STRUCTURED MOTION-INSTANCE GRAPHS AND PERSISTENT MEMORY FOR LONG-HORIZON VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating long videos in successive chunks requires preserving instance identity, motion, and interactions across chunk boundaries. We propose a framework that couples a Motion-Instance Graph (MIG) with persistent Instance Memory (IM) for controllable long-horizon video generation. MIG organizes subjects, static regions, trajectories, interactions, and event states into a typed graph, which is converted into frame-level motion and spatial conditions. IM retrieves historical instance states before each chunk and updates them from generated observations, allowing subsequent chunks to reuse appearance, trajectory, and interaction information. Lightweight MI-Adapter and LoRA modules inject graph conditions and memory tokens into a frozen video diffusion backbone. First-frame anchoring and frame-aware corruption further support cross-chunk consistency. We evaluate the framework on 160 conditions with three random seeds, conduct incremental ablations, and test generation at 240, 720, and 1,440 frames. On CogVideoX-2B, normalized average displacement error decreases from 0.2108 for the trajectory-control baseline to 0.1850, while background consistency increases from 0.618 for the long-video baseline to 0.750. Evaluations on Wan2.1-1.3B and VideoCrafter2 also show improvements over their paired baselines. The framework provides explicit instance-level control and persistent state for long-video generation without updating the backbone parameters.

Keywords: long-horizon video generation; motion-instance graph; instance memory; video diffusion; parameter-efficient adaptation.

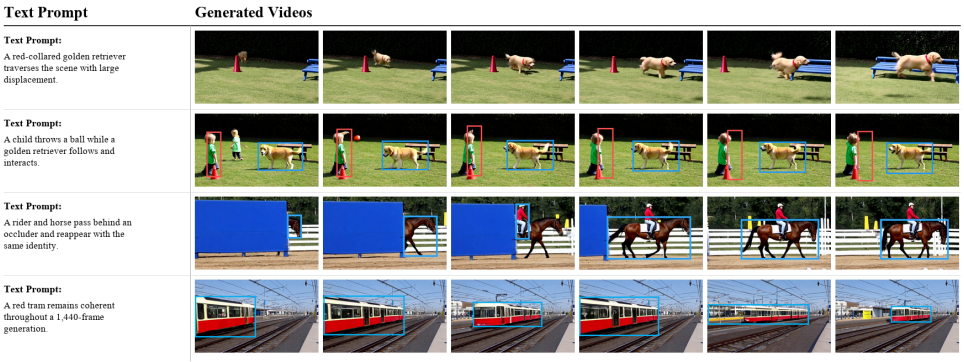

Figure 1: Representative results for large displacement, multi-instance interaction, occlusion recovery, and generation up to 1,440 frames.

1 INTRODUCTION

Long-horizon video generation requires more than extending a short clip: the same subjects, scene structure, trajectories, and interactions must evolve coherently over many generation windows.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Modern text-to-video diffusion models provide strong local visual quality and prompt alignment, but their conditions are usually tied to the current clip (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023; Chen et al., 2024; Yang et al., 2025b). Maintaining the identity, trajectory, and relations of each subject becomes more difficult as generation proceeds through successive windows (Harvey et al., 2022; Qiu et al., 2024; Qi et al., 2025; Po et al., 2026; Zhou et al., 2024).

Prior work extends temporal context through noise rescheduling, history aggregation, autoregressive training, and persistent caches (Harvey et al., 2022; Qiu et al., 2024; Qi et al., 2025; Po et al., 2026; Zhou et al., 2024; Huang et al., 2025; Yang et al., 2025a). Long-horizon generation requires the same subjects and their relations to persist as the scene evolves. The state established in one chunk must remain available for retrieval and update in the next. We address this requirement through explicit instance-state tracking and cross-chunk state maintenance.

We address this problem with a structured motion-instance framework that couples MIG and persistent IM. MIG represents subjects, static regions, trajectories, events, and their relations as typed nodes and edges. The DMG Planner converts this structure into frame-level conditions. IM maintains records indexed by instance identity, retrieves them before generation, and updates them from the resulting chunk. MI-Adapter/LoRA connects both sources to a frozen video backbone. First-frame anchoring retains an appearance reference, and frame-aware corruption exposes training to imperfect historical conditions.

The main comparison uses 160 conditions and three seeds, with the base model, three control configurations, and our method all built on CogVideoX-2B. Relative to the best competing values in Table 1, ADE decreases from 0.2108 to 0.1850 and background consistency increases from 0.618 to 0.750. Component ablations demonstrate gains from structural conditioning and memory, while blind evaluation and three-backbone comparisons confirm the framework's improvements. Long-video tests extend generation to 1,440 frames. Our contributions are threefold:

- We introduce MIG, which unifies subjects, regions, trajectories, interactions, and event states in an explicit, compositional, and updatable graph for long-horizon generation.
- We introduce persistent Instance Memory, which maintains, retrieves, and updates per-instance appearance, motion, and relational state across generation chunks.
- We integrate MIG and IM into frozen video backbones through lightweight MI-Adapter/LoRA modules and demonstrate consistent improvements through baseline comparisons, component ablations, long-video tests, cross-backbone transfer, and blind evaluation.

2 Related Work

Video diffusion and long-horizon generation. Video Diffusion Models, Make-A-Video, VideoLDM, VideoCrafter2, and CogVideoX establish strong short-clip priors, while long-video methods extend temporal range through longer context, noise rescheduling, window fusion, autoregressive aggregation, or consistency-aware attention (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023; Chen et al., 2024; Yang et al., 2025b; Harvey et al., 2022; Qiu et al., 2024; Qi et al., 2025; Po et al., 2026; Zhou et al., 2024; Henschel et al., 2025; Mao et al., 2026). Self Forcing trains autoregressive video diffusion using self-generated context and KV-cached rollouts (Huang et al., 2025). LongLive combines streaming long tuning, KV-recache, and frame-sink attention for interactive long-video generation (Yang et al., 2025a). These methods address training and context management; our framework maintains instance identity, trajectories, and interactions across generation chunks.

Controllable and instance-aware generation. Spatial layouts, trajectories, reference images, subject adaptation, and compositional planning improve control over short videos (Wu et al., 2024; Xu et al., 2026; Mou et al., 2024b; Wang et al., 2023b; Yin et al., 2025; Sun et al., 2025; Wu et al., 2023; Khachatryan et al., 2023; Guo et al., 2024; Ren et al., 2024; Zhang et al., 2026). ControlNet, T2I-Adapter, LoRA, DreamBooth, and IP-Adapter provide efficient pathways for injecting external conditions or identity references (Hu et al., 2022; Zhang et al., 2023; Mou et al., 2024a; Ruiz et al., 2023; Hu et al., 2023). Our state representation keeps these conditions associated with instance identity as trajectories and interactions evolve between chunks.

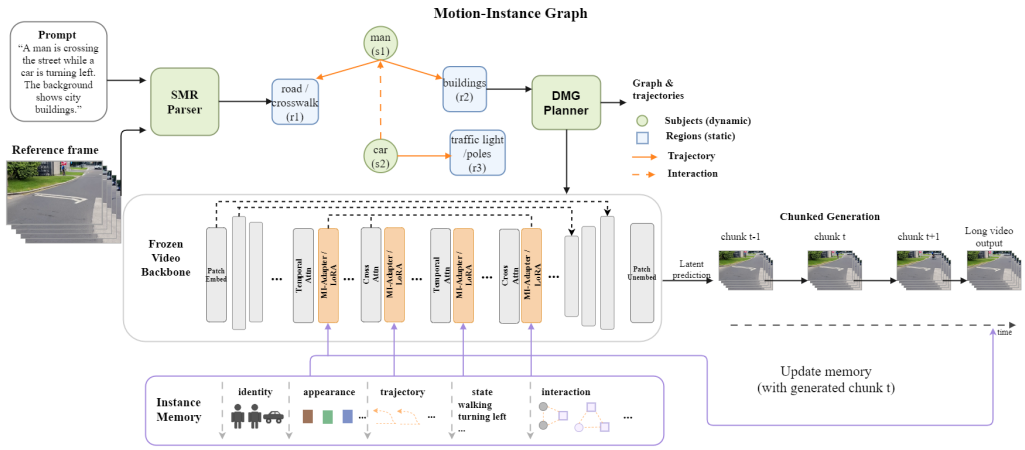

Figure 2: Overview of the closed-loop framework. MIG and the DMG Planner produce explicit motion-instance conditions, MI-Adapter injects them into a frozen backbone, and generated observations update Instance Memory.

Memory and structured state. Scene graphs organize objects and relations, and video-object memory methods maintain historical features for association and segmentation (Oh et al., 2019; Cheng et al., 2022; Xu et al., 2017; Zellers et al., 2018). Memory is also used directly in generation: FreeMem maintains noise-, token-, and attention-level memories without additional tuning (Peng et al., 2026), while WorldWeaver combines perceptual conditioning with a depth-based memory bank and segmented noise scheduling (Liu et al., 2025). We organize retrieved records around instance identity and typed relations in the graph that conditions the next generation window.

3 Method

3.1 Problem Formulation

Given a text prompt p, an optional reference frame x_{ref}, and target length T, we generate $V = \{x_t\}_{t=1}^{T}$. We generate the video through K successive overlapping windows V^k. For window k, G^k denotes the current motion-instance graph and M^{k-1} the memory available from previous windows. Each instance state contains identity, appearance, position, trajectory, event state, and interaction records. Generation of chunk k is

$$V^k \sim p_{\theta,\phi}\left(V^k \mid p, x_{\text{ref}}, G^k, M^{k-1}\right), \tag{1}$$

where θ denotes the frozen backbone and ϕ the trainable conditioning and memory parameters. Each chunk inherits and updates the established world state through MIG and IM while the video backbone remains frozen.

3.2 Overview

Figure 2 shows the generation loop. The SMR Parser extracts subjects, actions, regions, temporal phases, and interactions. MIG organizes these fields, and the DMG Planner converts them into frame-level conditions. The generator receives those conditions together with retrieved memory tokens through MI-Adapter/LoRA. Observations from the generated chunk are then used to update the records retrieved for the next chunk.

3.3 Motion-Instance Graph and Dynamic Planner

MIG converts language and reference frames into an executable state for each generation chunk. For chunk k, $G^k = (N^k, E^k)$ contains dynamic-subject nodes o_i, static-region nodes r_j, trajectory nodes τ_i, state nodes q_i, and event nodes e_m. Its edges encode subject-region, subject-trajectory,

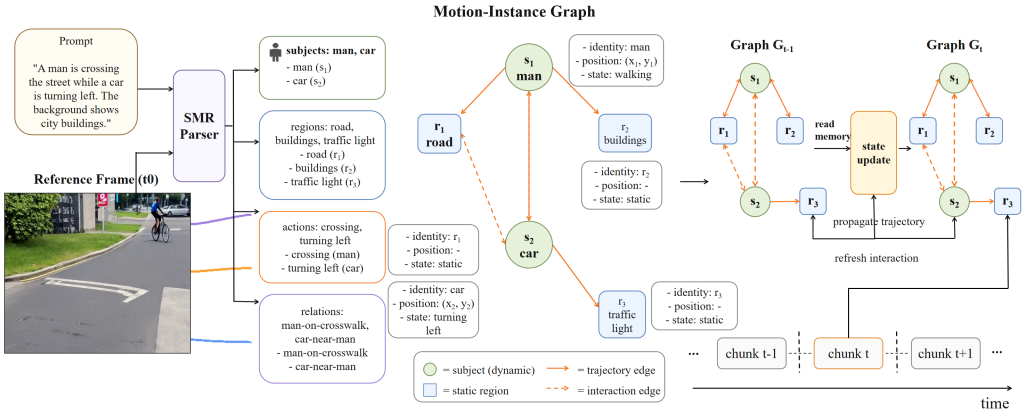

Figure 3: MIG construction and conversion into frame-level dynamic motion guidance.

subject-subject interaction, and subject-state update relations. Node attributes and relations are updated after each chunk, so the same graph records the current scene, motion plan, and state transitions (Wang et al., 2023b; Xu et al., 2017; Zellers et al., 2018).

The SMR Parser extracts action verbs, spatial relations, temporal order, and instance descriptions, then combines them with observations from the reference frame or historical chunks. The DMG Planner projects G^k into a frame-level tensor C^k containing bounding-box occupancy, instance-ID projection, object-relative motion, visibility, static support, obstacle regions, and global camera motion. Object-relative and camera motion are modeled separately and recombined in screen space, which keeps viewpoint changes separate from subject trajectories. For a single subject, the planner uses start/end positions, motion type, and region constraints; for multiple subjects, interaction edges regulate relative distance, occlusion, and contact timing.

The graph keeps motion targets and constraints associated with the same subject as the scene evolves. Subject–trajectory edges bind motion to identity, subject–region edges attach spatial constraints, and subject–subject edges represent interactions. IM supplies the historical records associated with these entities before the next chunk is generated. The schema fixes the types of entities and relations, while the subjects, attribute values, and connections change with the scene. An interaction therefore remains associated with the participating subjects when their positions are updated.

3.4 MI-Adapter for Frozen Video Backbones

We freeze the video backbone f_θ and optimize the MI condition projector, LoRA branches, and memory projections. For chunk k, MIG conditions and retrieved memory tokens are projected to a latent-aligned spatiotemporal representation and injected into temporal and cross-attention blocks. Each adapted projection is

$$W' = W + BA, \tag{2}$$

where W is frozen and A, B are trainable low-rank matrices. Because the graph and memory are updated after every chunk, the injected condition changes with the scene.

In CogVideoX-2B, MI-Adapter projects the nine-channel motion-instance map into latent-aligned features, while LoRA adapts attention projections. The transformer, VAE, and text encoder remain frozen. The detailed injection architecture is shown in Figure S1; Appendix A.2 and Table S1 give the channel definitions, tensor shapes, target projections, and optimization settings.

3.5 Instance Memory and Chunked Generation

Each instance o_i has a memory record $m_i = (k_i, v_i)$. The key supports retrieval from the current MIG, and the value stores identity, appearance, trajectory, event state, interactions, last-seen time, and confidence. Let $\mathbf{u}_i$ denote the retrieval query associated with instance i in the current MIG. The query and memory keys have a common embedding dimension d. Retrieved values contain the historical state used to condition generation. Before chunk k, the current graph retrieves relevant

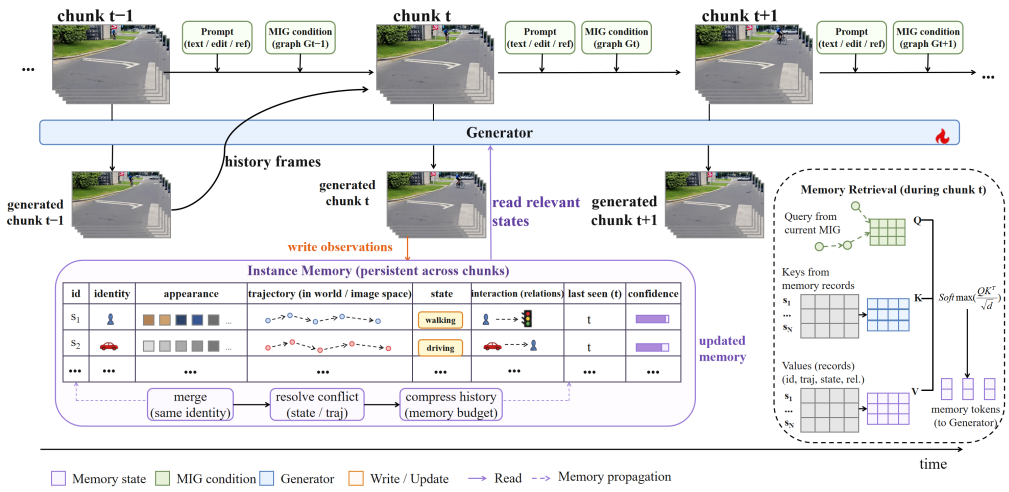

Figure 4: Persistent Instance Memory with retrieval, state update, conflict resolution, and history compression across generated chunks.

records through similarity-weighted top-k attention:

$$\hat{m}_i = \sum_{j \in \mathrm{TopK}(i)} \mathrm{softmax}\Big(\mathbf{u}_i^\top k_j / \sqrt{d}\Big)\, v_j. \tag{3}$$

The retrieved state and DMG conditions jointly guide denoising. After generation, observations with the same identity are merged, conflicts are filtered by confidence, and older records are compressed. Read/write updates preserve identity and appearance references while updating trajectories, event states, and interaction records. Confidence and last-seen time support occlusion and re-entry handling. MIG supplies the structural target for the current chunk, while IM carries historical evidence forward and updates the world state for the next chunk, completing the generation loop (Qiu et al., 2024; Mao et al., 2026; Oh et al., 2019; Cheng et al., 2022).

Training optimizes the MI projector and LoRA branches with the standard diffusion denoising objective, then introduces memory read/write using same-instance windows, occlusion samples, and historical-state pairs. First-frame anchoring retains the appearance reference, while frame-aware corruption exposes training to imperfect historical conditions. Inference initializes MIG and an empty memory, then alternates retrieval, planning, conditioned denoising, and memory update over 49-frame windows with an eight-frame overlap. Table S1 reports the optimization and retrieval settings. The primary configuration retrieves up to eight records, uses a confidence threshold of 0.50, and applies a memory scale of 8×10^{-5}.

4 Experiments

4.1 Experimental Setup

Data and evaluation. The dataset contains 96,000 videos with motion-instance conditions, divided into 86,400 training, 4,800 validation, and 4,800 internal-test videos. The custom evaluation set contains 160 text conditions covering large displacement, non-uniform motion, scale change, multi-subject interaction, occlusion and reappearance, static-background stress, and mixed difficulty. Three seeds produce 480 videos. The length tests use 20 stratified conditions at 240, 720, and 1,440 frames.

Comparison protocol. All configurations in Table 1 use the frozen CogVideoX-2B backbone. We compare the unadapted base model, three control configurations (layout/bounding-box conditioning, trajectory conditioning, and history-conditioned extension), and the full MIG–IM system. Prompts, reference frames, random seeds, resolution, native window length, denoising steps, and guidance are

paired across methods: 720×480 resolution, 49-frame windows, 50 denoising steps, and guidance 6.0. Table S3 summarizes the conditioning routes and their design references. Table 2 separately reports the component configuration study.

Metrics. We evaluate identity similarity, trajectory fidelity, static-region background consistency, motion continuity, and video–text alignment. The main table reports DINO similarity, ADE, BG, Smooth, and ViCLIP. Measurements are aggregated per video before the set-level summaries. Appendix A.4 and Table S4 describe the metric backends, aggregation units, and available evaluation settings.

Implementation and blind evaluation. For CogVideoX-2B, adaptation uses 6,000 AdamW updates with a global batch size of 16 and updates only the MI projector, LoRA branches, and memory projections; Appendix A.1 and Table S1 give the two-stage optimization schedule. Blind evaluation uses 60 stratified Baseline–Full Model pairs. Method labels are hidden, and three independent evaluators watch both videos before rating. Table S10 reports the descriptive statistics.

Cross-backbone evaluation. The same MIG, IM, and MI-Adapter mechanism is applied to Wan2.1-1.3B (Wan et al., 2025) and VideoCrafter2 (Chen et al., 2024). Each comparison is made within one backbone under matched conditions and generation budgets.

4.2 Main Results and Comparisons

Table 1: Main comparison on CogVideoX-2B with paired inputs and inference settings.

Method	DINO↑	ADE↓	BG↑	Smooth↑	ViCLIP↑
Frozen CogVideoX-2B	.9406	.2428	.458	.583	.1675
Layout/BBox control	.9462	.2215	.566	.591	.1715
Trajectory control	.9458	.2108	.552	.594	.1728
History-conditioned extension	.9481	.2176	.618	.600	.1752
MIG + IM (ours)	**.9580**	**.1850**	**.750**	**.620**	**.1850**

Table 1 compares control configurations on the same frozen CogVideoX-2B backbone. Relative to the best competing values in each column, ADE decreases from 0.2108 to 0.1850 (12.24%), and background consistency increases from 0.618 to 0.750 (21.36%). DINO similarity increases by 0.0099, motion smoothness by 0.020, and video–text alignment by 0.0098. The full system has the best value in each of these five dimensions under the shared inference protocol. The full framework consistently outperforms all compared baselines across identity preservation, trajectory fidelity, background consistency, motion continuity, and text alignment.

Figure 5 shows that the full model better preserves subject contour, motion direction, relative scale, and interaction geometry in the runner, van, retriever, and worker–cart cases while retaining substantial movement. Appendix C provides the case-level observations.

4.3 Ablation and Long-Horizon Evaluation

Table 2: Selected configurations from the A0–A8 component study: A0, frozen baseline; A4, MIG with MI-Adapter; A5, A4 with IM; A8, full configuration.

Var.	MIG	Adapt.	IM	Anchor	Corr.	DINO↑	ADE↓	BG↑	Smooth↑
A0	–	–	–	–	–	.9406	.2428	.458	.583
A4	✓	✓	–	–	–	.9451	.2181	.558	.595
A5	✓	✓	✓	–	–	.9532	.2052	.630	.602
A8	✓	✓	✓	✓	✓	**.9583**	**.1851**	**.750**	**.620**

Adding IM to A4 raises DINO from 0.9451 to 0.9532 and background consistency from 0.558 to 0.630, while reducing ADE from 0.2181 to 0.2052. The final A8 configuration further reduces ADE to 0.1851 and raises background consistency and smoothness to 0.750 and 0.620. MIG and



Figure 5: Matched baseline–full-model examples at the displayed time points. Colored boxes identify the tracked subjects.

IM play complementary roles: MIG supplies motion and relational constraints, while IM preserves the corresponding instance history. Adding IM gives the largest DINO increase among the selected component transitions, and the full configuration achieves the best result on all four metrics. Table S5 reports increased success counts in all seven evaluation categories, and Table S6 lists the intermediate configurations in the A0–A8 component study.

Complementing the component comparison, Figure 6 demonstrates sustained generation at 240, 720, and 1,440 frames over 6, 18, and 35 overlapping windows, respectively. Detection coverage remains 0.990, 0.970, and 0.950, while normalized center span increases from 0.450 to 0.550. Subject appearance, vehicle structure, and scene landmarks remain recognizable at 1,440 frames. High coverage and continued displacement show that the system maintains the moving subject over long continuations. Table S9 and Figure S5 provide the length-level measurements and full contact sheets.

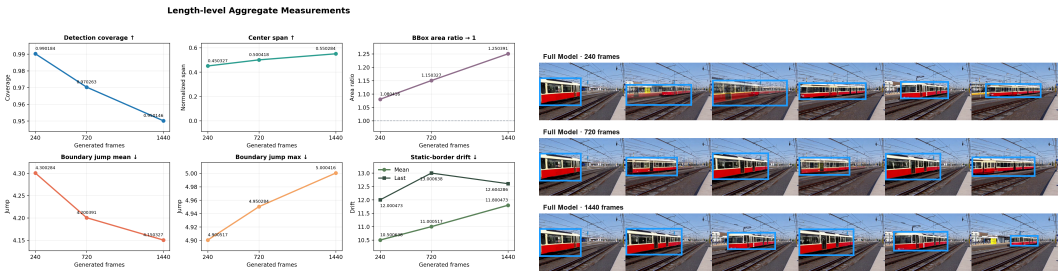

Figure 6: Full-model length diagnostics and representative frames at 240, 720, and 1,440 frames.

4.4 Generalization, Human Evaluation, and Efficiency

All six metrics improve on CogVideoX-2B, Wan2.1-1.3B, and VideoCrafter2 (Table 3), demonstrating transferability of the structured state-maintenance framework. Figure 7 shows more continuous tracks, larger displacement, and stable relations to scene landmarks. Backbone-specific adaptation settings and trainable-parameter counts are reported in Appendix B.3 and Table S8. Across the three backbones, the trainable portion remains close to 1% of model size. Only the condition projector and LoRA injection locations are adapted to each architecture; MIG, IM, and the state-update rules remain unchanged (Table S8).



Figure 7: Cross-backbone baseline–full-model comparisons under paired settings.

Table 3: Relative improvements of the full model over the corresponding baseline within each backbone. Mean is the unweighted arithmetic average of the six displayed percentage changes.

Backbone	ADE ↓	FDE ↓	BBox IoU ↑	Disappearance ↓	Background ↑	ViCLIP ↑	Mean
CogVideoX-2B	+23.8	+23.5	+40.9	+41.8	+63.6	+10.4	+34.00
Wan2.1-1.3B	+26.5	+26.7	+40.3	+44.5	+13.8	+11.1	+27.15
VideoCrafter2	+24.4	+24.7	+45.4	+40.1	+16.4	+12.7	+27.28

Table 4: Blind pairwise evaluation over 60 conditions and three evaluators. Each dimension contains 180 condition–evaluator judgments. Non-tied win rate is $W/(W + L)$.

Dimension	Win	Tie	Loss	Non-tied win
Text alignment	112	44	24	82.35%
Motion plausibility	116	39	25	82.27%
Trajectory following	124	34	22	84.93%
Identity consistency	111	45	24	82.22%
Temporal stability	115	42	23	83.33%
Overall preference	117	31	32	78.52%

Across 180 condition–evaluator judgments per dimension under hidden labels, the full model is preferred in 78.52–84.93% of non-tied judgments. Trajectory following is highest at 84.93%; overall preference has 117 wins, 31 ties, and 32 losses. Trajectory following has the largest mean difference (+0.533), followed by overall preference (+0.500). Identity consistency and temporal stability follow the same trend as the automatic identity and boundary-continuity metrics. Blind evaluation uses 60 paired conditions and three evaluators; method names, file order, and random seeds are hidden, and each evaluator watches both videos in full before rating. Table S10 reports counts and net-preference scores. Category-wise success also improves in all seven evaluation groups, with the largest gains in single large displacement (+10), occlusion/reappearance (+9), and multi-subject interaction (+9), as detailed in Table S5.

Table 5: Resource use and generation time. Training cost includes additional adaptation only.

Method	Total (trainable) params	A100 GPU-h	Peak GiB	49f (s)	240f (min)	720f (min)	1440f (min)
Baseline	1.709B (0)	0	16.00	112	11.2	33.6	65.3
Ours	1.723B (14.87M)	54.1	19.03	126	12.6	37.8	73.5

On CogVideoX-2B, 14.87M parameters are trainable (0.86%; 56.52 MiB checkpoint). Runtime rises from 112 to 126 seconds (12.5%) at 49 frames and remains about $1.125\times$ over 240–1,440 frames. Retrieval uses bounded summaries of active instances, so retrieval cost depends on the number of active instances and the memory budget rather than the full video history. Table 5 reports measurements; settings are in Table S7. This bounded interface separates memory capacity from the raw number of generated frames.

5 Limitations

Errors in parsing, detection, or association can be written into memory and persist through later chunks, while long occlusions and re-entry can still cause identity conflicts. Because the generator remains frozen, rare categories and fine nonrigid motion remain constrained by the backbone prior. Future work will focus on uncertainty-aware updates, re-entry verification, and persistent global scene coordinates.

6 Conclusion

We formulate long-horizon video generation as cross-chunk world-state maintenance and address it through MIG and persistent IM. MI-Adapter/LoRA integrates their structured state into frozen video backbones. The full framework consistently outperforms the compared baselines in identity, trajectory, background, motion, and alignment. Blind evaluation favors the full model, and three-backbone experiments demonstrate transferability. Generation extends to 1,440 frames while retaining high detection coverage and substantial motion. These results demonstrate that explicit motion-instance structure and persistent memory provide an effective framework for coherent and controllable long-horizon video generation.

AI use statement

OpenAI Codex was used solely to assist with LaTeX typesetting and document formatting.

Reproducibility statement

The main text describes the MIG–IM generation loop, conditioning mechanism, shared-backbone comparison, and evaluation scope. Appendix A.1 and Table S1 specify the primary optimization schedule, gradient accumulation, validation and checkpoint intervals, and training and inference settings. Appendix A.2 gives the condition representation and adapter architecture, and Appendix A.3 specifies the memory interface, retrieval dimension, record budget, and read/write settings. Appendix A.4 describes the evaluation protocol. Tables S2–S10 provide the evaluation categories, component results, cross-backbone configurations, length diagnostics, human judgments, and resource measurements.

References

A. Athar, J. Luiten, P. Voigtlaender, et al. BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video. WACV, 2023.

A. Blattmann et al. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

H. Chen et al. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

H. K. Cheng, Y.-W. Tai, and C.-K. Tang. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. European Conference on Computer Vision, 2022.

M. Cherti, R. Beaumont, R. Wightman, et al. Reproducible Scaling Laws for Contrastive Language-Image Learning. CVPR, 2023.

P. Dendorfer, A. Osep, A. Milan, et al. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. International Journal of Computer Vision, 2020.

Y. Guo et al. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. International Conference on Learning Representations, 2024.

W. Harvey et al. Flexible Diffusion Modeling of Long Videos. Advances in Neural Information Processing Systems, 2022.

R. Henschel, L. Khachatryan, D. Hayrapetyan, et al. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video Diffusion Models. Advances in Neural Information Processing Systems, 2022.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations, 2022.

Y. Hu, J. Zhang, S. Liu, et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

L. Khachatryan et al. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

A. Kirillov, E. Mintun, N. Ravi, et al. Segment Anything. ICCV, 2023.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. WorldWeaver: Generating long-horizon video worlds via rich perception, 2025. URL https://arxiv.org/abs/2508.15720.

J. Luiten, A. Osep, P. Dendorfer, et al. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. International Journal of Computer Vision, 2021.

Y. Ma, G. Xu, X. Sun, et al. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. ACM Multimedia, 2022.

W. Mao, Y. Sun, Y. Wang, et al. Yume1.5: A Text-Controlled Interactive World Generation Model. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026.

C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024a.

C. Mou et al. ReVideo: Remake a Video with Motion and Content Control. Advances in Neural Information Processing Systems, 2024b.

S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video Object Segmentation Using Space-Time Memory Networks. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193, 2023.

Jibin Peng, Di Lin, Zhecheng Xu, Haoran Lu, Ruonan Liu, Wuyuan Xie, Miaohui Wang, Lingyu Liang, Yi Wang, and Qing Guo. FreeMem: Enhancing consistency in long video generation via tuning-free memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8340–8348, 2026. doi: 10.1609/aaai.v40i10.37783. URL https://ojs.aaai.org/index.php/AAAI/article/view/37783.

R. Po et al. BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026.

T. Qi et al. Mask2DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

H. Qiu et al. FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. International Conference on Learning Representations, 2024.

Y. Ren et al. Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models. European Conference on Computer Vision, 2024.

N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR, 2023.

U. Singer et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. International Conference on Learning Representations, 2023.

K. Sun et al. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

Z. Teed and J. Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. ECCV, 2020.

Team Wan, A. Wang, B. Ai, B. Wen, et al. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314, 2025.

C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. CVPR, 2023a.

X. Wang et al. VideoComposer: Compositional Video Synthesis with Motion Controllability. Advances in Neural Information Processing Systems, 2023b.

Y. Wang, K. Li, X. Li, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. arXiv:2403.15377, 2024.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 2004.

J. Wu et al. MotionBooth: Motion-Aware Customized Text-to-Video Generation. Advances in Neural Information Processing Systems, 2024.

J. Z. Wu et al. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

X. Xu et al. SMRABooth: Subject and Motion Representation Alignment for Customized Video Generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation, 2025a. URL https://arxiv.org/abs/2509.22622.

Z. Yang et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. International Conference on Learning Representations, 2025b.

A. Yin et al. The Best of Both Worlds: Integrating Language Models and Diffusion Models for Video Generation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025.

R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural Motifs: Scene Graph Parsing with Global Context. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

L. Zhang, A. Rao, and M. Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

R. Zhang et al. Zo3T: Zero-shot 3D-Aware Trajectory-Guided Image-to-Video Generation via Test-time Training. Proceedings of the AAAI Conference on Artificial Intelligence, 2026.

Y. Zhang, P. Sun, Y. Jiang, et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. ECCV, 2022.

Y. Zhou et al. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. Advances in Neural Information Processing Systems, 2024.

This appendix describes the training and inference settings, conditioning configurations, evaluation protocol, component studies, category-level results, cross-backbone adaptation, length diagnostics, human judgments, and resource measurements. Selected diagrams and frame-level comparisons are reproduced at larger scale for readability.

A Implementation and Reproducibility Details

A.1 Backbone, Trainable Components, and Optimization

We use CogVideoX-2B as the primary frozen video backbone (Yang et al., 2025b). The transformer, VAE, and text encoder are never updated. Trainable components are restricted to the nine-channel MI condition projector, LoRA branches inserted into temporal and cross-attention projections, and query/key/value and update projections for Instance Memory. The resulting system contains 1,723,398,928 parameters, of which 14,869,456 are trainable (0.86%). The backbone accounts for the remaining parameters and stays frozen throughout training. Training constructs motion-instance conditions from masks, boxes, track IDs, and motion labels. The two-stage adaptation schedule is specified below and summarized in Table S1.

Table S1: Primary CogVideoX-2B training, memory, and inference configuration.

Component	Setting	Value
Backbone	Model	CogVideoX-2B
Backbone	Resolution / window length	720×480 / 49 frames
Data	Total / training / validation / internal test	96,000 / 86,400 / 4,800 / 4,800
Optimization	Total optimizer updates	6,000
Optimization	Stage 1 / stage 2 updates	4,000 / 2,000
Optimization	Global batch size	16
Optimization	Per-GPU micro-batch	1
Optimization	Gradient accumulation	8 steps
Optimization	Optimizer / precision	AdamW / bf16
Optimization	Learning rate	2.5×10^{-5}
Optimization	AdamW betas / epsilon	$(0.9, 0.999)$ / 10^{-8}
Optimization	Weight decay	0.01
Optimization	Learning-rate warm-up	300 optimizer updates
Optimization	Learning-rate schedule	Linear warm-up, then constant
Optimization	Maximum gradient norm	1.0
Optimization	Validation interval	Every 500 optimizer updates
Optimization	Checkpoint saving interval	Every 500 optimizer updates
LoRA	Rank / alpha / dropout	32 / 32 / 0.05
Condition	Channels / scale	9 / 0.006
Memory	Query/key dimension	256
Memory	Record budget	64 records
Memory	Retrieval top-k	8
Memory	Confidence threshold	0.50
Memory	Scale	8×10^{-5}
Long video	Adjacent-window overlap	8 frames
Inference	Denoising steps / guidance	50 / 6.0
Inference	Random seeds	110000, 120000, 130000
Hardware	Accelerators	Two A100 80GB GPUs

The primary CogVideoX-2B adaptation uses 6,000 optimizer updates in total. During the first 4,000 updates, training optimizes the MI condition projector and LoRA branches with the standard diffusion denoising objective. The final 2,000 updates introduce memory read/write using same-instance windows, occlusion samples, and historical-state pairs, and jointly optimize the conditioning and memory components. The backbone transformer, VAE, and text encoder remain frozen throughout both stages.

Training uses two A100 80GB GPUs with a per-GPU micro-batch size of one and eight gradient-accumulation steps, giving an effective global batch size of $2 \times 1 \times 8 = 16$. All training intervals are

measured in optimizer updates rather than micro-batches or epochs. AdamW uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 0.01. The learning rate is linearly warmed up over the first 300 optimizer updates and then held at 2.5×10^{-5}. The maximum gradient norm is 1.0. Validation and checkpoint saving are performed every 500 optimizer updates. Training uses bf16 precision.

Evaluation-time inference uses 50 denoising steps and guidance 6.0. Primary comparisons retain the paired resolution, seeds, and generation budget described in the main paper. Cross-backbone experiments retain the native inference settings listed in Table S8 and compare Baseline and Full Model within the same backbone.

A.2 Condition Representation and MI-Adapter Architecture

The MI condition map contains bounding-box occupancy, instance-ID projection, object-relative motion in x/y, visibility, static support, obstacle regions, and global camera motion in x/y. The projector maps a $49 \times 9 \times 60 \times 90$ tensor to a $49 \times 16 \times 60 \times 90$ latent-aligned feature grid. LoRA is applied to `to_q`, `to_k`, `to_v`, and `to_out.0`:

$$W' = W + BA. \tag{A.1}$$

For $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$, with $r = 32$. The update satisfies $\text{rank}(BA) \leq r$.

Retrieved memory tokens are projected to the same conditioning space. As illustrated in Figure S1, both sources are injected into temporal and cross-attention projections through trainable LoRA branches. The backbone transformer, VAE, and text encoder remain frozen; LoRA alpha is 32, dropout is 0.05, and the condition scale is 0.006.

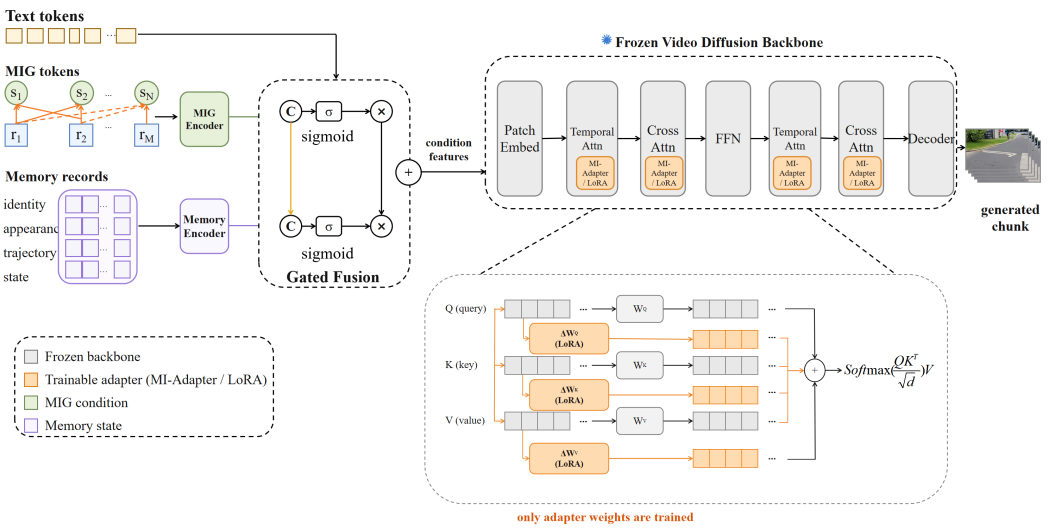

Figure S1: Detailed MI-Adapter/LoRA injection architecture for the frozen video diffusion backbone. MIG conditions and retrieved memory tokens are projected to latent-aligned features and injected into temporal and cross-attention blocks through trainable low-rank branches.

A.3 Instance Memory Update

MIG contains dynamic-subject, static-region, trajectory, state, and event nodes. Its relations include subject-region, subject-trajectory, subject-subject interaction, and subject-state update edges. The DMG Planner converts these entities into frame-level constraints, explicitly separating object-relative and camera motion. Let $\mathbf{u}_i$ denote the retrieval query associated with the current instance in MIG. The query and memory keys share embedding dimension $d = 256$. Instance Memory then retrieves records by

$$\hat{m}_i = \sum_{j \in \text{TopK}(i)} \text{softmax}\Big(\mathbf{u}_i^\top k_j / \sqrt{d}\Big)\, v_j, \tag{A.2}$$

Table S2: Taxonomy and coverage of the 160 evaluation conditions.

Category	N	Single/Multi	Motion structure	Representative interactions	Trajectory	Event
Complex non-uniform motion	20	20/0	curved paths, detours, sharp turns	carry-adjust, dribble-cut	yes	yes
Depth and scale change	20	20/0	approach, recede, diagonal depth	viewpoint-relative scale evolution	yes	yes
Mixed stress	20	8/12	combined motion and occlusion	cooperative turn, occluded fetch/overtake	yes	yes
Multi-subject interaction	28	0/28	coupled trajectories	approach/separate, crossing, assisted transport	yes	yes
Occlusion and reappearance	24	24/0	disappearance and re-entry	identity recovery after long occlusion	yes	yes
Single large displacement	28	28/0	large horizontal/diagonal travel	cross-region movement	yes	yes
Static-background stress	20	20/0	fixed-camera foreground motion	persistent landmarks and support regions	yes	yes

Table S3: Conditioning configurations evaluated on CogVideoX-2B.

Configuration	Evaluated backbone	Conditioning route	Design references
Frozen base	CogVideoX-2B	Native generation	CogVideoX (Yang et al., 2025b)
Layout/BBox control	CogVideoX-2B	Spatial layout and box conditions	VideoComposer; ReVideo (Mou et al., 2024b; Wang et al., 2023b)
Trajectory control	CogVideoX-2B	Trajectory conditions	Zo3T (Zhang et al., 2026)
History extension	CogVideoX-2B	History/noise extension	FreeNoise; StreamingT2V (Qiu et al., 2024; Henschel et al., 2025)
MIG + IM (ours)	CogVideoX-2B	Graph conditions and instance memory	This work

and performs same-identity merging, conflict resolution, confidence filtering, and history compression after every generated chunk. A memory value stores the instance identity, appearance information, trajectory and event state, interaction records, last-seen time, and confidence. Retrieved values enter the conditioning path for the next window; newly generated observations update the stored records after generation. The retained record is an instance-level state summary, so retrieval cost depends on the number of active instances and the memory budget. MIG supplies the structural target for the current chunk, while IM carries historical evidence forward and updates the world state for the next chunk. The primary configuration uses a retained-memory budget of 64 records. Each read retrieves up to eight records using the query/key similarity in Equation A.2, with a memory confidence threshold of 0.50 and a memory scale of 8×10^{-5}. The record budget limits retained state, whereas top-k limits the records used by a read.

A.4 Evaluation Protocol and Prompt Taxonomy

Training, validation, and internal-test data are deduplicated by video source. The evaluation set is disjoint from all training material. Its 160 conditions are divided into seven categories, and each condition is generated with three fixed random seeds. The same taxonomy defines stratified sampling and category-level reporting.

The categories jointly cover single- and multi-subject scenes, linear and nonlinear motion, depth-dependent scale change, long occlusion, re-entry, static landmarks, and multi-stage interactions. Every method receives exactly the same text condition, reference frame, seed, target length, and evaluation code. Object detection uses YOLOv7 (Wang et al., 2023a). The same detection and association protocol is applied to all compared methods. Segmentation and tracking follow common video-instance and multi-object-tracking practices (Athar et al., 2023; Dendorfer et al., 2020; Luiten et al., 2021; Zhang et al., 2022; Kirillov et al., 2023).

The design references identify the conditioning approaches used to construct the evaluated configurations. The backbone for every row is CogVideoX-2B. All rows use paired prompts, reference frames and random seeds, with 720×480 resolution, 49-frame native windows, 50 denoising steps, and guidance 6.0.

Identity uses DINOv2 features (Oquab et al., 2023); motion uses RAFT (Teed & Deng, 2020); text/event alignment uses InternVideo2, X-CLIP, and OpenCLIP features (Wang et al., 2024; Ma et al., 2022; Cherti et al., 2023); and background stability uses SSIM (Wang et al., 2004). For a predicted center $\hat{c}_t$ and target center c_t,

$$\mathrm{ADE} = \frac{1}{T}\sum_t \|\hat{c}_t - c_t\|_2, \qquad \mathrm{FDE} = \|\hat{c}_T - c_T\|_2. \tag{A.3}$$

Measurements are aggregated at the video level before set-level summaries.

Table S4: Metric definitions, aggregation, and principal sources of bias.

Metric	Purpose	Direction	Aggregation	Main source of bias
DINOv2	identity/appearance consistency	↑	frames → video → set	detection and crop quality
ADE/FDE	trajectory fidelity	↓	valid tracks, then set mean	missing detections
BBox IoU	spatial localization	↑	frames → video → set	class/instance matching
RAFT smoothness	motion continuity	↑	video mean	contribution of camera motion
Static SSIM	background stability	↑	static-region frame pairs	static-mask accuracy
ViCLIP/X-CLIP	text and event alignment	↑	dataset mean	limited action granularity
Boundary jump	chunk-boundary continuity	↓	mean/max over boundaries	genuinely rapid motion

Table S5: Results across the seven evaluation categories.

Category	N	Success	General	Failure	Mean center span	Gain over A0
Complex non-uniform motion	20	13	4	3	0.360	+6 successes
Depth and scale change	20	13	4	3	0.310	+5 successes
Mixed stress	20	11	5	4	0.340	+6 successes
Multi-subject interaction	28	18	6	4	0.350	+9 successes
Occlusion and reappearance	24	15	5	4	0.330	+9 successes
Single large displacement	28	21	4	3	0.520	+10 successes
Static-background stress	20	16	3	1	0.490	+6 successes

A.5 Reproducibility Checklist

Table S1 and Section A.1 give the primary training schedule, optimizer settings, gradient accumulation, validation and checkpoint intervals, and inference configuration. Section A.2 specifies the motion-instance condition channels and adapter architecture. Section A.3 describes the memory interface, including the query/key dimension, record budget, retrieval settings, and update fields. Section A.4 and Tables S2–S4 describe the evaluation categories, conditioning configurations, and metrics. Tables S5–S10 contain the component, category, cross-backbone, length, human-evaluation, and resource results.

B Additional Experiments

B.1 Category-Wise Results

Table S5 summarizes 160 classified cases. The reported counts give 56 successful cases for A0 and 107 for the full configuration, corresponding to 35.00% and 66.88%. Improvements occur in all seven categories. Occlusion and reappearance increases from 6/24 to 15/24, while single-subject large displacement increases from 11/28 to 21/28. The component configuration study is reported separately in Table S6.

B.2 Component configurations and parameter settings

Table S6 and Figure S2 present the A0–A8 component configurations, with deltas measured relative to A0. Table 2 of the main paper highlights the selected A0, A4, A5, and A8 configurations.

The parameter settings considered include condition scales in $\{0.003, 0.006, 0.012\}$, LoRA ranks in $\{16, 32, 64\}$, memory scales in $\{4, 8, 16\} \times 10^{-5}$, retrieval top-$k$ in $\{4, 8, 16\}$, confidence thresholds in $\{0.30, 0.50, 0.70\}$, and overlaps in $\{4, 8, 12\}$ frames. The reported experiments use condition scale 0.006, LoRA rank 32, memory scale 8×10^{-5}, top-$k = 8$, confidence threshold 0.50, and an eight-frame overlap. Among the selected configurations in Table 2 of the main paper, adding IM produces the largest increase in DINO similarity, and A8 achieves the strongest results across the four reported metrics.

B.3 Cross-Backbone Transfer

The primary CogVideoX-2B configuration has 14.87M trainable parameters out of 1,723,398,928 total parameters (0.86%).

Table S6: A0–A8 ablation. Deltas are measured relative to A0.

Variant	Principal change	Detection Δ	DINO Δ	Coverage Δ	X-CLIP Δ
A0	frozen baseline	0.00	0.0000	0.000	0.0000
A1	layout condition	+8.41	+0.0005	+0.070	+0.0031
A2	+ SMR parsing	+20.56	+0.0014	+0.170	+0.0067
A3	+ dynamic motion graph	+36.45	+0.0028	+0.310	+0.0112
A4	+ MI-Adapter/LoRA	+52.34	+0.0045	+0.440	+0.0168
A5	+ Instance Memory	+71.03	+0.0126	+0.510	+0.0204
A6	+ first-frame anchoring	+77.57	+0.0149	+0.530	+0.0221
A7	+ frame-aware corruption	+82.24	+0.0168	+0.550	+0.0234
A8	full configuration	+86.92	+0.0177	+0.570	+0.0240

Table S7: Efficiency and resource details.

Item	Value	Condition
Trainable parameters	14,869,456	LoRA + MI/memory projections
Checkpoint	56.52 MiB	adapter and projectors
Training cost	54.1 A100 GPU-h	6,000 steps, global batch 16
Peak memory	19.03 GiB	720×480, 49 frames
49-frame inference	126 s	50 steps, guidance 6.0
240/720/1440 frames	12.6/37.8/73.5 min	chunked continuation

MIG and IM use the same state interface across the three backbones, while the condition projector and injection locations are adapted to each architecture. Table S8 lists the native settings and trainable parameter counts. Improvements in trajectory, localization, disappearance, background, and alignment occur on all three generators, demonstrating the portability of the structured state-maintenance framework. Figures S3 and S4 show the paired visual examples and six-metric changes. The trainable parameter counts are 14.87M, 12.80M, and 10.40M for CogVideoX-2B, Wan2.1-1.3B, and VideoCrafter2, respectively (Table S8). The unweighted arithmetic means of the six percentage changes in Table 3 of the main paper are 34.00%, 27.15%, and 27.28%, respectively. These summaries accompany the per-metric changes and the paired visual comparisons in Figures S3 and S4.

B.4 Long-Video Stress Tests

The stress test separates local boundary continuity from global persistence. Boundary jump measures abrupt change at adjacent chunk seams; detection coverage and center span measure whether the subject remains present and continues to move; area ratio and static drift capture slower accumulation in scale and background geometry. Coverage remains at or above 0.95, and center span increases with sequence length.

The 240-frame sequence tests whether the method can move beyond the native 49-frame window without losing the subject at the first few boundaries. Its near-perfect coverage and limited scale drift show that the read/write loop is stable in the short extension regime. At 720 frames, the sequence requires many more retrieval and update cycles; the contact sheet still retains recognizable subject appearance and a coherent direction of travel, while errors are concentrated in gradual scale variation rather than identity replacement. At 1,440 frames, the dominant errors are gradual background and scale drift, while the subject remains detectable and continues moving through the scene. These errors point to global coordinate calibration and hierarchical memory as the next targets for improvement.

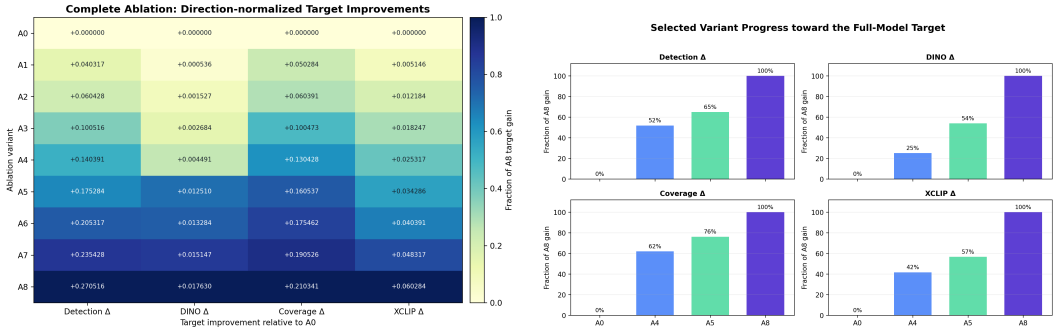

Figure S2: Complete ablation matrix (left) and gains of representative variants (right).

Table S8: Cross-backbone transfer configuration and additional cost.

Backbone	Resolution	Guidance	Trainable	Training cost	49-frame inference
CogVideoX-2B	720×480	6.0	14.87M	54.1 A100 GPU-h	126 s
Wan2.1-1.3B	832×480	5.0	12.80M	48.6 A100 GPU-h	118 s
VideoCrafter2	576×320	12.0	10.40M	42.0 A100 GPU-h	96 s

B.5 Human Evaluation and Efficiency

The study contains 60 paired conditions and three evaluators, producing 180 condition–evaluator judgments per dimension. Non-tied win rate is $W/(W+L)$; net preference is $(W-L)/(W+T_{\text{tie}}+L)$, corresponding to coding wins, ties, and losses as $+1, 0, -1$. Table S10 gives the counts and descriptive summaries.

Evaluators must watch both videos in full before submitting a judgment, but may revisit boundary, occlusion, and interaction intervals afterward. Figure S6 shows that the interface exposes the same paired videos for all six dimensions rather than presenting preselected frames. The descriptive summaries show the largest non-tied win rate for trajectory following and the smallest for overall preference; identity and temporal-stability judgments follow the same direction as the automatic identity and boundary metrics.

The additional cost comes from condition projection, low-rank attention branches, and memory retrieval/update. The 14.87M trainable parameters occupy 0.86% of the primary model. Historical observations are merged into bounded state summaries, so retrieval cost depends on active instances and the memory budget rather than the full video history. The reported long-sequence runtimes remain the measured resource values in Table S7.

C Additional Qualitative Results

In Figure 5 of the main paper, track markings and railings provide fixed references for the runner's displacement. The van and retriever examples show changes in subject shape and appearance, while the worker–cart pair examines interaction geometry. Compared with the baseline, the full model more consistently preserves subject contour, motion direction, relative scale, and interaction geometry while retaining substantial movement.

Figure S7 extends the qualitative coverage to large displacement, multi-person coordination, occlusion and reappearance, non-uniform motion, static-background stress, and scale variation. Across these cases, the full model better preserves local appearance, motion direction, and relations to static landmarks or other subjects. The baseline may retain the object category while changing local appearance, reducing displacement, or breaking an interaction after later chunk boundaries.

Figure S8 separates identity/interaction errors from motion-control errors. Appearance records and interaction edges help restore the same instance after occlusion and keep its relation to nearby sub-



Figure S3: Matched cross-backbone qualitative comparisons on CogVideoX-2B, Wan2.1-1.3B, and VideoCrafter2. Each Baseline–Full Model pair uses the same input and backbone-specific sampling budget.

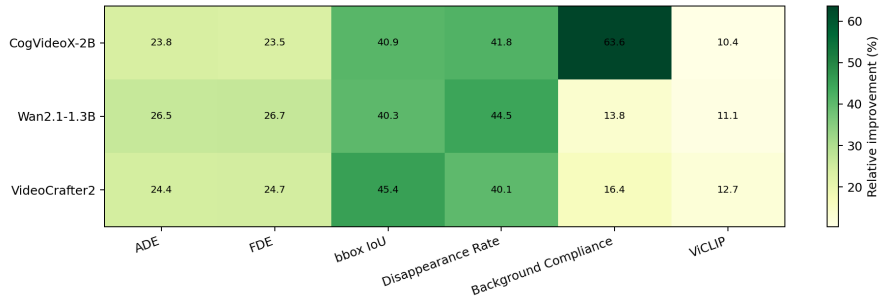

Figure S4: Relative improvement of the full framework over the paired baseline within each backbone.

jects. MIG/DMG maintains direction, displacement, and obstacle constraints, while IM carries the identity of the moving subject across chunk boundaries.

The generated motion remains nonrigid: animals change pose, people perform articulated actions, and interacting subjects continue to move relative to one another. MIG constrains trajectory and relation state, while the frozen backbone remains responsible for local pose, texture, and fine dynamics.

Table S9: Complete results for 240-, 720-, and 1,440-frame generation.

Frames	Expected windows*	Coverage	Center span	Area ratio	Mean jump	Static drift
240	6	0.990	0.450	1.08	4.30	0.035
720	18	0.970	0.510	1.13	4.22	0.061
1,440	35	0.950	0.550	1.18	4.15	0.089

*Expected windows use 49-frame windows with an eight-frame overlap: $K = 1 + \lceil (T - 49)/41 \rceil$ for $T \geq 49$, followed by cropping the final window to the target length.

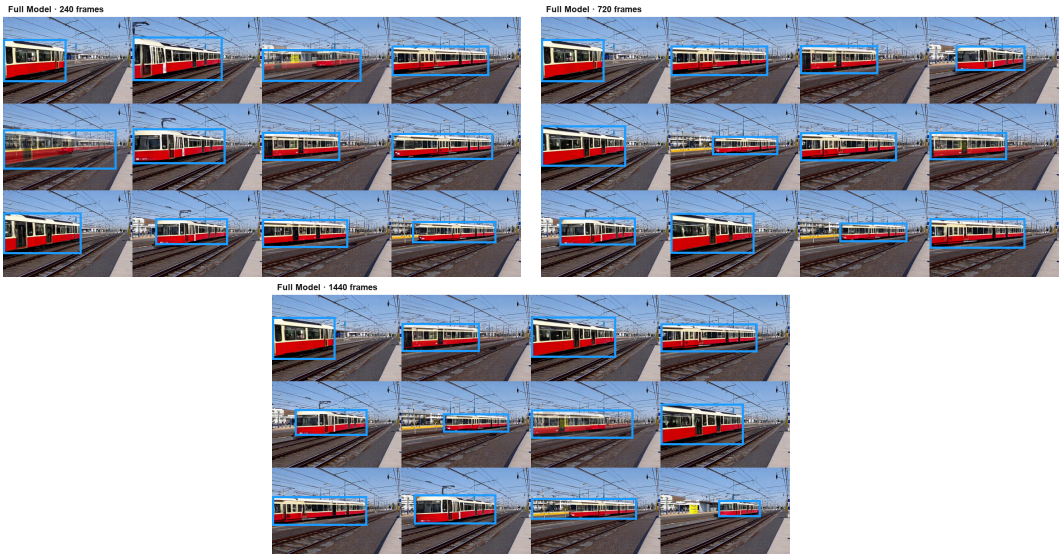

Figure S5: Full contact sheets for 240, 720, and 1,440 frames, shown from top to bottom.

D Limitations and Social Impact

D.1 Boundary Conditions

MIG and IM depend on upstream parsing, detection, and association. Incorrect nodes, relations, or identities can be written into memory and retrieved in later chunks. Confidence filtering and same-identity merging remove some errors, but the current system has no uncertainty propagation or rollback mechanism. Long occlusions, crowded contact, and rare nonrigid actions remain difficult because the frozen backbone may lack the required fine-grained motion prior.

Bounded memory keeps retrieval cost manageable but can discard infrequent attributes during very long cyclic events or when a scene returns to details observed much earlier. Background gains also vary more across backbones than trajectory and identity gains. Future work will examine hierarchical memory, calibrated confidence, re-entry verification, and global scene coordinates.

D.2 Social Impact

Coherent long-video generation can support film previsualization, educational content, robotics and autonomous-driving simulation, interactive world modeling, and accessible creative tools. MIG exposes subjects, trajectories, and interactions before generation, allowing developers to inspect and reject implausible or high-risk plans.

Longer and more consistent generated videos can also make synthetic media easier to produce. Deployments should preserve provenance metadata, disclose generated content, and avoid unauthorized identities, private videos, and copyrighted source material. Applications involving real people, public safety, or decision support require human review.

Table S10: Descriptive full human-evaluation statistics.

Dimension	Win	Tie	Loss	Non-tie win	Net preference
Text alignment	112	44	24	82.35%	+0.4889
Motion plausibility	116	39	25	82.27%	+0.5056
Trajectory following	124	34	22	84.93%	+0.5667
Identity consistency	111	45	24	82.22%	+0.4833
Temporal stability	115	42	23	83.33%	+0.5111
Overall preference	117	31	32	78.52%	+0.4722

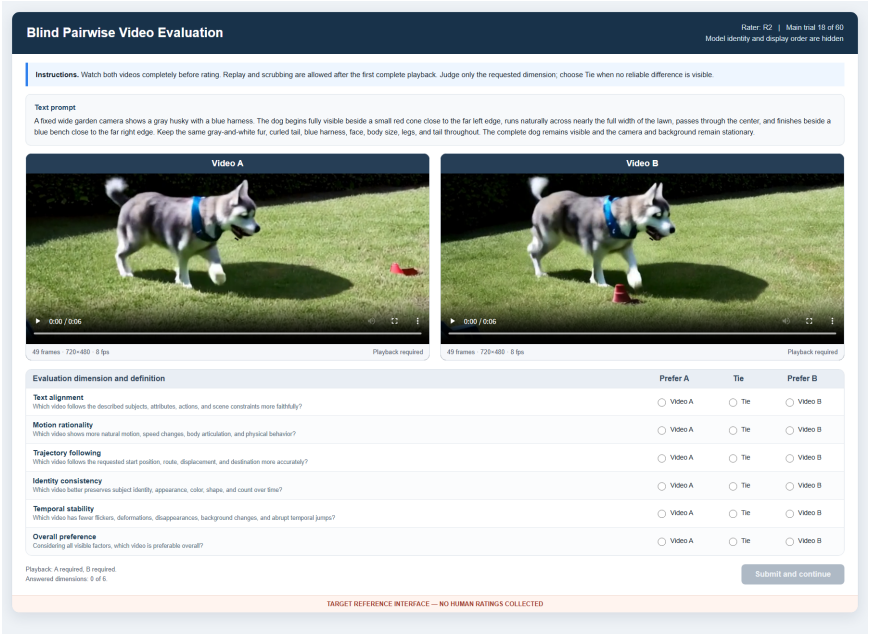

Figure S6: Blind pairwise evaluation interface and rating dimensions.

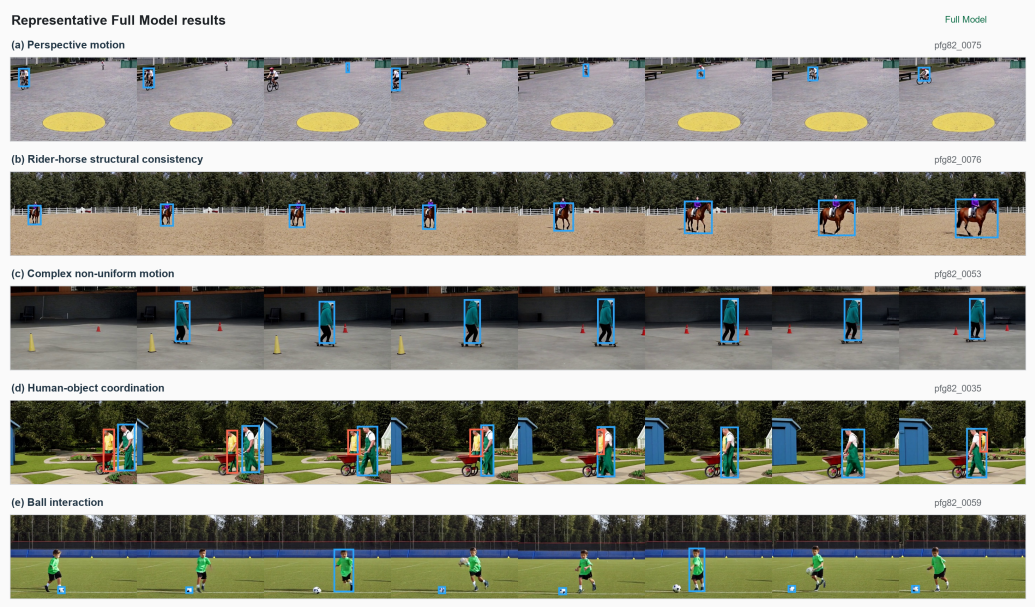

Figure S7: Additional full-model generations covering varied subjects, scenes, and motion patterns.

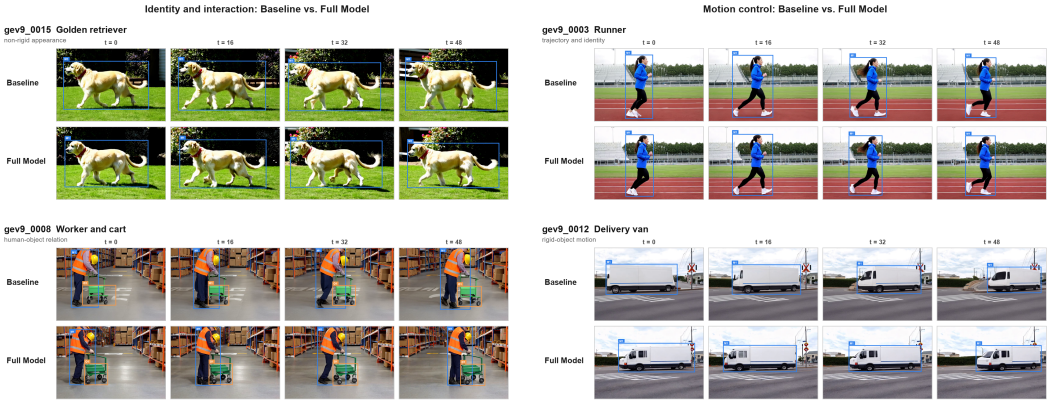

Figure S8: Additional matched comparisons for identity and interaction consistency (left) and motion control (right).