OneStory: Coherent Multi-Shot Video Generation with Adaptive Memory
Supplementary Material

Zhaochong An[1,2], Menglin Jia[1], Haonan Qiu[1], Zijian Zhou[1], Xiaoke Huang[1], Zhiheng Liu[1], Weiming Ren[1], Kumara Kahatapitiya[1], Ding Liu[1], Sen He[1], Chenyang Zhang[1], Tao Xiang[1], Fanny Yang[1], Serge Belongie[2], Tian Xie[1,*]
[1]Meta AI, [2]University of Copenhagen
[*]Project lead

A. Additional Training Details

This section provides expanded details on the training formulation used in our model, including the unified three-shot training setup and the construction of frame-level pseudo-labels. These details complement Sec. 4.2 and Sec. 4.4 of the main paper.

A.1. Unified Three-Shot Training

As discussed in Sec. 3 of the main paper, the dataset contains videos with varying numbers of shots, with two-shot sequences being the most common and three-shot sequences relatively fewer. Training directly on sequences of non-uniform length leads to unstable optimization. To mitigate this, we unify all training samples into a *three-shot format* by synthesizing an additional shot for two-shot videos.

Synthetic shot construction. Given a two-shot sequence $(S_{\text{first}}, S_{\text{last}})$, we create a synthetic shot S_{syn} using one of:

(i) Cross-video insertion: inserting a shot randomly sampled from another video.

(ii) Augmented-first-shot variant: applying spatial or color transformations to S_{first}.

This results in synthetic triplets that, for each sample, take one of the two forms:

$$\big(S_{\text{first}}, S_{\text{syn}}, S_{\text{last}}\big) \quad \text{or} \quad \big(S_{\text{syn}}, S_{\text{first}}, S_{\text{last}}\big), \tag{1}$$

while the real triplets are represented in the structure $(S_{\text{first}}, S_{\text{second}}, S_{\text{last}})$. In all cases, S_{last} serves as the prediction target.

Training objective. The model is trained to generate the final shot S_{last} conditioned on the first two shots and its caption C_{last}:

$$\mathcal{L}_{\text{shot}} = \mathbb{E}\big[\mathcal{L}_{\text{diff}}\big(\mathcal{G}(S_{\text{first}},\, S_{\text{syn}/second},\, C_{\text{last}}),\; S_{\text{last}}\big)\big]\,, \tag{2}$$

where $\mathcal{L}_{\text{diff}}$ denotes a rectified-flow diffusion loss [1–3]. This unified formulation standardizes all training samples to a consistent three-shot structure and enables unified three-shot training, improving optimization stability.

A.2. Frame Relevance Pseudo-Labels

To assist the learning of the frame relevance scores $\mathbf{S}$, we construct frame-level pseudo-labels $\mathbf{y} = \{y_r\}_{r=1}^{F}$ that approximate the relevance of each historical frame in $\mathbf{M}$ to the target shot. The pseudo-labels incorporate both real and synthetic frames introduced in Sec. A.1.

Real historical frames. For frames originating from real historical shots, we compute cosine similarity between each historical frame and the target shot using DINOv2 [4] and CLIP [5] embeddings, producing a scalar relevance score. These pseudo-labels help the Frame Selection module prioritize visually and semantically aligned frames while down-weighting irrelevant ones.

Synthetic frames. Frames from synthetic shots introduced in Sec. A.1 are assigned coarse relevance labels: $y_r = -1$ for randomly inserted shots to indicate clear irrelevance, and $y_r = 0$ for augmented-first-shot variants to reflect partial relevance. These labels explicitly guide the selector to down-weight non-informative or misleading frames.

Supervision loss. The predicted relevance scores $\mathbf{S}$ are supervised using a regression loss:

$$\mathcal{L}_{\text{sel}} = \frac{1}{F}\sum_{r=1}^{F}(s_r - y_r)^2, \tag{3}$$

where $s_r = \mathbf{S}[r]$ denotes the predicted relevance score for the r^{th} historical frame. The full training objective is given by:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{shot}} + \lambda\,\mathcal{L}_{\text{sel}}, \tag{4}$$

with λ controlling the weight of the selector supervision loss. This joint optimization encourages the model to identify informative context frames while maintaining high-fidelity generation.

B. Additional Details on Evaluation Benchmark

We construct a human-centric benchmark for both T2MSV and I2MSV to evaluate multi-shot video generation under realistic narrative conditions. As introduced in Sec. 3 of the main paper, each shot is paired with a referential caption following a progressive narrative flow, reflecting real-world storytelling. To comprehensively evaluate MSV performance, the benchmark spans three canonical multi-shot storytelling patterns:

1. **Main-subject consistency.** Multiple shots focus on the same character(s), who may appear in different environments or perform different actions. This pattern evaluates the model's ability to preserve identity under various cross-shot changes.
2. **Insert-and-recall with an intervening shot.** A shot introducing a new scene, such as an environment-only view or a new character, is inserted mid-sequence, after which the narrative returns to the primary subject(s) and later revisits the intervening shot. This pattern stresses the model's ability to maintain long-range memory and remain robust to temporal distractors.
3. **Composable generation.** Characters introduced separately in earlier shots are composed together in later shots. This tests whether the model can correctly integrate multiple narrative threads into a coherent shared scene.

In total, we curate 64 six-shot test cases for T2MSV and 64 six-shot test cases for I2MSV, covering a diverse range of subjects, environments, and complex cross-shot relationships, thereby ensuring comprehensive MSV performance evaluation. More examples are provided in our Project Page.

C. Additional Qualitative Results

Generating *coherent multi-shot videos* that faithfully follow narrative captions is essential for real-world storytelling. Here, we analyze our model from three perspectives, using examples from the main paper to illustrate its ability to maintain continuity across shots. Additional video qualitative results are available our Project Page.

Identity consistency. Our model preserves character identity across long-range shots under diverse variations. In the 1^{st} example of Fig. 1 in the main paper, the same subject remains consistent across changes in viewpoint (Shots 4, 5) and actions (Shots 1, 3, 8). This illustrates the effectiveness of our adaptive memory in maintaining stable long-range identity cues.

Background details. Beyond character fidelity, our model maintains consistent background details across shots, enabling spatially coherent story progression. In the 2^{nd} example of Fig. 1 in the main paper, fine-grained elements such as plants and fences remain aligned from Shot 1 to Shot 7 despite large cross-shot dynamics. Similarly, in the 3^{rd} example, the red flowers reappear consistently across Shots 1, 4, 5, 6, 7, and 9, demonstrating the model's ability to preserve scene layout and spatial structure.

Reappearance and composition. Realistic narratives often involve disappear–reappear patterns and the merging of multiple narrative threads through composable generation. Our model effectively recalls characters or environments that reemerge after several intervening shots, *e.g.*, Shots 4 and 9, or Shots 2 and 6 in the 2^{nd} example of Fig. 1 in the main paper. Furthermore, in Shot 7 of the same example, the woman from Shot 1 and the man from Shot 4 appear together, demonstrating the model's capacity to unify distinct visual narratives into a coherent multi-subject scene.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1

[2] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint*, 2022.

[3] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint*, 2022. 1

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint*, 2023. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1