historical events. It is explicitly stated that the generated stories do not represent or endorse any particular values, beliefs, cultural perspectives, or political positions in the real world. Users should be aware that the results of this research should not be considered factual information or used as a reference for decision-making.

A.3 COMPUTATION TIME AND MEMORY CONSUMPTION

Model	Step	Memory Consumption	Computation Time
Full Model	Inference	31.2276 GB / 80.0 GB (29,781 MiB)	880 sec
w/o DIPW	Inference	31.2108 GB / 80.0 GB (29,765 MiB)	357 sec
w/o TWB	Inference	31.2234 GB / 80.0 GB (29,777 MiB)	917 sec
w/o SAR	Inference	30.5775 GB / 80.0 GB (29,161 MiB)	1,205 sec
w/o ALL (Mochi)	Inference	25.2801 GB / 80.0 GB (24,109 MiB)	126 sec

Table 2: **Step-wise Memory Consumption and Computation Time Analysis**. This table presents a detailed comparison of memory consumption and computation time for our model and its ablated variants during inference, evaluated on an NVIDIA H100 GPU (80GB). The “Full Model” includes all proposed components, achieving a balanced trade-off between efficiency and quality. Removing TWB slightly increases inference time, while excluding DIPW significantly reduces computation time. In contrast, removing SAR leads to a substantial increase in computation time, highlighting its role in optimization. The baseline “w/o ALL (Mochi)” configuration has the lowest memory and fastest inference time but lacks all of the benefits we mentioned in this paper.

Table 2 presents a detailed comparison of memory consumption and computation time across different configurations of our model, evaluated using an NVIDIA H100 GPU (80GB). The analysis highlights the impact of various components on inference efficiency, demonstrating how each contributes to overall computational requirements.

Our full model achieves high-quality synthesis with a memory consumption of 31.21 GB and an inference time of 869 seconds. The result demonstrates that while individual components contribute

to different aspects of computational performance, our full model strikes a balance between efficiency and performance. The ablation study confirms that SAR plays a crucial role in speeding up inference, while TWB and DIPW contribute to reducing overall computational time. Notably, the baseline model (Mochi) is the most lightweight but lacks the high-fidelity outputs achieved by the full model.

Overall, our approach effectively balances computational cost and memory efficiency, making it suitable for real-world applications where both scalability and high-quality synthesis are critical.

A.4 DETAILED QUALITATIVE RESULTS

Left one in Figure 3 presents a qualitative comparison of a multi-character interaction sequence where Tom Cruise meets Taylor Swift at the Tidal Basin. Our model effectively preserves relative positioning and interactions between the two characters, while the baseline methods distort spatial relationships or fail to maintain interaction consistency. The background remains visually stable in our results, while competing methods introduce inconsistencies in the cherry blossom setting. Our model also captures emotional progression, with Tom Cruise transitioning from smiling to running and catching his breath, while other methods often fail to maintain facial expression consistency across frames. Furthermore, logical action continuity is evident in our results, as Tom Cruise’s movement smoothly follows the pursuit-and-capture narrative, whereas baseline models frequently introduce abrupt or disjointed transitions. Also, the bottom one illustrates a sequential movement sequence where Tom Cruise walks through Washington, D.C., passing key landmarks such as the White House and the Washington Monument. Our model ensures smooth temporal continuity, where each motion naturally leads into the next, while some baselines generate erratic or inconsistent movements. Unlike other methods that frequently reset poses abruptly between frames, our approach maintains the logical impact of prior movements on subsequent actions. Scene awareness is also preserved, ensuring architectural landmarks remain stable, while competing methods often introduce background distortions or misplaced elements. The right one demonstrates an interaction between a corgi and a red ball in Central Park. Our model ensures object permanence, keeping the ball consistently positioned and preventing unnatural displacement, while other methods often fail to track the object properly. The action sequence follows a natural progression from the corgi seeing the ball, biting it, and then kicking it, whereas some baselines introduce inconsistencies by skipping intermediate actions or generating illogical motion transitions. Our model also captures realistic canine behavior, such as tail wagging and playful spinning, while other methods often produce rigid or unnatural movements.

A.5 ABLATION STUDY

Without Time-weighted Blending (TWB), the generated frames lack temporal consistency, resulting in abrupt scene transitions where Tom Cruise and Taylor Swift appear as different identities across frames. Additionally, there is no meaningful interaction between them, making the sequence feel disconnected. The full model utilizes TWB to enforce bidirectional constraints, preserving spatial and temporal continuity across video segments. Without Dynamics-Informed Prompt Weighting (DIPW), the model fails to integrate Prompt 2 into the scene, leading to incomplete or inaccurate action sequences. The absence of DIPW prevents the model from smoothly interpolating between different prompt levels, causing a loss of intended action details and reducing narrative control. Our approach leverages DIPW to guide structured prompt blending, ensuring accurate action progression aligned with the evolving scene. Without Semantic Action Representation (SAR), the continuity of motion and action sequences deteriorates. The absence of SAR leads to disjointed character actions, where movements do not logically connect across frames, disrupting motion coherence. The full model incorporates SAR to encode high-level action semantics, ensuring that character behaviors evolve naturally and respond dynamically to preceding movements. These results highlight the necessity of each component: TWB maintains scene and identity consistency, DIPW enables structured prompt-driven action transitions, and SAR ensures smooth motion continuity and logical action sequences. The full integration of these modules allows for semantically aligned, visually coherent, and perceptually smooth long-form video generation.

Table 1 shows our evaluation results for ablation study. The full model (Ours) achieves the best overall performance, demonstrating that each module plays a crucial role in generating coherent

Method	CLIP-add \uparrow	CLIP-combined \uparrow	BLIP \uparrow	DINO \uparrow	LPIPS ($V_{12} - V_{(2N-1)(2N)}$) \downarrow
Vlogger	0.2889 \pm 0.0133	0.2998 \pm 0.0122	28.8509 \pm 1.2651	0.9254 \pm 0.0188	0.6193 \pm 0.0214
Mochi	0.2940 \pm 0.0139	0.3103 \pm 0.0180	29.5948 \pm 1.8717	0.9530 \pm 0.0098	0.6811 \pm 0.0386
Ours (Full Model)	0.3055 \pm 0.0095	0.3211 \pm 0.0122	29.5117 \pm 1.6253	0.9697 \pm 0.0056	0.6733 \pm 0.0501
Ours w/o DIPW	0.2883 \pm 0.0214	0.3013 \pm 0.0213	28.9450 \pm 1.7390	0.9418 \pm 0.0131	0.6412 \pm 0.0224
Ours w/o TWB	0.2944 \pm 0.0152	0.3030 \pm 0.0157	29.1403 \pm 1.1490	0.9655 \pm 0.0093	0.6803 \pm 0.0440
Ours w/o SAR	0.2926 \pm 0.0107	0.3005 \pm 0.0124	28.7187 \pm 0.8393	0.9526 \pm 0.0150	0.6396 \pm 0.0324

Table 3: **Quantitative evaluation of generated videos (mean \pm std).** Our method shows the best or comparable performance across multiple metrics. Lower LPIPS indicates better realism and temporal consistency. Standard deviations are omitted for brevity.

long-form videos. Although LPIPS is lower when SAR is removed (0.6396), this does not indicate better video quality. Instead, it reflects reduced motion complexity and less dynamic transitions, as SAR enhances character interactions and logical action continuity at the cost of slightly increased perceptual differences. The drop in CLIP-add (0.2926) and CLIP-combined (0.3005) without SAR further confirms that it is essential for maintaining text-video alignment. Similarly, removing DIPW leads to weaker prompt-based scene transitions, while TWB removal results in the highest LPIPS (0.6887), indicating degraded temporal smoothness. These results highlight that SAR, DIPW, and TWB must be combined to ensure text-aligned, semantically structured, and perceptually coherent video generation. The method for quantitative evaluation follows Section 4.4

A.6 DETAILED QUANTITATIVE EVALUATION

Table 3 reports the same quantitative evaluation as Table 1, but includes standard deviation values (mean \pm std) computed across 8 validation stories. The inclusion of standard deviations provides insight into the stability and consistency of each method across diverse prompts and scenes. Our full model not only achieves the best average performance across all metrics—including CLIP-add, CLIP-combined, BLIP, DINO, and LPIPS—but also shows stable results with relatively low variance. Each ablation variant (w/o DIPW, TWB, SAR) demonstrates noticeable drops in performance or increased variability, highlighting the contribution of each component to the overall video generation quality.

A.7 DATASET PLOT AND CHARACTER

Our dataset is designed to comprehensively represent both animate (e.g., humans, animals) and inanimate objects (e.g., balls, buses, cars, boats, airplanes) to ensure diversity in story generation. Each story plot includes at least one visually distinguishable action performed by an entity, such as throwing a ball, boarding a vehicle, running, pressing a button, or walking, to enhance dynamic storytelling.

The dataset covers eight diverse locations: New York City, Washington D.C., Paris, London, Los Angeles, San Francisco, Chicago, and Las Vegas. Among these, seven locations (excluding New York City) feature at least two characters (or two animals) per plot, introducing interactions and multi-character dynamics. Each story plot consists of 12 to 13 sequential scenes, with Chicago, Las Vegas, and New York City including 12 scenes per story, while all other locations include 13 scenes per story. To ensure diversity in character representation, we introduce two distinct character sets: a celebrity set (Tom Cruise & Taylor Swift, Elon Musk & Angelina Jolie) and an animal set (Corgi Dog & Siamese Cat, Panda & Fox). Given eight different locations, 12-13 scenes per story, and four distinct character settings, the dataset consists of a total of 404 prompts. ($12 \times 3 + 13 \times 5 = 101$, $101 \times 4 = 404$)

Prompt 1 provides a broad scene description that establishes the setting and context, while Prompt 2 introduces specific actions performed by the character within that scene. For example, in a New York City subway scenario, Prompt 1 may be “Tom Cruise is inside of the subway train,” setting up the environment, whereas Prompt 2 specifies an action such as “Tom Cruise is sitting.” This structure enables fine-grained control over character movement and interactions while maintaining coherence in scene transitions.

This structured approach allows for a broad range of environments, interactions, and character-driven narratives, making it well-suited for evaluating story generation models. Actions that humans can perform but animals cannot may be adapted accordingly. For example, since a dog cannot pick up and throw a ball with both hands, such an action is replaced with the dog kicking the ball instead.

To be specific, in a scene where the character arrives at Central Park, Prompt 1 describes, “Tom Cruise sees a red ball in Central Park,” providing situational context, while Prompt 2 refines the action in detail with, “Tom Cruise is picking up a red ball in Central Park” and later “Tom Cruise is throwing a red ball in Central Park.” Similarly, in an animal-based variation, Prompt 1 states, “A corgi dog sees a red ball in Central Park,” while Prompt 2 adapts the action appropriately, such as “A corgi dog is biting a red ball in Central Park” and “A corgi dog is kicking a red ball in Central Park.”

This approach ensures that actions are naturally adapted for different entities, particularly when an action performed by a human (e.g., gripping and throwing a ball) must be substituted with a more plausible behavior for an animal (e.g., biting or kicking the ball). The dataset maintains consistency in narrative progression while allowing for variations based on both character type and setting. Some examples of dataset plots can be found below.

```
prompt_nyc = [
    "Tom Cruise is inside of the subway train", "Tom Cruise is sitting",
    "Tom Cruise is looking out the subway window", "Tom Cruise now stands out",
    "Tom Cruise is getting off the NYC subway train", "Tom Cruise is walking",
    "Tom Cruise is walking up the subway exit stairs", "Tom Cruise is looking around",
    "Tom Cruise is looking at the streets of Times Square, NYC", "Tom Cruise is tilting his head curiously",
    "Tom Cruise is walking on the streets of Times Square, NYC", "Tom Cruise is walking",
    "Tom Cruise is waiting for a bus at the Times Square bus stop in NYC", "Tom Cruise is standing",
    "Tom Cruise is getting on a bus at the Times Square bus stop", "Tom Cruise is walking",
    "Tom Cruise is looking out the bus window at the city view", "Tom Cruise is sitting",
    "Tom Cruise has arrived at Central Park", "Tom Cruise is strolling",
    "Tom Cruise sees a red ball in Central Park", "Tom Cruise is observing it curiously",
    "Tom Cruise is picking up a red ball in Central Park", "Tom Cruise is gripping it firmly",
    "Tom Cruise is throwing a red ball in Central Park", "Tom Cruise is watching its trajectory"
]
prompt_nyc_corgi = [
    "a corgi dog is inside of the subway train", "a corgi dog is sitting",
    "a corgi dog is looking out the subway window", "a corgi dog now stands out",
    "a corgi dog is getting off the subway train", "a corgi dog is walking",
    "a corgi dog is looking at the streets of Times Square, NYC", "a corgi dog is tilting its head curiously",
    "a corgi dog is walking on the streets of Times Square, NYC", "a corgi dog is wagging its tail",
    "a corgi dog is waiting for a bus at the Times Square bus stop in NYC", "a corgi dog is standing",
    "a corgi dog is getting on a bus at the Times Square bus stop", "a corgi dog is walking",
    "a corgi dog is looking out the bus window at the city view", "a corgi dog is sitting",
    "a corgi dog has arrived at Central Park", "a corgi dog is sniffing the ground",
```



```

918     "a corgi dog sees a red ball in Central Park", "a corgi dog is
919         looking up at the sky",
920     "a corgi dog is biting a red ball in Central Park", "a corgi dog is
921         wagging its tail",
922     "a corgi dog is kicking a red ball in Central Park", "a corgi dog is
923         playfully spinning"
924 ]
925 prompt_dc = [
926     "Tom Cruise is walking around the White House", "Tom Cruise is
927         observing the architecture",
928     "Tom Cruise is strolling along the streets near the U.S. Capitol in
929         Washington D.C.", "Tom Cruise is looking around with curiosity",
930     "Tom Cruise is passing by the Washington Monument", "Tom Cruise is
931         walking",
932     "Tom Cruise is walking along the Tidal Basin surrounded by cherry
933         blossoms", "Tom Cruise is taking a leisurely stroll",
934     "Tom Cruise stops for a moment at the Tidal Basin surrounded by
935         cherry blossoms", "Tom Cruise is sitting",
936     "Tom Cruise is jogging along the Tidal Basin surrounded by cherry
937         blossoms", "Tom Cruise is enjoying the fresh air",
938     "Tom Cruise encounters Taylor Swift at the Tidal Basin surrounded by
939         cherry blossoms", "Tom Cruise is smiling",
940     "Tom Cruise is trying to catch Taylor Swift at the Tidal Basin
941         surrounded by cherry blossoms", "Tom Cruise is running",
942     "Taylor Swift is eventually caught up with by Tom Cruise at the Tidal
943         Basin surrounded by cherry blossoms", "Tom Cruise is catching
944         his breath and grinning",
945     "Tom Cruise and Taylor Swift are enjoying the cherry blossoms
946         together", "Tom Cruise and Taylor Swift are sitting side by side
947         ",
948     "Tom Cruise and Taylor Swift are admiring the cherry blossoms at the
949         Tidal Basin", "Tom Cruise and Taylor Swift are lying on the grass
950         looking up at the sky",
951     "Tom Cruise and Taylor Swift are lying on the grass at the Tidal
952         Basin surrounded by cherry blossoms, slowly closing their eyes",
953     "Tom Cruise and Taylor Swift are resting peacefully",
954     "Tom Cruise and Taylor Swift have fallen asleep at the Tidal Basin
955         surrounded by cherry blossoms", "Tom Cruise and Taylor Swift are
956         peacefully dozing off"
957 ]
958 prompt_dc_corgi = [
959     "a corgi dog is walking around the White House", "a corgi dog is
960         sniffing the ground",
961     "a corgi dog is strolling along the streets near the U.S. Capitol in
962         Washington D.C.", "a corgi dog is walking",
963     "a corgi dog is passing by the Washington Monument", "a corgi dog is
964         looking around",
965     "a corgi dog is walking along the Tidal Basin surrounded by cherry
966         blossoms", "a corgi dog is taking a leisurely stroll",
967     "a corgi dog stops for a moment at the Tidal Basin surrounded by
968         cherry blossoms", "a corgi dog is sitting",
969     "a corgi dog is running along the Tidal Basin surrounded by cherry
970         blossoms", "a corgi dog is excitedly running",
971     "a corgi dog encounters a siamese cat at the Tidal Basin surrounded
972         by cherry blossoms", "a corgi dog is tilting its head curiously",
973     "a corgi dog is chasing a fleeing siamese cat with at the Tidal Basin
974         surrounded by cherry blossoms", "a corgi dog and a siamese cat
975         are running",
976     "a siamese cat is finally caught by a corgi dog at the Tidal Basin
977         surrounded by cherry blossoms", "a corgi dog is gently wagging
978         its tail",
979     "a corgi dog and a siamese cat are sniffing the scent of cherry
980         blossoms together at the Tidal Basin", "a corgi dog and a siamese
981         cat are sitting",

```

```
972 "a corgi dog and a siamese cat are enjoying the cherry blossoms at
973 the Tidal Basin", "a corgi dog and a siamese cat are lying down",
974 "a corgi dog and a siamese cat are lying down at the Tidal Basin
975 surrounded by cherry blossoms, slowly closing their eyes", "a
976 corgi dog and a siamese cat are lying down",
977 "a corgi dog and a siamese cat have fallen asleep at the Tidal Basin
978 surrounded by cherry blossoms", "a corgi dog and a siamese cat
979 are peacefully sleeping"
980 ]
981 (...)
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
```



Figure 4: **Ablation Study** Each row shows a different setting: Full Model, w/o TWB, w/o DIPW, w/o SAR, and a baseline (Mochi). The full model produces the most coherent motion and character interactions.