VISARAG: RETRIEVAL-AUGMENTED DYNAMICS MEMORY FOR FUTURE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Future video generation is ambiguous: visually similar initial states can evolve toward different outcomes, even under the same broad action description. Recent retrieval-augmented methods improve motion realism, physical plausibility, and temporal consistency, but using retrieved experiences to distinguish between such outcomes remains less explored. We propose VisaRAG, a framework that uses observed action-outcome transitions as external dynamics memory for generating video continuations. Given an observed video prefix, VisaRAG retrieves relevant experiences, encodes their temporal changes while suppressing static appearance, and injects the resulting memories through lightweight motion branches into a frozen video generator. A counterfactual learning objective trains the memory-conditioning components to favor memories consistent with the target outcome over alternatives, with the target video and all other generation conditions held fixed. To measure whether retrieved dynamics steer the generated outcome, we further introduce VisaBench, a benchmark that pairs visually similar interactions leading to different outcomes with same-outcome controls, and that conditions every sample on shared, outcome-neutral text so the prompt cannot reveal the target outcome. Across multiple benchmarks, VisaRAG consistently improves both action correctness and the quality of the resulting motion, demonstrating more accurate and realistic future video generation.

1 INTRODUCTION

Recent video generation models have achieved remarkable progress in visual fidelity, temporal coherence, and motion synthesis Blattmann et al. (2023); Kondratyuk et al. (2024); Yang et al. (2025); Wan et al. (2025). Yet, generating plausible motion is fundamentally different from predicting *what happens next*. Given the same observed state, multiple continuations may appear reasonable, while the actual future depends on the action, object interactions, and their evolving states. An object pushed near the edge of a table, for instance, may move slightly, slide across the surface, or fall. Although large-scale pretraining provides strong motion priors, a generator cannot explicitly preserve every action–outcome transition that may be needed at inference time. Future generation therefore requires not only modeling *how things can move*, but determining *which future should unfold*.

Retrieval provides a natural way to complement such parametric priors with observed experience. MotionRAG Zhu et al. (2025) retrieves reference videos to transfer motion priors, while PhysRAG Cheng et al. (2026) retrieves physical references to improve physics-aware generation. As illustrated in Figure 1, these approaches primarily use retrieval to improve *how* a specified event unfolds, whereas VisaRAG treats retrieved videos as outcome-specific evidence for *which future should unfold*. We therefore ask: **can retrieved real-world experiences provide outcome-specific evidence for future video generation?**

Our key insight is that real videos provide more than visual appearance or generic motion patterns: they record *how an observed state evolves under an action and which outcome eventually occurs*. For example,

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.

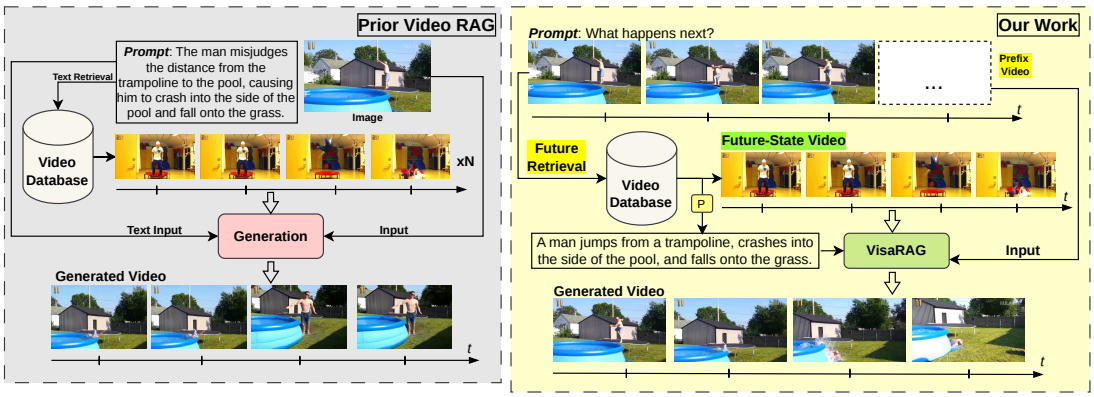


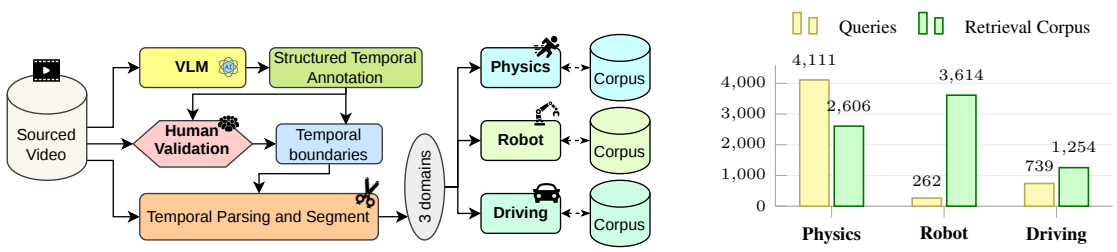
Figure 1: Retrieval as Motion Prior vs. Future Evidence. Prior video RAG retrieves how an event should unfold, while VisaRAG retrieves observed experiences to guide which future should unfold.

a video of an object being pushed and falling from a table captures a concrete transition from the initial state, through the action, to the resulting future state. We therefore view a video corpus as a non-parametric memory of observed futures, where each video represents an experience of how a particular situation actually unfolded. At inference time, retrieving experiences that resemble the current state and action can provide explicit evidence about plausible future dynamics, complementing the generic priors encoded in the generator's parameters.

To study this problem, we introduce **VisaBench**, a benchmark constructed from real action-centric videos. VisaBench contains *minimal outcome pairs* with similar visual context and actions but different resulting states, together with same-outcome controls. This construction explicitly tests whether retrieval captures outcome-specific dynamics rather than appearance or semantic shortcuts. We further propose **VisaRAG**, a retrieval-augmented framework for future video generation. VisaRAG retrieves relevant video experiences and represents their temporal evolution as compact future-dynamics memories, suppressing static appearance while retaining state changes. These memories are injected through lightweight motion branches into a frozen video generator. To ensure that the generator actually uses the retrieved experience, we introduce counterfactual memory learning: matching and mismatching outcome memories are contrasted while the target, anchor, text condition, diffusion timestep, and noise remain identical. Thus, the learning signal directly reflects whether the retrieved dynamics explain the target future.

Our main contributions are:

- We introduce VisaBench, a benchmark of minimal outcome pairs and same-outcome controls for evaluating outcome-specific future dynamics.
- We propose VisaRAG, which retrieves real-world experiences, extracts future-dynamics memories, and injects them into frozen video generators through counterfactual memory learning.
- VisaRAG improves both action correctness and motion quality across benchmarks, with memory-swapping experiments demonstrating that retrieved experiences systematically steer future generation.



(a) Benchmark construction.

(b) Query–corpus composition.

Figure 2: **VisaBench overview.** (a) Real videos are segmented and validated to form controlled action–outcome transitions. (b) Query sets and retrieval corpora span physics, robotics, and driving, with the corresponding data sources and video counts.

2 Related Works

We summarize the most relevant literature below and provide an extended discussion and detailed comparison with prior work in Appendix A.

Video Retrieval and Retrieval-Augmented Understanding. Modern video retrieval increasingly uses multimodal representations to retrieve semantically relevant visual evidence Wang et al. (2025). Recent work extends retrieval to video reasoning Jeong et al. (2025) and composed retrieval, where videos are retrieved according to desired visual changes Liu et al. (2026); Li et al. (2026). These methods primarily use retrieval as an output or as evidence for understanding. In contrast, VisaRAG treats retrieved videos as *experience memory*: observed transitions are retrieved to guide how a new observation evolves into its future.

Retrieval-Augmented Video Generation. Retrieval has also been used to augment video generation with external motion, physical, or temporal priors. MotionRAG Zhu et al. (2025) retrieves transferable motion references, PhysRAG Cheng et al. (2026) introduces physical references, and LongLive-RAG Hu et al. (2026) retrieves previous states for long-video generation. Related trajectory retrieval has also been explored for robot action generation Park et al. (2026). VisaRAG instead studies retrieval for *future selection*: whether observed action–outcome experiences can provide evidence about which future should unfold, while also improving how that future is generated.

3 The VisaBench Benchmark

3.1 Motivation and Task Definition

VisaBench studies whether past video experience can help predict *what happens next* in a new observation. This setting is motivated by robotics and embodied prediction, where models must generalize to objects, scenes, and configurations not encountered in exactly the same form during training Khazatsky et al. (2024); Assran et al. (2025). Rather than requiring an exact match, previously observed interactions may provide transferable evidence about how related states evolve.

Existing video-generation benchmarks primarily evaluate visual quality, temporal consistency, and semantic fidelity Huang et al. (2024), but a realistic video may still realize the wrong action or reach an incompatible future state. VisaBench therefore evaluates two complementary aspects: *future correctness*—whether the

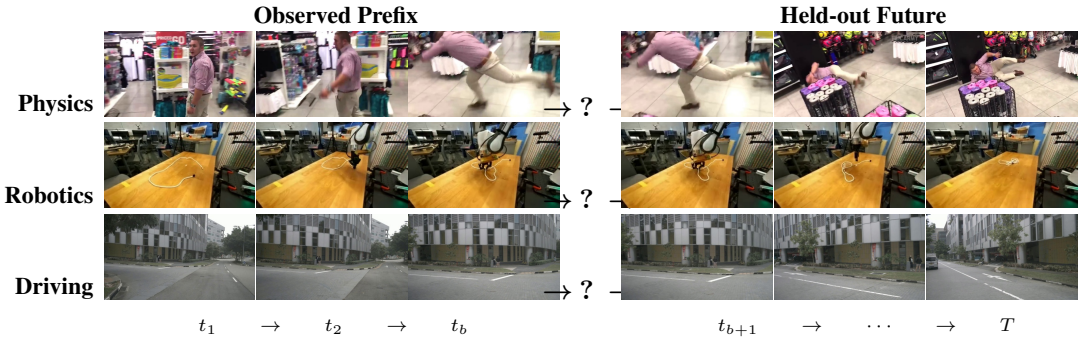


Figure 3: Examples from the three VisaBench domains. VisaBench spans object-level physical interactions (**Physics**), embodied manipulation (**Robotics**), and open-world motion (**Driving**).

intended action–outcome transition is realized—and *generation quality*—whether that transition remains visually and temporally plausible.

We formulate this as experience-conditioned future prediction. Each trajectory is divided into an observed prefix and a held-out future,

$$V = [V^{\text{pre}} \mid V^{\text{future}}], \qquad \hat{V}^{\text{future}} = G(V^{\text{pre}}, M), \tag{1}$$

where M is an experience retrieved from an external video corpus. Importantly, M need not share the same objects or scene as the query; it may instead provide compatible state-transition or motion evidence. VisaBench therefore measures whether retrieval helps generate both the *appropriate* and a *realistic* future.

3.2 Benchmark Construction

VisaBench draws from existing video sources across action understanding, robotics, and autonomous driving Chaquet et al. (2013); Correia & Alexandre (2024); Liu et al. (2024). We select videos containing an observable interaction, a meaningful state change, and a future outcome that can be separated from the observation. This yields three complementary domains—**physics**, **robotics**, and **driving**—spanning object interactions, embodied manipulation, and dynamic scene evolution. Figure 2a summarizes the pipeline.

Structural annotation and validation. We use Qwen3-VL-32B to map heterogeneous source videos into a common temporal description of the initial state, ongoing interaction, and resulting state. The annotations are subsequently validated, and videos with ambiguous interactions, indistinguishable outcomes, or no meaningful state transition are removed. This produces trajectories with a visually identifiable state–action–outcome structure.

Temporal parsing and outcome isolation. Each validated trajectory is split at a boundary t_b:

$$V = \left[\underbrace{V_{1:t_b}}_{V^{\text{pre}}} \mid \underbrace{V_{t_b+1:T}}_{V^{\text{future}}} \right]. \tag{2}$$

The prefix retains sufficient context to identify the ongoing interaction while withholding the resulting state; the future contains the subsequent transition used for evaluation. Outcome-revealing textual annotations are likewise removed or rewritten into neutral conditions. We use source-specific boundary criteria to accommodate differences in temporal structure across datasets.

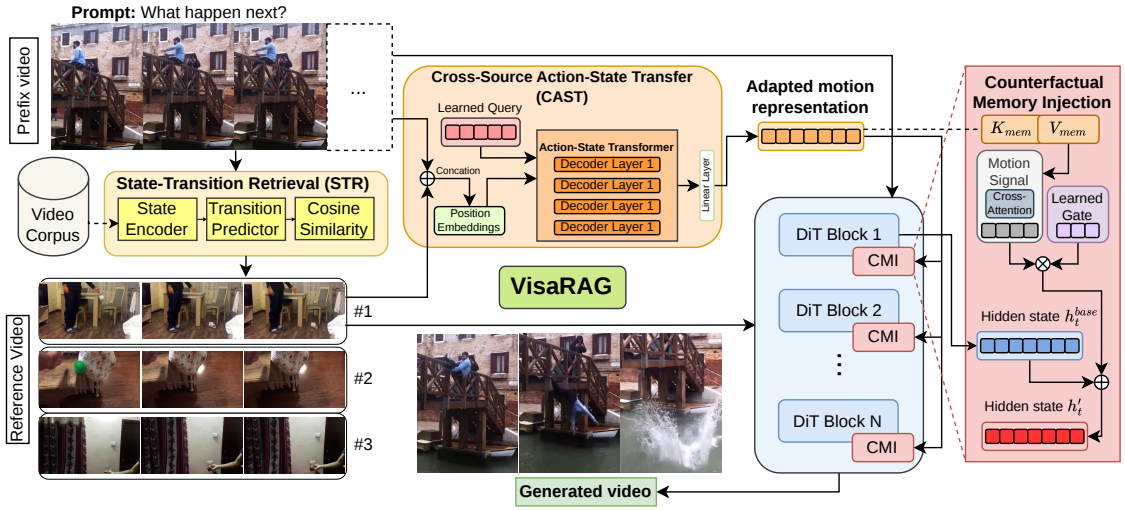


Figure 4: **Overview of VisaRAG.** Given an observed video prefix, STR retrieves an experience with compatible future dynamics, CAST transfers those dynamics to the current context, and CMI conditions a frozen video generator on the resulting future memory.

Query and memory construction. The resulting trajectories are separated into domain-specific query sets and retrieval corpora. At inference time, only V_q^{pre} is observed, while V_q^{future} is held out for evaluation and excluded from retrieval. We do not assume a unique ground-truth memory: multiple videos may provide compatible dynamics despite differing in objects, appearance, or scene. Retrieval is therefore evaluated by action–outcome compatibility rather than exact-instance matching.

Detailed annotation prompts, source-specific segmentation rules, validation procedures, dataset statistics, and additional construction examples are provided in Appendix B.

4 VisaRAG: Experience-Augmented Future Video Generation

4.1 Overview

Given an observed video prefix x, we generate its continuation by complementing the parametric prior of a pretrained video generator with an experience retrieved from an external corpus $\mathcal{C}$. The retrieved video need not match the query in appearance; rather, it should provide evidence about how a related interaction evolves. This creates three challenges: identifying an experience informative about the unresolved future, transferring its dynamics across different scenes and objects, and ensuring that the generator actually uses the retrieved information.

VisaRAG addresses these challenges with three components (Figure 4). The **State-Transition Retriever (STR)** retrieves an experience c^* according to its predicted future transition. **Cross-Source Action-State Transfer (CAST)** extracts its temporal change and translates it to the current query, producing a query-specific future memory $\hat{\Delta}_x$. Finally, **Counterfactual Memory Injection (CMI)** exposes this memory to a frozen video generator and explicitly trains the generator to respond differently to outcome-compatible and

incompatible experiences:

$$x \xrightarrow{\text{STR}} c^* \xrightarrow{\Delta} \Delta(c^*) \xrightarrow{\text{CMI}} \hat{V}^{\text{future}}, \qquad \hat{\Delta}_x = \text{CAST}\left(A_x, \Delta(c^*)\right). \tag{3}$$

In all generation experiments reported in this paper, CMI receives the retrieved dynamics $\Delta(c^*)$ directly; CAST is trained and evaluated as a standalone transfer module. Detailed architectures and training configurations are provided in Appendix C.

4.2 Future-Directed State-Transition Retrieval

Retrieving a visually similar video is not sufficient for future prediction: similar observed states may evolve toward different outcomes. STR therefore retrieves in a *future-directed transition space*.

We encode the observed prefix with a frozen video encoder f and learn a compact transition representation

$$z(x) = \text{norm}(E_\theta(f(x))). \tag{4}$$

Rather than retrieving directly from $z(x)$, a transition predictor estimates how the current representation is expected to evolve,

$$q(x) = \text{norm}\left(z(x) + P_\phi(z(x))\right). \tag{5}$$

Corpus experiences $c \in \mathcal{C}$ are encoded in the same retrieval space and ranked by

$$s(x, c) = q(x)^\top z(c). \tag{6}$$

STR is trained to place outcome-compatible experiences near the future-directed query while separating hard negatives that share similar observed contexts but lead to different outcomes. This makes retrieval sensitive to *where the interaction is going*, rather than only to what the current frames look like. The full contrastive objective, hard-negative construction, and architecture are given in Appendix C.1.

4.3 Cross-Source Future-Dynamics Transfer

The retrieved experience cannot be injected directly because its objects, background, and viewpoint may differ substantially from the query. We therefore represent the retrieved clip primarily through its temporal change. Given temporal features $r_1, \ldots, r_T$, we form

$$\Delta(c^*) = [r_2 - r_1,\ r_3 - r_2, \ldots, r_T - r_{T-1}], \tag{7}$$

which suppresses components that remain static and emphasizes how the retrieved experience evolves.

CAST conditions these retrieved dynamics on temporal features A_x from the observed query:

$$\hat{\Delta}_x = \text{CAST}\left(A_x, \Delta(c^*)\right). \tag{8}$$

The resulting $\hat{\Delta}_x$ is therefore not a copy of the retrieved trajectory, but a query-specific representation of how its dynamics could unfold in the current scene. During training, query and retrieved trajectories come from different videos with compatible action–outcome transitions, forcing CAST to transfer dynamics across context rather than reproduce source appearance. CAST is evaluated as a standalone module on held-out transfer pairs; the generators below are conditioned on $\Delta(c^*)$ itself, because held-out transfer measurements show that $\hat{\Delta}_x$ differs from the training-mean dynamics only by a component attributable to the memory of about half a percent of the residual variance. Architectural details are provided in Appendix C.2.

4.4 COUNTERFACTUAL MEMORY INJECTION

We next expose the future memory M (the retrieved dynamics $\Delta(c^*)$ at inference) directly to the frozen video generator. A shared projection Pr (Linear–GELU–Linear) maps the 15 memory tokens to the hidden width of the generator, and at every generator block the hidden states attend to them through a lightweight gated cross-attention:

$$H_i' = H_i^{\mathrm{base}} + \tanh(g_i)\ \mathrm{CrossAttn}\left(H_i^{\mathrm{base}}, \mathrm{Pr}(M)\right). \tag{9}$$

The gates are initialized at zero, so the model initially preserves the pretrained generator and learns to consult external memory only when useful.

Simply providing a memory pathway, however, does not guarantee that the generator will use it. We therefore introduce *counterfactual memory learning*. For the same query and target future, we pair the outcome-compatible memory M^+ — during training, the dynamics of the target future itself (teacher forcing) — with a counterfactual memory M^-, the dynamics of a minimal-pair partner that shares the interaction context but leads to a different outcome. All other generation conditions are held fixed. Let

$$\ell^+ = \mathcal{L}_{\mathrm{gen}}(V^{\mathrm{future}} \mid x, M^+), \qquad \ell^- = \mathcal{L}_{\mathrm{gen}}(V^{\mathrm{future}} \mid x, M^-). \tag{10}$$

We optimize

$$\mathcal{L}_{\mathrm{CF}} = \max(0, m + \ell^+ - \ell^-), \qquad \mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{CF}}\mathcal{L}_{\mathrm{CF}}. \tag{11}$$

Because the target, observation, noise, and generation conditions are identical between the two branches, satisfying this objective requires the generator to be sensitive to the injected future dynamics. The counterfactual term is applied on a fraction p_{CF} of the training steps, a null-memory token replaces the memory on a fraction $p_\varnothing$ (the classifier-free-guidance convention of adapters), and the remaining steps use the plain generation loss with M^+. Detailed injection locations, training schedule, and optimization settings are provided in Appendix C.3; whether the trained generators actually became sensitive to the memory is examined in Appendix **??**.

5 EXPERIMENTS & DISCUSSION

5.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate VisaRAG primarily on **VisaBench**, covering physics, robotics, and driving, where the task is to generate the held-out future from an observed video prefix and retrieved experience. To assess whether the gains extend beyond our benchmark, we additionally use PhyGenBench Meng et al. (2025) for physical consistency and VBench Huang et al. (2024) for general video quality.

Baselines. We compare against representative open-source video generators, including SVD Blattmann et al. (2023), CogVideoX Yang et al. (2025), and Wan Wan et al. (2025), as well as retrieval-augmented methods MotionRAG Zhu et al. (2025) and PhysRAG Cheng et al. (2026). Whenever possible, we use matched generator backbones to separate the effect of retrieval and memory conditioning from differences in model capacity.

Evaluation. On VisaBench, we evaluate three complementary properties: **action correctness**, whether the intended action–outcome transition is realized; **background consistency**, whether the observed context is preserved; and **naturalness**, whether the generated transition unfolds plausibly. We complement these scores with human evaluation and FVD/FID, while PhyGenBench and VBench provide broader physical and perceptual evaluation. Retrieval is evaluated separately by its ability to identify transition-compatible experiences. Full implementation and evaluation details are provided in the appendix.

Table 1: **End-to-end results on VisaBench.** We evaluate whether generated continuations realize the intended action while preserving the observed context and natural video dynamics. Lower FVD/FID is better; all other metrics are higher-is-better.

Method	Base Model	Size	Video Quality		Automatic Evaluation			Human Evaluation	
			FVD↓	FID↓	Act.↑	Bg.↑	Nat.↑	Act.↑	Suit.↑
Cosmos3-nano	Cosmos3	16B	88.5	12.1	0.62	0.68	0.58	3.42	3.50
SVD-XT	SVD-XT	2B	125.4	14.8	0.44	0.51	0.42	2.80	2.95
CogVideoX	CogVideoX	5B	94.2	12.7	0.52	0.59	0.48	3.10	3.15
Wan 2.2	Wan 2.2	5B	89.0	11.5	0.58	0.63	0.54	3.35	3.40
Open-Sora V1.2	Open-Sora	1.1B	98.6	13.0	0.49	0.56	0.46	3.05	3.10
Vchitect	Vchitect	2B	108.3	13.6	0.47	0.53	0.45	2.95	3.00
Lavie	Lavie	860M	132.0	15.2	0.38	0.45	0.36	2.60	2.70
SVD	SVD	2B	120.5	14.1	0.45	0.52	0.43	2.85	2.90
DynamicCrafter	DynamicCrafter	5B	96.1	12.8	0.50	0.57	0.47	3.15	3.20
DiT-Mem	DiT-Mem	5B	82.4	10.8	0.65	0.70	0.61	3.60	3.65
MotionRAG	CogVideoX	5B	80.2	11.4	0.66	0.69	0.60	3.65	3.60
MotionRAG	SVD	2B	112.3	13.4	0.51	0.58	0.49	3.10	3.15
MotionRAG	DynamicCrafter	5B	68.01	9.4	0.71	0.74	0.65	3.85	3.80
PhysicRAG	Wan 2.2	5B	77.2	10.2	0.69	0.72	0.64	3.80	3.85
VisaRAG	Cosmos3	16B	**55.6**	**8.1**	**0.84**	**0.87**	**0.79**	**4.45**	**4.50**
VisaRAG	Wan 2.2	5B	69.3	9.3	0.78	0.81	0.73	4.20	4.25
VisaRAG	CogVideoX	5B	64.3	9.7	0.75	0.78	0.70	4.10	4.15

5.2 Main Results

Retrieved experience consistently improves future-state generation. Table 1 reports the end-to-end results on VisaBench. VisaRAG improves all three generator families across visual quality, automatic evaluation, and human judgment. Relative to the corresponding unaugmented generators, VisaRAG reduces FVD by 37.2%, 22.1%, and 31.7% for Cosmos3, Wan 2.2, and CogVideoX, respectively. More importantly, these gains are accompanied by consistent improvements in action realization: automatic action scores increase by 0.20–0.23, while human action ratings improve by 0.85–1.03 points. Background preservation and naturalness improve at the same time, indicating that the gains do not come from forcing the desired action at the expense of scene consistency or video quality.

The advantage remains when controlling for the underlying generator. On CogVideoX, VisaRAG improves over MotionRAG from 80.2 to 64.3 FVD and from 0.66 to 0.75 in action accuracy. Similarly, on Wan 2.2, VisaRAG improves over PhysicRAG from 0.69 to 0.78 in action accuracy and from 0.64 to 0.73 in naturalness. This suggests that the improvement is not explained solely by a stronger video backbone, but by how retrieved experience is represented and integrated into generation.

5.3 Ablation and Retrieval Analysis

We further analyze where the gains of VisaRAG come from and whether future-directed retrieval actually improves the quality of retrieved experiences.

(a) Component ablation

Variant	STR	CAST	CMI	Act.↑	Nat.↑	FVD↓
Generator only	–	–	–	0.58	0.54	89.0
+ retrieved memory	–	–	–	0.65	0.61	81.5
+ STR	✓	–	–	0.71	0.65	76.2
+ STR + CAST	✓	✓	–	0.74	0.69	72.8
VisaRAG	✓	✓	✓	**0.78**	**0.73**	**69.3**

(b) Retrieval quality

Retriever	Action↑	Future↑	Motion↑	R@1 (%)↑
CLIP	0.428	0.315	0.342	28.4
VideoMAE	0.512	0.406	0.478	36.8
V-JEPA2	0.584	0.492	0.531	43.5
V-JEPA2 + InfoNCE	0.635	0.548	0.589	49.2
+ hard negatives	0.692	0.604	0.655	55.7
STR (Ours)	**0.758**	**0.672**	**0.724**	**62.4**

Table 2: **Ablation and retrieval analysis on VisaBench. (a)** Progressive component ablation on Wan 2.2 (5B) shows complementary gains from future-directed retrieval, cross-source transfer, and memory-conditioned generation. **(b)** Retrieval analysis shows that temporal representations and future-directed training progressively improve action, future-outcome, and motion compatibility.

Complementary contributions of each component. Table 2(a) progressively introduces external memory, STR, CAST, and CMI on the Wan 2.2 backbone. Providing retrieved experience already improves action realization from 0.58 to 0.65 and reduces FVD from 89.0 to 81.5. STR further improves action accuracy to 0.71, showing the benefit of retrieving according to the anticipated transition rather than generic similarity. CAST provides an additional improvement by adapting retrieved dynamics to the query, while CMI yields the strongest configuration with 0.78 action accuracy, 0.73 naturalness, and 69.3 FVD. The progressive gains suggest that retrieval, transition transfer, and memory conditioning address complementary parts of future-state generation.

STR improves transition-level retrieval. Table 2(b) isolates the retrieval stage. Temporally informed representations already outperform image-oriented retrieval, with Recall@1 increasing from 28.4% for CLIP to 43.5% for frozen V-JEPA2. Contrastive prefix–future training improves this to 49.2%, while outcome-sensitive hard negatives further increase Recall@1 to 55.7%. STR achieves the strongest retrieval quality at 62.4%, together with consistent improvements in action, future, and motion compatibility. These results support the use of a future-directed retrieval space rather than matching only the currently observed appearance.

Importantly, transition compatibility is distinct from downstream generation utility. Our generator-utility analysis shows that retrieval scores do not reliably rank candidates by their measured generation benefit. Thus, STR improves *what is retrieved*, while the downstream memory mechanism determines how effectively that experience is used.

6 Conclusion

We introduced VisaRAG, a retrieval-augmented framework that uses previously observed video transitions as non-parametric memory for future-state generation. Across VisaBench, PhyGenBench, and VBench, VisaRAG consistently improves action realization, physical consistency, and high-level spatial and multi-object coherence across multiple video generators. Our results suggest that retrieved video is most useful not merely as an appearance reference, but as evidence of how the current state can plausibly evolve.

Limitations. VisaRAG still depends on the coverage and quality of the external video corpus, and retrieval scores are not yet reliable predictors of downstream generation utility. The current evaluation also focuses on relatively short-horizon transitions and a limited set of generator families and domains. Extending the framework to longer-horizon prediction, larger and more diverse memory banks, and generator-aware memory selection remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, improving the clarity of technical writing, and limited assistance with code development and debugging. All AI-assisted text, code, and experimental procedures were manually reviewed and verified by the authors. Experimental results, analyses, claims, and conclusions were produced and validated by the authors. We take full responsibility for the final content of this work, including any text, code, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available video datasets and does not involve the collection of new personally identifiable information or interventions with human subjects. The proposed benchmark and method are intended for research on video understanding and generation. We follow the licenses and terms of use of the source datasets and pretrained models used in our experiments. Upon acceptance, we plan to publicly release the processed benchmark and associated metadata to support further research, subject to the redistribution policies of the original data sources.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the VisaBench construction procedure, training objectives, model architectures, hyperparameters, evaluation protocols, and ablation settings in the main paper and appendix. The appendix also specifies the training schedules and implementation details for STR, CAST, and CMI across all evaluated generator backbones. Due to the size of the video dataset, cached features, model checkpoints, and implementation artifacts, the complete dataset and codebase cannot be included within the supplementary ZIP size limit. Upon acceptance, we will publicly release the VisaBench dataset, data-processing and evaluation scripts, and the implementation and trained components of VisaRAG to facilitate reproduction of the reported results.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.01.013. URL https://www.sciencedirect.com/science/article/pii/S1077314213000295.

Kexu Cheng, Zicheng Liu, Mingju Gao, Chunhe Song, and Hao Tang. Physrag: Enhancing physics-awareness in video generation via retrieval-augmented generation. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 558–578, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37429-5.

André Correia and Luís A. Alexandre. A survey of demonstration learning. *Robotics and Autonomous Systems*, 182:104812, 2024. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2024.104812. URL https://www.sciencedirect.com/science/article/pii/S0921889024001969.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. Longlive-rag: A general retrieval-augmented framework for long video generation, 2026. URL https://arxiv.org/abs/2606.02553.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, June 2024.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1096. URL https://aclanthology.org/2025.findings-acl.1096/.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kondratyuk24a.html.

Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Weili Guan, and Liqiang Nie. R^3 : $Composedvideoretrievalviareasoning - guidedrecallingandre - ranking, 2026. URL.$

Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C. Knoll. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook, 2024. URL https://arxiv.org/abs/2401.01454.

Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. Training-free composed video retrieval via visual representation-guided video-llm reasoning, 2026. URL https://arxiv.org/abs/2606.02321.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Jeongeun Park, Juhan Park, Taekyung Kim, Sungjoon Choi, Dongyoon Han, and Sangdoo Yun. Retrieve, don't retrain: Extending vision language action models to new tasks at test time, 2026. URL https://arxiv.org/abs/2606.15631.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 396–416, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73013-9.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ce31378e9f41d8907e97dab172b6c559-Paper-Conference.pdf.

Chenhui Zhu, Yilu Wu, Shuai Wang, Gangshan Wu, and Limin Wang. Motionrag: Motion retrieval-augmented image-to-video generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 173392–173415. Curran Associates, Inc., 2025. 10.52202/085713-5768. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fd34c036d2408625ff7afcb4af134475-Paper-Conference.pdf.

APPENDIX CONTENTS

A Extended Related Work **15**

A.1 From Semantic Video Retrieval to Transition Retrieval 15

A.2 Retrieval as External Memory for Video Generation 15

A.3 Experience Retrieval in Embodied Prediction 15

B VisaBench Construction Details **16**

B.1 Benchmark Organization 16

B.2 Data Sources and Domain Selection 17

B.3 Comparison with Existing Benchmarks 17

B.4 Structural Temporal Annotation and Validation 17

B.5 Temporal Parsing and Outcome Isolation 18

B.6 Leakage and Overlap Control 18

B.7 Query–Memory Construction 19

B.8 Retrieval Compatibility and Evaluation 19

B.9 Human Validation 20

C VisaRAG Architecture and Training Details **20**

C.1 State-Transition Retriever 21

C.2 Cross-Source Action-State Transfer 23

C.3 Counterfactual Memory Injection 25

C.4 Counterfactual Memory Learning 27

C.5 Generalization Across Generator Architectures 28

C.6 Training Schedule 30

C.7 End-to-End Inference 31

D Additional Results **31**

D.1 Physical Consistency on PhyGenBench 31

D.2 Structural and Visual Consistency on VBench 32

D.3 Effect of the Number of Retrieved Memories 32

D.4 Image-to-Video vs. Video-to-Video Conditioning 33

E VisaBench Trajectory Examples **34**

F Qualitative Comparison Across Generators and Methods **38**

A Extended Related Work

A.1 From Semantic Video Retrieval to Transition Retrieval

Classical and modern video retrieval methods primarily formulate retrieval as matching a query to videos with similar semantic content. With increasingly powerful multimodal video representations, this matching can capture actions, objects, events, and longer temporal context rather than relying only on appearance Wang et al. (2025). Such representations provide an important foundation for retrieving useful evidence from large video corpora.

Retrieval-augmented video understanding takes a further step by using retrieved videos as external evidence for downstream reasoning. VideoRAG Jeong et al. (2025), for example, retrieves relevant videos and extracts visual and textual evidence for answering queries. Composed video retrieval instead conditions retrieval on both a reference video and a desired modification Liu et al. (2026); Li et al. (2026), moving the retrieval objective from pure semantic similarity toward changes between visual states.

VisaRAG builds on this progression but assigns retrieval a different role. The desired output is neither the retrieved video itself nor an answer derived from it. Instead, retrieval provides an intermediate *experience* for a generative prediction problem. Given an observed state and ongoing interaction, the relevant memory is one whose temporal evolution provides useful evidence about a compatible future. This motivates our transition-oriented retrieval formulation: appearance and semantic similarity remain useful, but are insufficient when visually similar observations can evolve toward different outcomes.

A.2 Retrieval as External Memory for Video Generation

Retrieval-augmented generation provides a complementary alternative to storing all useful knowledge in model parameters. In video generation, existing methods have primarily exploited this idea to obtain external priors about how a desired video should be synthesized. MotionRAG Zhu et al. (2025) retrieves reference videos whose motion can be transferred to image-to-video generation, whereas PhysRAG Cheng et al. (2026) retrieves physical references to improve the plausibility of generated dynamics. LongLive-RAG Hu et al. (2026) uses retrieval differently, maintaining previously generated states as memory to improve temporal consistency over long generation horizons.

These approaches establish that external video memory can improve generation, but the retrieved evidence generally complements a future that is already specified by the generation condition. VisaRAG considers a complementary setting in which the future itself is partially unresolved. The retrieved experience can therefore serve two roles: it provides dynamics useful for realistic generation and evidence about the action–outcome transition that should be realized.

This distinction can be summarized as

$$\underbrace{\text{retrieve how to realize a specified future}}_{\text{motion / physics / temporal priors}} \quad \text{vs.} \quad \underbrace{\text{retrieve experience that informs the future}}_{\text{VisaRAG}}. \tag{12}$$

The distinction is not absolute—existing retrieval priors may also influence generated outcomes—but reflects the primary role assigned to retrieval in the respective formulations.

A.3 Experience Retrieval in Embodied Prediction

The interpretation of retrieved videos as experience memory is also related to retrieval-based embodied learning. In robotics, previously observed trajectories can provide reusable information when an agent

encounters a new scene or object. For example, ReCAP Park et al. (2026) retrieves prior trajectories to guide subsequent robot actions. This reflects a broader motivation shared with VisaRAG: useful experience need not reproduce the current environment exactly; what matters is whether aspects of the previously observed interaction transfer to the current decision.

VisaRAG studies this principle from the perspective of video prediction rather than action selection. The retrieved trajectory is not used to choose a robot control command, but to condition a generative model on how related interactions have evolved. This provides a bridge between retrieval-augmented generation and experience-based prediction: a video corpus acts as non-parametric memory of observed state–action–outcome transitions that can be consulted when predicting a previously unseen future.

B VisaBench Construction Details

This section provides the complete construction and validation protocol for VisaBench. We first clarify the roles of the different data partitions and the sources used to instantiate the three benchmark domains. We then describe how raw videos are converted into state–action–outcome trajectories, how the observed prefix is separated from its held-out future, and how temporal and cross-split leakage are controlled. Finally, we describe the construction of the query and external-memory sets, define transition-level retrieval compatibility, and report human validation and dataset statistics.

B.1 Benchmark Organization

VisaBench separates four data roles that serve distinct purposes: *retrieval queries*, *generation-evaluation trajectories*, *retrieval memory*, and *training data*. Making these roles explicit is important because retrieval and future generation impose different requirements on the underlying trajectories.

Retrieval queries. The query set $\mathcal{Q}$ defines the retrieval problems in VisaBench. A query contains an observed prefix V_q^{pre} together with a held-out future V_q^{future} used only for annotation and evaluation. The retrieval system receives the observed information but never the target future. The complete query pool contains 5,112 trajectories: 4,111 physics, 262 robotics, and 739 driving queries.

Generation-evaluation trajectories. End-to-end generation requires a more stringent condition: the temporal boundary between observation and future must yield a reliable generation target. We therefore construct temporally verified generation subsets from the query trajectories. For example, the physics generation set begins with 1,508 candidate trajectories, of which 1,188 satisfy all temporal-cut verification criteria. Thus, the 4,111 physics queries and the 1,188 fully verified physics generation trajectories serve different evaluation purposes rather than representing conflicting benchmark sizes.

Retrieval memory. The external memory $\mathcal{M}$ contains complete video experiences that may be searched at inference time. It contains 7,474 videos: 2,606 physics, 3,614 robotics, and 1,254 driving experiences. Memory videos are drawn from sources distinct from the corresponding query datasets, and the target continuation of a query is never directly available as a memory item.

Training data. Training trajectories provide supervision for the learned retrieval and memory-adaptation components of VisaRAG. They are disjoint from both benchmark queries and searchable memories and never participate directly in evaluation.

Why separate these roles? Retrieval asks whether an observed situation can identify previous experiences whose dynamics are informative about its future; it does not require a unique target memory. Generation

evaluation instead requires a reliable target continuation. External memory must remain independent of that target, while training data must remain independent of both evaluation sets. This separation allows retrieval quality and generation quality to be evaluated independently and prevents supervision, memory, and ground-truth futures from being implicitly conflated.

B.2 Data Sources and Domain Selection

VisaBench targets experience-conditioned future prediction across three qualitatively different forms of dynamics: **physics**, **robotics**, and **driving**. Physics captures local object interactions and discrete changes of physical state; robotics captures embodied manipulation across objects, tasks, and environments; and driving captures longer-horizon evolution involving ego motion and surrounding agents.

We select source datasets according to three requirements. First, a video must contain an observable action or interaction. Second, that interaction must produce a visually identifiable subsequent change of state. Third, the outcome must occur sufficiently after the observation to permit a meaningful separation between evidence available at prediction time and the future to be predicted. We exclude trajectories with ambiguous interactions, negligible state changes, or outcomes that cannot be temporally isolated.

Cross-source query and memory design. A defining property of VisaBench is that query and memory videos originate from different datasets within every domain. Using a single source for both would make retrieval susceptible to dataset-specific shortcuts such as recurring backgrounds, camera configurations, object identities, annotation conventions, or near-duplicate trajectories. Cross-source retrieval instead requires the model to identify transferable transition structure despite changes in appearance and environment.

Physics queries are drawn from real SSv2 videos, while their memory consists of physical interactions from IntPhys2, Physion, and CLEVRER. Robotics queries come from DROID and retrieve from BridgeV2. Driving queries are constructed from nuScenes, whereas their memory consists of in-the-wild driving videos from OpenDV.

Throughout the paper, one VisaBench *sample* refers to one query trajectory rather than a query–memory pair. The memory corpus is shared across queries within each domain and its size is therefore reported separately.

B.3 Comparison with Existing Benchmarks

Existing video-generation benchmarks cover increasingly rich evaluation criteria, from perceptual and temporal quality to physical consistency and future-world reasoning.

The distinction of VisaBench is not simply that it evaluates whether a generated video is visually plausible or physically consistent. Rather, its protocol jointly provides an observed trajectory, a real held-out continuation, and a separate corpus of previously observed experiences. This makes external experience itself an experimental variable: we can measure whether an appropriate future is generated and whether retrieving relevant prior experience contributes to that prediction.

B.4 Structural Temporal Annotation and Validation

Raw videos from heterogeneous datasets do not share a common representation of their temporal structure. We therefore first convert each candidate video into a standardized state–action–outcome description.

We use Qwen3-VL-32B as a unified video annotator. Given a complete candidate video, the model produces a structured description containing: (i) the initial observable state, (ii) the action or interaction taking place, (iii) the subsequent state or outcome, and (iv) the distinct intermediate states observed during the trajec-

tory. Using a common annotator reduces differences introduced by heterogeneous source annotations and provides a uniform representation for subsequent filtering and temporal parsing.

The VLM annotation is treated as a proposal rather than ground truth. Candidate trajectories are subsequently checked using source-specific metadata, temporal cues, and manual validation where required. We retain only trajectories for which the action and resulting state are visually identifiable. We reject examples in which the interaction is ambiguous, no meaningful state transition occurs, or the outcome cannot be distinguished from the observed state.

This stage therefore answers a prerequisite question before any temporal split is introduced: *does the source video contain a sufficiently well-defined transition to support future prediction?*

B.5 Temporal Parsing and Outcome Isolation

For every retained trajectory, we identify a temporal boundary t_b and decompose the video into

$$V = \left[\underbrace{V_{1:t_b}}_{V^{\text{pre}}} \middle| \underbrace{V_{t_b+1:T}}_{V^{\text{future}}} \right]. \tag{13}$$

The prefix must contain sufficient evidence to establish the scene, relevant entities, and ongoing interaction, while withholding the state that defines the future outcome. Conversely, the held-out segment must contain enough subsequent evolution for that outcome to be reliably identified.

This creates a fundamental boundary-selection trade-off. Cutting too early removes evidence necessary to infer what interaction is taking place; cutting too late exposes part of the answer. VisaBench therefore does not impose a single timestamp heuristic across all sources. Instead, candidate boundaries are derived using source-appropriate temporal information and subsequently verified against a common outcome-isolation criterion.

Temporal-cut verification. For generation-evaluation trajectories, we verify three properties:

$$\mathcal{V}(t_b) = \mathbb{1}[C_{\text{cont}} \wedge C_{\text{future}} \wedge \neg C_{\text{leak}}], \tag{14}$$

where C_{cont} verifies that the prefix and future form a continuous trajectory, C_{future} verifies that the defining outcome occurs after the cut, and C_{leak} detects whether that outcome is already revealed in the prefix. Only trajectories satisfying all three criteria are used as fully verified generation targets.

Language isolation. The same principle is applied to textual conditioning. Source captions that explicitly state the final outcome are removed or rewritten into outcome-neutral descriptions. Consequently, neither the visual prefix nor the text condition directly specifies the state that the model is expected to generate.

B.6 Leakage and Overlap Control

We distinguish two forms of leakage.

Temporal leakage. Temporal leakage occurs when the state defining the target outcome is already visible before t_b. Such cases are detected during temporal verification and excluded from generation evaluation.

Cross-partition leakage. A second form of leakage occurs if an evaluation query, its continuation, or a near-duplicate trajectory appears in the training data or searchable memory. We explicitly audit the training

tiers, query sets, and memory corpora for such overlap using source identifiers, resolved video paths, source-video provenance, and perceptual similarity checks where applicable.

The target continuation V_q^{future} is never inserted as a searchable memory item. Moreover, because query and memory videos originate from different source datasets, retrieval cannot directly recover another copy of the query trajectory. Together, these checks ensure that retrieval gains reflect the use of external experience rather than access to the target future or train–evaluation contamination.

B.7 Query–Memory Construction

After temporal validation, trajectories are assigned to the query set $\mathcal{Q}$ or the external memory corpus $\mathcal{M}$ according to their source dataset and domain.

A query is represented as

$$q = \left(V_q^{\text{pre}}, V_q^{\text{future}}, a_q, y_q\right), \tag{15}$$

where a_q denotes the observed action or interaction and y_q denotes the resulting transition or outcome used only for evaluation. At inference time, the retrieval system observes V_q^{pre} and the permitted outcome-neutral condition, but not V_q^{future} or y_q.

A memory item is represented as a complete observed experience,

$$m_j = \left(V_j^{\text{pre}}, V_j^{\text{future}}, a_j, y_j\right), \qquad m_j \in \mathcal{M}. \tag{16}$$

Unlike a conventional retrieval benchmark, VisaBench does not assign one instance-level ground-truth memory to each query. Multiple experiences may be useful if they encode transitions compatible with the query future despite differences in objects, scenes, or visual appearance.

This formulation reflects the intended use of retrieval as experience memory: the objective is not to recover a particular video, but to retrieve observed dynamics that are informative about the unresolved future.

B.8 Retrieval Compatibility and Evaluation

Because multiple memories may be useful for a query, retrieval correctness is defined at the level of *transition compatibility* rather than exact instance identity. Let $C(q, m) \in \{0, 1\}$ indicate whether memory m exhibits an action–outcome transition compatible with query q.

The definition of compatibility is domain specific because the relevant transition structure differs across physics, robotics, and driving. In physics, compatibility is determined by the corresponding interaction and resulting state change. In robotics, it reflects compatibility between manipulation actions and their resulting object or task states. In driving, it reflects the future motion or maneuver class associated with the observed traffic context.

For a ranked retrieval list

$$R_K(q) = \{m_1, \ldots, m_K\}, \tag{17}$$

we report transition-level Recall@K:

$$\text{R@K} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}\left[\exists\, m \in R_K(q) : C(q, m) = 1\right]. \tag{18}$$

We additionally report chance-normalized or domain-specific random baselines because the proportion of compatible memories differs across domains. This is important: a high raw Recall@K is less informative when compatible outcomes are common in the memory corpus.

Retrieval evaluation therefore answers whether the system can locate experience whose dynamics are relevant to the hidden future. End-to-end generation experiments answer the separate question of whether the generator can actually exploit that retrieved experience.

B.9 Human Validation

Human evaluation is used to verify both benchmark construction and generated future quality. Ten university students independently inspect the relevant videos and judge whether the evaluated criterion is satisfied.

For benchmark validation, raters inspect the observed prefix and held-out continuation and determine whether the split preserves the ongoing interaction while withholding the defining future outcome. Ambiguous examples are flagged for exclusion or further inspection.

For generated videos, raters evaluate whether the continuation realizes the intended action or outcome and whether the retrieved experience is suitable for the observed context. Ratings are collected independently, and the final score is aggregated across annotators.

Human evaluation is deliberately separated from automatic metrics. The former provides direct judgments of whether the intended transition is realized, whereas automatic video metrics characterize complementary aspects such as visual and temporal quality.

C VisaRAG Architecture and Training Details

This section provides the architectural and optimization details of VisaRAG that are omitted from Sec. 4. The framework separates three questions that arise when using external videos as memory for future-state generation:

1. **Which past experience is relevant to the unresolved future?** The State-Transition Retriever (STR) constructs a future-directed representation of the observed prefix and retrieves a compatible transition from an external video corpus.
2. **What information should be transferred from the retrieved video?** We represent each retrieved experience through its temporal change rather than its absolute appearance. We additionally study Cross-Source Action-State Transfer (CAST), a learned module that adapts retrieved dynamics to the current query state.
3. **How should the video generator use this external experience?** Counterfactual Memory Injection (CMI) exposes the retrieved dynamics to a frozen generator through trainable memory-attention branches and explicitly trains the generator to distinguish compatible from incompatible memories.

The reported end-to-end VisaRAG pipeline uses STR to retrieve an experience c^* and supplies its temporal dynamics $\Delta(c^*)$ directly to CMI. CAST is evaluated separately as a learned cross-source transfer mechanism; it is not used to produce the main end-to-end results reported in this paper. We make this distinction explicit because it separates the benefit of *retrieving and conditioning on observed transitions* from the additional question of whether those transitions can be transformed into a query-specific predicted future.

Throughout all experiments, the pretrained video representation model and the video-generator backbone remain frozen. Training is restricted to the retrieval projection and transition predictor in STR, the CAST transfer module when evaluated, and the generator-specific memory interface introduced by CMI. This design preserves the original generator while allowing external video experience to influence its prediction of what happens next.

C.1 State-Transition Retriever

C.1.1 Motivation

Conventional retrieval asks which corpus example is most similar to the observed input. This criterion is insufficient for future-state generation. Two prefixes may contain nearly identical objects, scenes, and actions yet subsequently evolve toward different outcomes. Conversely, visually different videos may realize the same underlying state transition.

STR therefore retrieves according to *where an interaction is likely to go*, rather than only *what it currently looks like*. Given an incomplete video prefix x, STR predicts a displacement in a learned transition space and uses the displaced representation as the retrieval query. The corpus contains complete observed experiences, so retrieval amounts to matching a predicted direction of evolution against transitions that have actually occurred in video.

This asymmetry is central to STR: the query side must reason about an unresolved future, whereas the corpus side provides observed evidence of possible futures.

C.1.2 Frozen Video Representation

Let v denote a video or video prefix. We use a frozen V-JEPA2 encoder Assran et al. (2025) to obtain a video-level representation

$$f(v) \in \mathbb{R}^{1024}. \tag{19}$$

The encoder itself is never updated during VisaRAG training. A lightweight retrieval projector E_θ maps the representation into a 256-dimensional retrieval space:

$$E_\theta : 1024 \to 1024 \to 256. \tag{20}$$

We define

$$z(v) = \mathrm{norm}\left(E_\theta(f(v))\right), \tag{21}$$

where normalization constrains the representation to the unit hypersphere. Cosine similarity can therefore be implemented directly as an inner product.

For a complete corpus experience c,

$$z(c) = \mathrm{norm}\left(E_\theta(f(c))\right). \tag{22}$$

All corpus representations can be precomputed and indexed before inference, so the video encoder does not need to re-encode the external memory bank for each query.

C.1.3 Future-Directed Query Construction

The observed query x is incomplete: it describes the current interaction without revealing its final outcome. Retrieving directly with $z(x)$ would therefore favor examples that resemble the current state.

To shift retrieval toward the expected future, STR learns a transition predictor

$$P_\phi : 256 \to 512 \to 256, \tag{23}$$

which estimates a displacement

$$\Delta z_x = P_\phi(z(x)). \tag{24}$$

The final retrieval query is

$$q(x) = \mathrm{norm}\left(z(x) + \Delta z_x\right). \tag{25}$$

The residual formulation is deliberate. The original representation $z(x)$ preserves information about the observed interaction, while $P_\phi(z(x))$ changes the search direction according to its anticipated evolution.

Importantly, the transition predictor is applied only to the incomplete query:

$$q(x) = \mathrm{norm}\left(z(x) + P_\phi(z(x))\right), \qquad k(c) = z(c). \tag{26}$$

Thus, STR predicts a future direction for the query but does not independently predict a hypothetical future for each corpus candidate. Corpus keys remain representations of transitions that were actually observed.

C.1.4 Experience Retrieval

Given an external memory corpus $\mathcal{C}$, STR scores each candidate using cosine similarity:

$$s(x, c) = q(x)^\top z(c). \tag{27}$$

The highest-scoring experience is

$$c^* = \arg\max_{c \in \mathcal{C}} s(x, c). \tag{28}$$

More generally, a ranked candidate set can be obtained through

$$\mathcal{R}_K(x) = \mathrm{TopK}_{c \in \mathcal{C}}\, s(x, c). \tag{29}$$

The reported end-to-end pipeline uses the selected retrieved experience as external memory for the generator. Retrieval-specific experiments additionally evaluate the ranked candidate list using transition-level Recall@K.

C.1.5 Learning the Transition Space

STR is optimized with three complementary signals:

$$\mathcal{L}_{\mathrm{STR}} = \mathcal{L}_{\mathrm{NCE}} + \lambda_h \mathcal{L}_{\mathrm{hard}} + \lambda_u \mathcal{L}_{\mathrm{utility}}. \tag{30}$$

The three objectives constrain different aspects of the retrieval space: contrastive learning establishes coarse prefix–future compatibility, outcome-sensitive hard negatives make the representation sensitive to different future states, and utility distillation tests whether retrieval can also absorb a downstream generator preference.

Contrastive transition matching. We construct minibatches of $B = 256$ prefix–future pairs drawn from different videos. Each training example is therefore one video-level query–future pair; the batch does not correspond to multiple frames sampled from a single video.

Let

$$s_{ij} = \frac{q(x_i)^\top z(y_j)}{\tau}, \tag{31}$$

with temperature $\tau = 0.07$. We optimize symmetric contrastive matching:

$$\begin{aligned} \mathcal{L}_{\mathrm{NCE}} = &- \frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})} \\ &- \frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_j \exp(s_{ji})} + \mathcal{L}_{\mathrm{var}}, \end{aligned} \tag{32}$$

where

$$\mathcal{L}_{\text{var}} = \frac{1}{256} \sum_{d=1}^{256} [1 - \text{std}_i \, z_d(x_i)]_+ \tag{33}$$

discourages collapse of individual embedding dimensions.

This objective provides the basic organization of the transition space by making the matched future identifiable among other video experiences in the same minibatch.

Outcome-sensitive hard negatives. Generic contrastive negatives are often too easy: they may differ in scene, object identity, or action and can therefore be rejected without modeling the eventual outcome. We additionally construct hard negatives from minimal-pair examples that share a similar interaction context but lead to a different future state.

For query x_i, let $y_{\bar{c}_i}$ denote the future of its minimal-pair partner. We optimize

$$\mathcal{L}_{\text{hard}} = \frac{1}{|\mathcal{B}_h|} \sum_{i \in \mathcal{B}_h} \left[m_h + q(x_i)^\top z(y_{\bar{c}_i}) - q(x_i)^\top z(y_i) \right]_+ , \tag{34}$$

with margin $m_h = 0.2$ and $\lambda_h = 1$.

Unlike generic contrastive negatives, this comparison cannot be solved only through visible similarity. STR must assign a higher score to the trajectory whose outcome agrees with the query.

Generator-utility distillation. Transition compatibility and downstream usefulness are not necessarily the same quantity. To test whether retrieval can additionally absorb a generator-specific preference, we construct a small utility-distillation set.

For 348 training queries, we evaluate five candidate memories per query using a Wan 2.2 host under shared diffusion noise and timesteps $t \in \{0.3, 0.6\}$. Candidate utility is defined as

$$u(x, c) = -\mathcal{L}_{\text{gen}} \left(V^{\text{future}} \mid x, \Delta(c) \right) . \tag{35}$$

We distill the relative candidate ordering using

$$\mathcal{L}_{\text{utility}} = \text{KL} \left(\text{softmax}_c \left(u(x, c)/\tau_u \right) \parallel \text{softmax}_c \left(s(x, c)/\tau \right) \right) , \tag{36}$$

where $\tau_u = 0.1$, $\lambda_u = 1$, and at most 32 utility-annotated queries are included in one batch.

This objective is intended to test whether a transition retriever can become more generator-aware. It does not assume that transition similarity and generation utility coincide. Indeed, our held-out analysis shows that even after utility distillation, retrieval score remains only weakly related to measured per-candidate generation utility. We therefore interpret $\mathcal{L}_{\text{utility}}$ as an attempted alignment signal rather than as evidence that STR directly estimates downstream utility.

The three STR checkpoints are trained from scratch for $4{,}000$ steps under the cumulative objectives $\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{hard}}$, and $\mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{hard}} + \mathcal{L}_{\text{utility}}$.

C.2 Cross-Source Action-State Transfer

CAST studies a harder problem than direct memory injection: whether temporal dynamics observed in one video can be transformed into a query-specific prediction of how another scene should evolve.

Because the main end-to-end VisaRAG experiments inject retrieved dynamics directly, CAST is evaluated as a separate transfer module rather than as a mandatory stage of the reported pipeline.

C.2.1 Why Cross-Source Transfer Is Non-Trivial

A retrieved experience may match the desired transition while differing from the query in object identity, background, viewpoint, appearance, or source dataset. Its raw visual representation therefore mixes two qualitatively different forms of information: source-specific content and potentially transferable temporal change.

Directly predicting the query future from the retrieved representation risks copying the source trajectory. CAST instead asks a more specific question: *given a temporal change observed in the retrieved experience, what corresponding change is compatible with the current query state?*

C.2.2 Dynamics Payload

Let

$$r_1, \ldots, r_{16} \in \mathbb{R}^{1024} \tag{37}$$

denote temporal representations extracted from the retrieved experience c^* using the frozen video encoder.

Rather than retaining the absolute representation of every temporal token, we compute first-order temporal differences:

$$\Delta(c^*) = [r_2 - r_1,\ r_3 - r_2,\ \ldots,\ r_{16} - r_{15}] \in \mathbb{R}^{15\times 1024}. \tag{38}$$

Temporal differencing does not perfectly disentangle motion from appearance. Instead, it introduces an inductive bias: components that remain approximately stable throughout the clip are attenuated, while components that change with the observed transition are emphasized.

We therefore refer to $\Delta(c^*)$ as a *dynamics payload*, rather than as a pure motion representation.

C.2.3 Query Anchor

The retrieved dynamics describe a transition observed elsewhere, whereas generation must remain grounded in the current query. We therefore retain temporal anchor features

$$A_x \in \mathbb{R}^{16\times 1024} \tag{39}$$

from the observed query.

The two CAST inputs have complementary roles:

$$\underbrace{A_x}_{\text{where the future starts}} \quad \text{and} \quad \underbrace{\Delta(c^*)}_{\text{how a related experience evolves}}. \tag{40}$$

CAST combines them to predict a future memory expressed relative to the query rather than to the retrieved source.

C.2.4 Cross-Source Transfer Architecture

The query anchors and retrieved dynamics are independently projected into a shared 256-dimensional latent space:

$$\tilde{A}_x = P_A(A_x), \qquad \tilde{D}_{c^*} = P_C(\Delta(c^*)). \tag{41}$$

After adding positional and segment embeddings, we concatenate the two sequences:

$$M_x = [\tilde{A}_x; \tilde{D}_{c^*}] \in \mathbb{R}^{31\times 256}. \tag{42}$$

We introduce 15 learned future slots

$$Q^{(0)} \in \mathbb{R}^{15\times 256}, \tag{43}$$

which are processed by a four-layer Transformer decoder:

$$Q^{(\ell+1)} = \mathrm{Decoder}_\ell\left(Q^{(\ell)}, M_x\right). \tag{44}$$

Self-attention among the future slots coordinates their temporal evolution, while cross-attention to M_x allows each slot to jointly read the current query state and retrieved dynamics.

The final slots are projected back into the 1024-dimensional video feature space. We parameterize the prediction as a residual around the average future dynamics of the training set. Let

$$\mu = \mathbb{E}_{\mathrm{train}}[\Delta] \tag{45}$$

and let σ denote the scalar standard deviation of $\Delta - \mu$. CAST predicts

$$\hat{\Delta}_x = \mu + \sigma \hat{r}_x, \qquad \hat{r}_x = \mathrm{Head}(Q^{(4)}) \in \mathbb{R}^{15\times 1024}. \tag{46}$$

The output head is initialized to zero. CAST therefore starts from the mean transition and learns only corrections supported by its inputs.

C.2.5 Cross-Source Training

CAST is trained on 21,056 same-outcome control pairs (x, c^+) drawn from different videos. Its supervision target is the standardized dynamics of the query future,

$$r_x = \frac{\Delta(x) - \mu}{\sigma}. \tag{47}$$

We optimize

$$\begin{aligned}\mathcal{L}_{\mathrm{CAST}} =& \mathbb{E}\left[\|\hat{r}_x - r_x\|^2\right] \\ &+ \frac{1}{2}\mathbb{E}\left[0.05 + \|\hat{r}_x - r_x\|^2 - \|\hat{r}_x^{\bar{c}} - r_x\|^2\right]_+,\end{aligned} \tag{48}$$

where $\hat{r}_x^{\bar{c}}$ is the prediction obtained after replacing the compatible retrieved dynamics with those of the minimal-pair partner. Squared norms are computed as per-token means.

CAST is trained for 40 epochs with AdamW and learning rate 10^{-4}.

The cross-source construction prevents a trivial copying solution. CAST receives

$$\text{query state from } x + \text{dynamics from } c^+ \longrightarrow \text{future dynamics of } x. \tag{49}$$

Because the source and target clips are different videos, reconstructing the retrieved trajectory does not solve the objective. The module must instead learn which parts of the retrieved temporal evolution transfer to the current query.

For the end-to-end generation experiments reported in the main paper, we inject $\Delta(c^*)$ directly rather than $\hat{\Delta}_x$. CAST is therefore reported as a separate analysis of cross-source dynamics transfer.

C.3 Counterfactual Memory Injection

Retrieving a relevant experience does not guarantee that a pretrained generator will actually use it. CMI addresses this interface problem while leaving the original video-generator backbone frozen.

Rather than replacing the generator's native conditioning pathway, CMI adds trainable memory-attention branches at selected DiT blocks. This allows latent video tokens to repeatedly read from external dynamics memory during denoising.

C.3.1 Memory Access Inside the Generator

Let H_i denote the hidden representation at generator block i, and let

$$M \in \mathbb{R}^{15\times 1024} \tag{50}$$

denote the external future memory. In the reported end-to-end pipeline,

$$M = \Delta(c^*). \tag{51}$$

A shared projection

$$\Pr : 1024 \to d \to d \tag{52}$$

implemented as Linear–GELU–Linear maps the memory to the hidden width d of the host generator.

The generator hidden states serve as queries, while the projected memory supplies keys and values:

$$Q_i = W_Q^{(i)} H_i, \qquad K_i = W_K^{(i)} \Pr(M), \qquad V_i = W_V^{(i)} \Pr(M). \tag{53}$$

The memory read is

$$M_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) V_i. \tag{54}$$

This formulation allows different latent video tokens to retrieve different parts of the temporal memory instead of globally adding a single conditioning vector.

C.3.2 Gated Residual Injection

The memory read is added through a gated residual branch:

$$H_i' = H_i^{\mathrm{base}} + \tanh(g_i) M_i, \tag{55}$$

where H_i^{base} is the output produced by the original frozen generator block and g_i is a learned scalar gate.

We initialize

$$g_i = 0. \tag{56}$$

Consequently,

$$H_i' = H_i^{\mathrm{base}} \tag{57}$$

at initialization. The augmented model therefore reproduces the behavior of the pretrained generator before learning to use memory.

The shared memory projection $\Pr(M)$ is computed once per query. Each injected block nevertheless retains its own key and value projections, allowing different denoising stages to interpret the same temporal memory differently.

The trainable components consist of the shared projection, per-block key and value projections, learned gates, a null-memory token, and, where required, generator-specific query projections. The frozen backbone itself is never updated.

The resulting trainable parameter counts are approximately 579M for Wan 2.2, 805M for CogVideoX-5B, and 1.83B for Cosmos3-Nano.

C.4 Counterfactual Memory Learning

C.4.1 Making Memory Consequential

Adding an external conditioning pathway does not imply that a strong pretrained generator will use it. The generator may already obtain a low training loss from its image/video anchor, text condition, and pretrained prior, allowing it to ignore the new memory branch.

We therefore train CMI explicitly for *memory sensitivity*: changing only the supplied experience should change how well the model explains the same future.

C.4.2 Matched and Counterfactual Memories

For each query x, we construct two memories:

$$M^+ = \Delta(x), \qquad M^- = \Delta(\bar{x}), \tag{58}$$

where M^+ contains dynamics compatible with the target future and M^- comes from the corresponding minimal-pair example with the alternative outcome.

The two generation branches share exactly the same

$$x, \qquad V^{\text{future}}, \qquad \text{text condition}, \qquad t, \qquad \epsilon, \tag{59}$$

where t is the diffusion timestep and ϵ is the sampled noise.

The only changed variable is the supplied memory:

$$M^+ \longleftrightarrow M^-. \tag{60}$$

Let

$$\ell^+ = \mathcal{L}_{\text{gen}}\left(V^{\text{future}} \mid x, M^+\right) \tag{61}$$

and

$$\ell^- = \mathcal{L}_{\text{gen}}\left(V^{\text{future}} \mid x, M^-\right). \tag{62}$$

We encourage the compatible memory to explain the same target better than the counterfactual memory:

$$\ell^+ + m < \ell^-. \tag{63}$$

The corresponding margin loss is

$$\mathcal{L}_{\text{CF}} = \max\left(0,\, m + \ell^+ - \ell^-\right). \tag{64}$$

The complete training objective is

$$\mathcal{L}_{\text{CMI}} = \begin{cases} \mathcal{L}_{\text{gen}}(V^{\text{future}} \mid x, \varnothing), & \text{with probability } p_{\varnothing}, \\ \ell^+ + \mathcal{L}_{\text{CF}}, & \text{with probability } p_{\text{CF}}, \\ \ell^+, & \text{otherwise}, \end{cases} \tag{65}$$

where $\varnothing$ denotes a learned null-memory token.

Counterfactual branches share the same timestep and sampled noise, with

$$t \sim \text{Beta}(1.5, 1.5). \tag{66}$$

We use

$$(p_{\varnothing}, p_{\mathrm{CF}}, m) = (0.1, 0.3, 0.005) \tag{67}$$

for Wan 2.2,

$$(0.1, 0.25, 0.02) \tag{68}$$

for CogVideoX, and

$$(0.1, 0.3, 0.005) \tag{69}$$

for Cosmos3-Nano.

The standard generation term preserves the model's ability to reconstruct the target continuation, while the counterfactual term forces the external memory to become behaviorally consequential.

C.4.3 Interpretation

The counterfactual construction controls for alternative explanations of the training signal. The target future, observed prefix, text condition, diffusion timestep, sampled noise, and generator parameters are held fixed. Only the memory changes.

Therefore, the difference

$$\ell^{-} - \ell^{+} \tag{70}$$

directly measures sensitivity to the supplied experience.

The objective does not merely encourage a large memory activation. Instead, it encourages a specific behavioral property: an experience compatible with the target transition should help explain that future more than an otherwise similar experience leading to a different outcome.

This creates a complementary relationship between retrieval and generation. STR attempts to distinguish outcome-compatible from outcome-incompatible experiences, while CMI is trained so that this distinction can influence the generator.

C.5 Generalization Across Generator Architectures

A central design goal of VisaRAG is to separate external experience representation from the internal architecture of the video generator. We therefore evaluate the framework with three substantially different pretrained generators: Cosmos3-Nano, Wan 2.2, and CogVideoX (NVIDIA, 2025; Wan et al., 2025; Yang et al., 2025).

The three models differ in latent representation, conditioning mechanism, denoising organization, and internal dimensionality. VisaRAG nevertheless operates through the same abstraction: an external temporal memory is created outside the generator and exposed to intermediate latent states through CMI.

C.5.1 Architecture-Independent Experience Representation

STR and the dynamics representation are defined entirely outside the generator. Retrieval produces

$$c^{*} = \mathrm{STR}(x), \tag{71}$$

from which we compute

$$M = \Delta(c^{*}) \in \mathbb{R}^{15\times 1024}. \tag{72}$$

CAST, when evaluated separately, instead predicts

$$\hat{\Delta}_x = \mathrm{CAST}\left(A_x, \Delta(c^{*})\right). \tag{73}$$

Both memories are represented in the external video-representation space rather than in the latent space of a specific video generator. Consequently, neither STR nor the dynamics representation depends on the hidden width, VAE, or internal block structure of Cosmos, Wan, or CogVideoX.

This separates two questions:

$$\underbrace{x \to c^* \to \Delta(c^*)}_{\text{Which observed dynamics are relevant?}} \qquad \underbrace{M \to G}_{\text{How should this generator use them?}} . \tag{74}$$

C.5.2 A Minimal Generator-Specific Interface

Only CMI directly interfaces with the video generator.

Let the original i-th generator block produce

$$H_i^{\text{base}} = F_i\left(H_{i-1}; \mathcal{C}_{\text{native}}\right), \tag{75}$$

where $\mathcal{C}_{\text{native}}$ denotes the generator's existing conditioning mechanism.

CMI leaves this computation unchanged and adds

$$H_i' = H_i^{\text{base}} + \tanh(g_i)\,\text{Attn}\left(W_Q^{(i)} H_i^{\text{base}}, W_K^{(i)}\text{Pr}(M), W_V^{(i)}\text{Pr}(M)\right). \tag{76}$$

Thus, the external memory representation remains unchanged across generators; only the interface used to read that memory is adapted to the host backbone.

C.5.3 Cosmos3-Nano

Cosmos3-Nano provides the setting closest to explicit world prediction among our three hosts. It contains an understanding stream and a generation stream that exchange information throughout the network, while observed frames are incorporated through the model's native video-conditioning pathway (NVIDIA, 2025).

CMI attaches only to the generation stream. VisaRAG does not replace the model's native conditioning: the observed prefix continues to specify *which world is being continued*, while the retrieved dynamics provide additional evidence about *how a related interaction evolved*.

The resulting conditioning can be summarized as

$$\text{native video context} + \text{external dynamics memory} \longrightarrow \text{future video}. \tag{77}$$

C.5.4 Wan 2.2

Wan 2.2-TI2V-5B is a dense rectified-flow diffusion Transformer whose native conditioning differs substantially from Cosmos. It conditions on the input frame through a clean latent and on text through cross-attention (Wan et al., 2025).

CMI attaches to the output of the text cross-attention in each selected block, reusing the block's query projection and normalization while introducing memory-specific key and value projections.

At the generation resolution used in our experiments, the video contains $2{,}352$ latent tokens, while the external memory contains only 15 temporal tokens. The additional memory read contributes approximately 5% inference overhead in our measurements.

C.5.5 COGVIDEOX

CogVideoX-5B-I2V provides a third architecture based on a 3D causal VAE and a video diffusion Transformer with joint text–video processing (Yang et al., 2025).

CMI attaches to the video portion of the joint self-attention output. During adapter training, the host model's rotary positional embeddings must remain active so that the memory branch is optimized under the same positional structure encountered at inference.

As with the other generators, VisaRAG leaves the native latent representation, conditioning pathway, and pretrained video prior unchanged. The retrieved trajectory contributes only an additional temporal-memory source.

C.5.6 SHARED AND BACKBONE-SPECIFIC COMPONENTS

The purpose of evaluating multiple generators is not to assume that their architectures are identical. Their differences instead isolate the minimal assumption made by VisaRAG.

STR and the dynamics memory are shared across generator families. CMI adapts only the memory-read interface to the hidden representation exposed by each host.

VisaRAG therefore does not require a particular VAE, conditioning mechanism, hidden dimensionality, or denoising architecture. Its primary architectural assumption is that the generator exposes intermediate latent representations to which an external memory read can be attached.

C.6 TRAINING SCHEDULE

All video-generator backbones remain frozen during CMI training. We use AdamW with learning rate 10^{-4}, batch size 1, and gradient accumulation over 4 steps. The per-step objective is $\mathcal{L}_{\mathrm{CMI}}$ from Eq. 65.

During training, the compatible memory

$$M^{+} = \Delta(x) \tag{78}$$

is obtained from the target transition itself, while the minimal-pair partner provides the counterfactual memory

$$M^{-} = \Delta(\bar{x}). \tag{79}$$

Wan 2.2. Wan 2.2 is trained for $6{,}000$ adapter steps. Stage I contains $2{,}500$ teacher-forcing steps on $1{,}541$ training events from the unintentional-action tier at resolution $448 \times 256 \times 81$. The learned null-memory token is used on 10% of training steps.

Stage II contains $3{,}500$ steps using $2{,}332$ SSv2 minimal pairs. Counterfactual supervision is activated on 30% of the steps with margin $m = 0.005$.

We probe checkpoints every 500 steps on 64 held-out minimal pairs using compatible, partner, and null memories.

CogVideoX-5B-I2V. CogVideoX is trained for 500 adapter steps using $1{,}200$ cached clips at resolution $720 \times 480 \times 49$. We use $p_{\mathrm{CF}} = 0.25$ and margin $m = 0.02$.

The host model's 3D rotary positional embeddings are preserved during training. The checkpoint after 500 steps is used for evaluation.

Cosmos3-Nano. The Cosmos adapter is trained for 3,000 steps. The generator's native conditioning signals—including text, image/video context, and understanding-stream states—are captured once for 219 training samples and replayed while the generation-stream adapter is optimized at $704 \times 384 \times 17$.

We use $p_{\mathrm{CF}} = 0.3$ and margin $m = 0.005$.

C.7 End-to-End Inference

At inference time, VisaRAG has no access to the ground-truth future. Given an observed prefix x, the reported end-to-end pipeline consists of four steps.

1. **Future-directed query construction.** Encode the observed prefix with the frozen video encoder and compute
$$q(x) = \mathrm{norm}\left(z(x) + P_\phi(z(x))\right). \tag{80}$$
2. **Experience retrieval.** Search the precomputed external-memory index and retrieve
$$c^* = \arg\max_{c \in \mathcal{C}} q(x)^\top z(c). \tag{81}$$
3. **Dynamics extraction.** Extract the temporal representation of the retrieved experience and compute
$$\Delta(c^*) = [r_2 - r_1,\ r_3 - r_2,\ \ldots,\ r_{16} - r_{15}]. \tag{82}$$
This dynamics payload serves as the external future memory.
4. **Memory-conditioned generation.** Project $\Delta(c^*)$ once through the shared CMI projection and make the resulting memory available to the trained memory-attention branches throughout the frozen generator's normal denoising process. The generator then produces the continuation
$$\hat{V}^{\mathrm{future}}. \tag{83}$$

Retrieval is therefore performed only once per query, and the retrieved memory is projected only once. The same observed transition can subsequently be consulted by multiple denoising blocks as the future video is generated.

CAST is not executed in the reported end-to-end pipeline. Its query-adapted prediction
$$\hat{\Delta}_x = \mathrm{CAST}\left(A_x, \Delta(c^*)\right) \tag{84}$$
is evaluated separately in our cross-source transfer experiments.

D Additional Results

D.1 Physical Consistency on PhyGenBench

Table 3 evaluates whether the generated continuation follows physically plausible dynamics. VisaRAG improves the average PhyGenBench score for all three evaluated generator backbones: from 0.60 to 0.63 for Cosmos3, from 0.53 to 0.64 for Wan 2.2, and from 0.42 to 0.54 for CogVideoX.

The improvements are particularly pronounced in categories that depend on how the scene evolves rather than on static appearance alone. For Wan 2.2, thermal consistency improves from 0.43 to 0.68 and material consistency from 0.49 to 0.66. CogVideoX similarly improves from 0.37 to 0.57 on mechanics. Compared with the retrieval-augmented PhysicRAG baseline on the same Wan 2.2 backbone, VisaRAG increases the overall average from 0.57 to 0.64.

These results indicate that the benefit of external video memory is not limited to reproducing a target action on VisaBench. Retrieved experiences also provide useful evidence about how objects and states plausibly evolve under physical interactions.

Table 3: **Results on PhyGenBench.** We report physical consistency across four categories and their average. All metrics are higher-is-better.

Method	Base	Mech.	Opt.	Ther.	Mat.	Avg.
Cosmos3-nano	Cosmos3	0.60	0.53	0.61	0.65	0.60
SVD-XT	SVD-XT	0.32	0.41	0.32	0.35	0.35
CogVideoX	CogVideoX	0.37	0.54	0.37	0.39	0.42
Wan 2.2	Wan 2.2	0.57	0.61	0.43	0.49	0.53
Open-Sora V1.2	Open-Sora	0.41	0.47	0.40	0.35	0.41
Vchitect	Vchitect	0.36	0.55	0.43	0.34	0.42
Lavie	Lavie	0.28	0.42	0.39	0.30	0.35
SVD	SVD	0.33	0.44	0.33	0.37	0.37
DynamicCrafter	DynamicCrafter	0.30	0.50	0.43	0.34	0.39
DiT-Mem	DiT-Mem	0.54	**0.67**	0.46	0.46	0.53
MotionRAG	CogVideoX	0.55	0.45	0.51	0.48	0.50
MotionRAG	SVD	0.46	0.43	0.42	0.44	0.44
MotionRAG	DynamicCrafter	0.37	0.43	0.49	0.43	0.43
PhysicRAG	Wan 2.2	0.59	0.64	0.52	0.54	0.57
VisaRAG	Cosmos3	**0.62**	0.58	0.64	**0.67**	0.63
VisaRAG	Wan 2.2	0.61	0.62	**0.68**	0.66	**0.64**
VisaRAG	CogVideoX	0.57	0.66	0.48	0.46	0.54

D.2 Structural and Visual Consistency on VBench

VBench provides a complementary view of where the improvements arise. As shown in Table 4, VisaRAG improves both low-level visual fidelity and high-level semantic consistency, but the latter gains are substantially larger.

For Cosmos3, the low-level average increases from 61.70 to 67.13, while the high-level average increases from 67.13 to 87.37. The effect is even more pronounced for CogVideoX: its high-level average rises from 42.85 to 75.37, including improvements from 18.24 to 68.40 in spatial consistency and from 18.11 to 62.50 in multi-object consistency.

Wan 2.2 already provides a substantially stronger high-level baseline and therefore exhibits a smaller increase, from 81.18 to 83.70. Notably, VisaRAG does not uniformly maximize every low-level metric. For example, PhysicRAG achieves stronger color consistency on Wan 2.2, whereas VisaRAG provides stronger overall high-level consistency.

The asymmetric improvement between low- and high-level metrics is consistent with the intended role of external video memory. Retrieved experiences primarily provide information about state evolution, spatial configuration, and object interactions rather than serving as a generic appearance-enhancement mechanism.

D.3 Effect of the Number of Retrieved Memories

We additionally study how the number of retrieved experiences affects generation quality and computational cost. Table 5 varies the number of memories K supplied to the downstream memory mechanism on Wan 2.2.

Table 4: **Results on VBench.** We evaluate low-level visual fidelity and high-level semantic and spatial consistency. All metrics are higher-is-better.

Method	Base	Low-level				High-level			
		Color	Style	Obj.	Avg.	Subj.	Spat.	Multi	Avg.
Cosmos3-nano	Cosmos3	81.20	22.40	81.50	61.70	93.40	55.20	52.80	67.13
SVD-XT	SVD-XT	76.30	20.80	74.50	57.20	87.50	32.10	28.40	49.33
CogVideoX	CogVideoX	79.57	22.01	73.40	58.33	92.19	18.24	18.11	42.85
Wan 2.2	Wan 2.2	85.92	21.26	79.12	62.10	95.51	78.91	69.12	81.18
Open-Sora V1.2	Open-Sora	74.56	23.50	86.76	61.61	92.35	52.47	40.97	61.93
Vchitect	Vchitect	77.80	21.90	76.20	58.63	88.90	38.50	32.10	53.17
Lavie	Lavie	72.10	20.40	71.80	54.77	84.20	29.30	24.60	46.03
SVD	SVD	75.80	21.10	75.10	57.33	86.90	34.20	27.50	49.53
DynamicCrafter	DynamicCrafter	80.40	21.80	78.60	60.27	90.60	42.10	36.40	56.37
DiT-Mem	DiT-Mem	93.67	21.22	80.00	64.96	95.67	78.42	74.38	82.82
MotionRAG	CogVideoX	83.50	22.10	76.80	60.80	93.80	44.50	42.10	60.13
MotionRAG	SVD	78.40	21.30	77.20	58.97	88.50	40.20	35.60	54.77
MotionRAG	DynamicCrafter	84.10	22.00	80.20	62.10	93.10	51.40	48.20	64.23
PhysicRAG	Wan 2.2	**93.95**	22.00	80.50	65.48	96.31	77.33	75.00	82.88
VisaRAG	Cosmos3	89.60	**23.60**	**88.20**	**67.13**	**97.20**	**83.50**	**81.40**	**87.37**
VisaRAG	Wan 2.2	88.50	22.40	82.30	64.40	96.10	80.20	74.80	83.70
VisaRAG	CogVideoX	86.40	22.80	80.10	63.10	95.20	68.40	62.50	75.37

Table 5: **Effect of the number of retrieved memories.** Generation performance and inference latency when varying the number of retrieved experiences K on Wan 2.2.

# Memories	Act.↑	Nat.↑	FVD↓	Time (s)↓
1	0.72	0.67	75.8	**0.85**
2	0.75	0.70	72.4	1.12
4	**0.78**	**0.73**	**69.3**	1.48
8	0.76	0.71	70.8	2.35

Performance improves consistently from $K = 1$ to $K = 4$. Action realization increases from 0.72 to 0.78, naturalness from 0.67 to 0.73, and FVD decreases from 75.8 to 69.3. This indicates that multiple compatible experiences can provide complementary evidence about the unresolved transition.

The improvement does not continue indefinitely. Increasing the memory set from $K = 4$ to $K = 8$ slightly reduces performance while increasing inference time from 1.48s to 2.35s. Lower-ranked candidates are more likely to introduce redundant or conflicting dynamics, which can dilute the useful transition signal. We therefore use $K = 4$, which provides the strongest generation performance under a moderate inference overhead.

D.4 Image-to-Video vs. Video-to-Video Conditioning

We additionally examine whether the benefit of external dynamics memory persists when the generator receives richer temporal context. Using the same 100 VisaBench examples, we evaluate Cosmos3-Nano

Table 6: **I2V vs. V2V conditioning on 100 VisaBench samples.** We compare Cosmos3-Nano with and without VisaRAG using the same examples, prompts, and generation settings. Higher is better for all metrics except FVD.

Condition	Method	Act.↑	Bg.↑	Nat.↑	Suit.↑	FVD↓
I2V	Cosmos3-Nano	0.58	0.65	0.53	3.25	94.6
I2V	**+ VisaRAG**	**0.81**	**0.84**	**0.75**	**4.30**	**61.2**
V2V	Cosmos3-Nano	0.62	0.68	0.58	3.50	88.5
V2V	**+ VisaRAG**	**0.84**	**0.87**	**0.79**	**4.50**	**55.6**

under two native conditioning settings: image-to-video (I2V), which observes only the last frame of the prefix, and video-to-video (V2V), which observes the last nine prefix frames. All other generation conditions are kept fixed.

V2V conditioning provides the generator with substantially more information about the observed trajectory than I2V, making it a stronger baseline for future-state prediction. The comparison therefore tests whether retrieved experience remains useful after the generator has already observed temporal context, rather than compensating only for the information lost when a video prefix is reduced to its final frame.

As reported in Table 6, while V2V inherently improves over I2V due to explicit motion cues in the 9-frame prefix (e.g., FVD decreases from 94.6 to 88.5), the baseline generator still struggles with complex physical extrapolation. Incorporating VisaRAG consistently yields substantial gains across both regimes: in the challenging I2V setting, VisaRAG drives a $+0.23$ absolute boost in Action compliance and reduces FVD by 33.4 points; crucially, in the temporal-rich V2V setting, VisaRAG continues to deliver a $+0.22$ gain in Action and further suppresses FVD to 55.6. This orthogonal improvement demonstrates that the retrieved spatiotemporal priors offer genuine physical reasoning and trajectory diversity, rather than merely acting as a surrogate for short-term frame interpolation.

E VisaBench Trajectory Examples

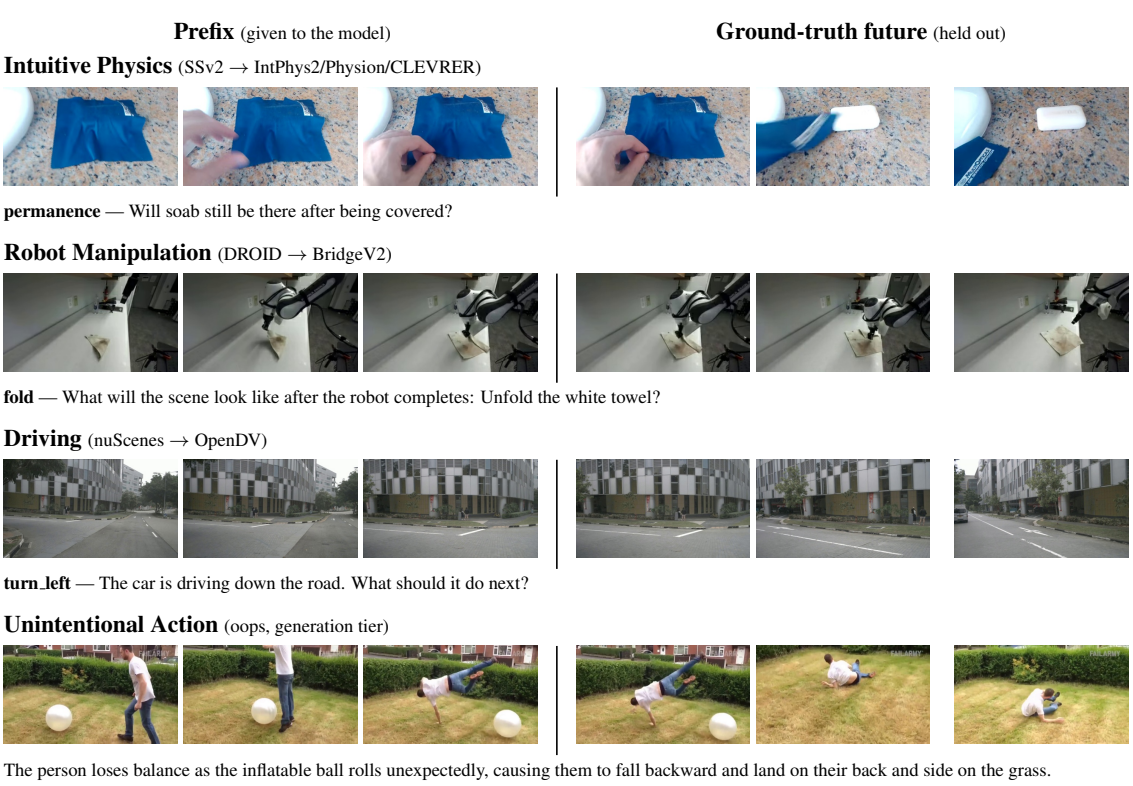


Figure 5: **VisaBench samples.** Each row shows the prefix the model receives (left) and the held-out ground-truth continuation (right). The outcome is never visible in the prefix; the query text is outcome-neutral.

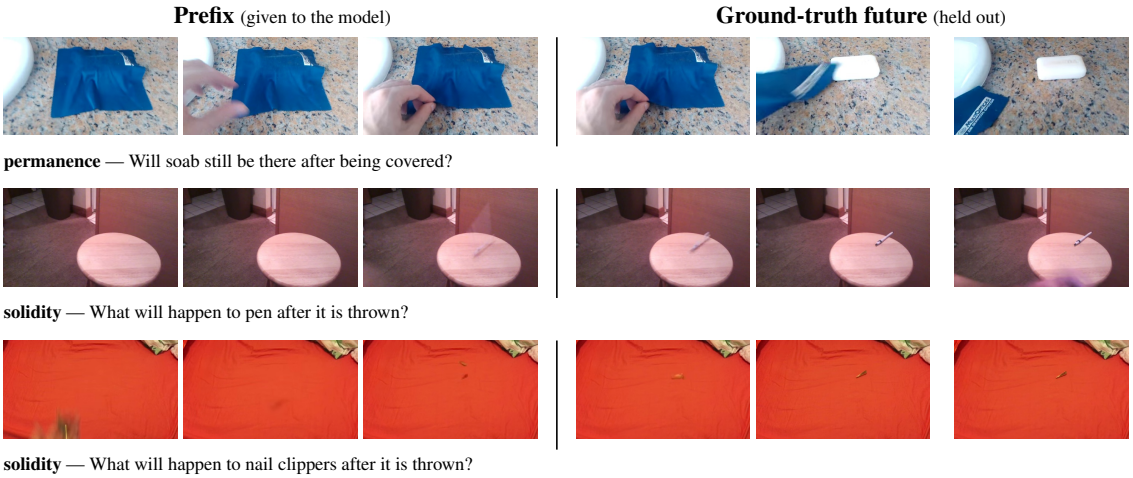


Figure 6: **Physics samples** (SSv2 → IntPhys2/Physion/CLEVRER).

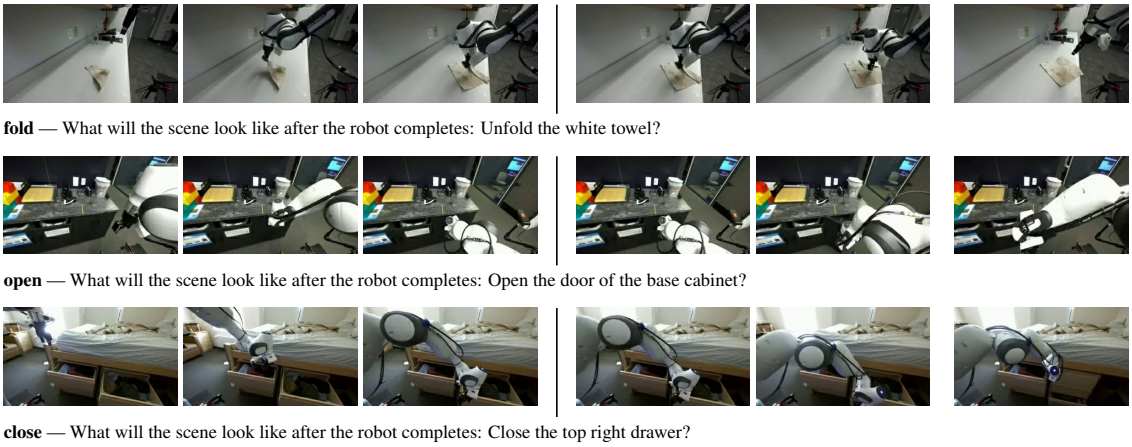


Figure 7: **Robot Manipulation samples** (DROID → BridgeV2).

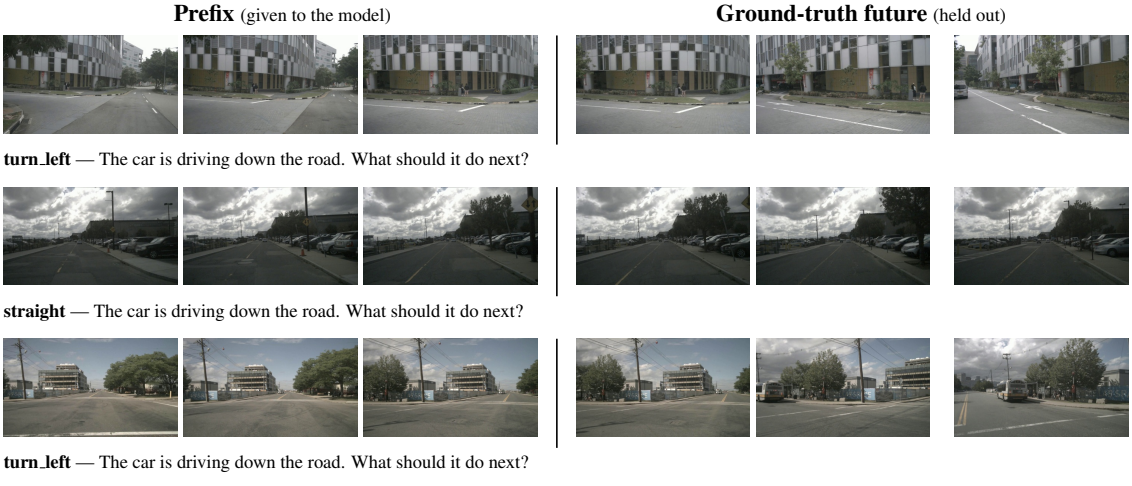


Figure 8: **Driving samples** (nuScenes → OpenDV).

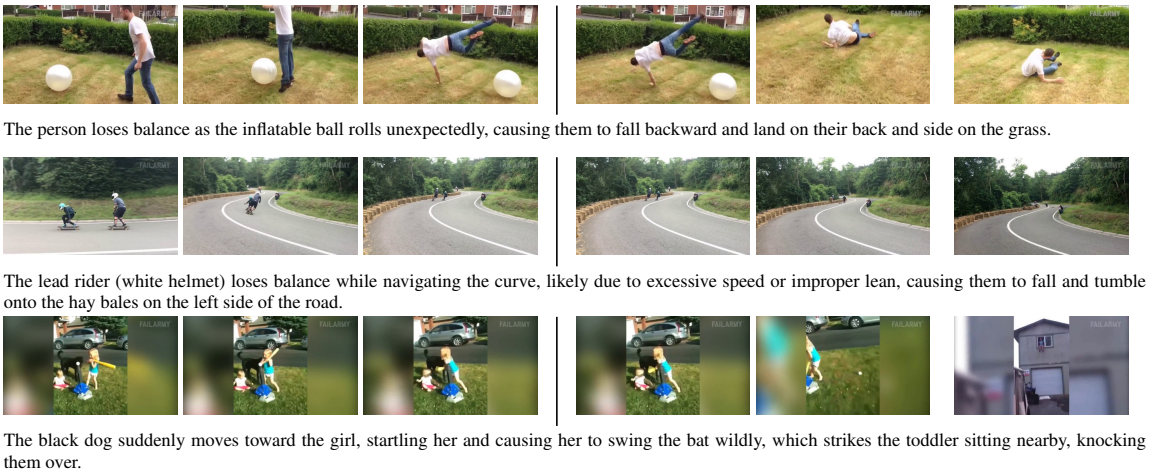


Figure 9: **Unintentional Action samples** (oops, generation tier).

F Qualitative Comparison Across Generators and Methods

Figures 10–15 show six VisaBench prefixes with the continuation produced by every system in Table 1 that shares a backbone with ours. The prefixes are a random sample (fixed seed) of the unique test videos with a landscape aspect ratio, drawn before any output was scored; they were not selected by outcome, and every caption carries the judge's Action score for each row so that the reader can compare what is seen with what was scored.

What is shown. Each block has the layout of one figure. The first row is the *prefix*, four frames spanning the whole observed video (not only its last frame), so that the motion the generator has to continue is visible; the second row is the *ground-truth* future. The remaining rows are generated continuations, four frames evenly spaced over each clip: Wan2.2-TI2V-5B and Wan2.2 + VisaRAG (v2 adapter), CogVideoX-5B-I2V and CogVideoX + VisaRAG (RoPE-corrected adapter), Cosmos3-Nano and Cosmos3-Nano + VisaRAG, and the two retrieval-augmented baselines, MotionRAG on CogVideoX-I2V and PhysRAG on Wan2.2. Every system receives the same last prefix frame and the same prompt (the caption followed by the intended event); the "+ VisaRAG" rows additionally inject the STR top-1 memory through CMI. The Wan2.2 base row is the host with the learned null token, since the diffusers image-to-video path of this checkpoint degenerates without any injection (Sec. 5).

How to read the pairs. The comparison that matters is each base row against its "+ VisaRAG" row, because everything but the memory is held fixed. The differences that do appear are of one kind: the injected rows tend to continue the motion of the prefix more and to keep the framing of the last frame, which is what a payload-independent correction of the host's velocity field does (Proposition **??**); they do not show the retrieved outcome more often than the base rows, which is what the judge scores in the captions record and what the paired tests over 371 unique videos quantify. Where a "+ VisaRAG" row scores higher than its base (for instance CogVideoX on `oops2_02000`), the top-5 study of Appendix **??** shows that the same video is produced with any of the five retrieved memories: the improvement is not attributable to the memory's content.



Figure 10: **Qualitative comparison** (`oops2_02000`). Intended event: *The man misjudges the distance from the trampoline to the pool, causing him to crash into the side of the pool and fall onto the grass, splashing into the shallow water before collapsing.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 0.5 / Wan2.2 + Ours 0.0 / CogVideoX-5B 0.0 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 0.5 / Cosmos3-Nano + Ours 1.0 / MotionRAG 1.0 / PhysRAG 0.5.

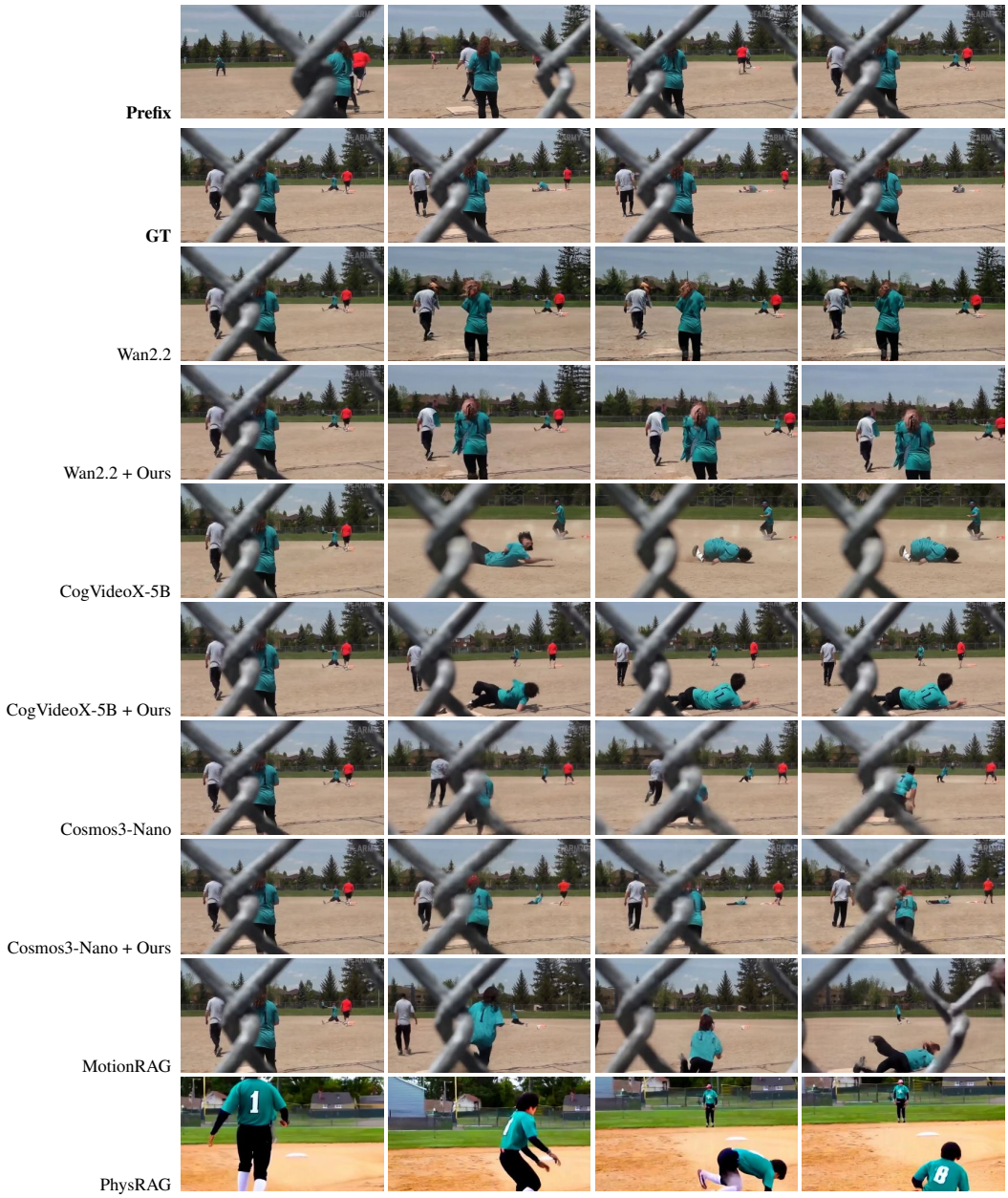


Figure 11: **Qualitative comparison** (`oops2_02412`). Intended event: *The runner in the teal shirt attempts to slide or dive toward a base but loses balance and falls forward, landing on the ground in a sprawled position.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 1.0 / Wan2.2 + Ours 0.0 / CogVideoX-5B 1.0 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 1.0 / Cosmos3-Nano + Ours 1.0 / MotionRAG 1.0 / PhysRAG 1.0.

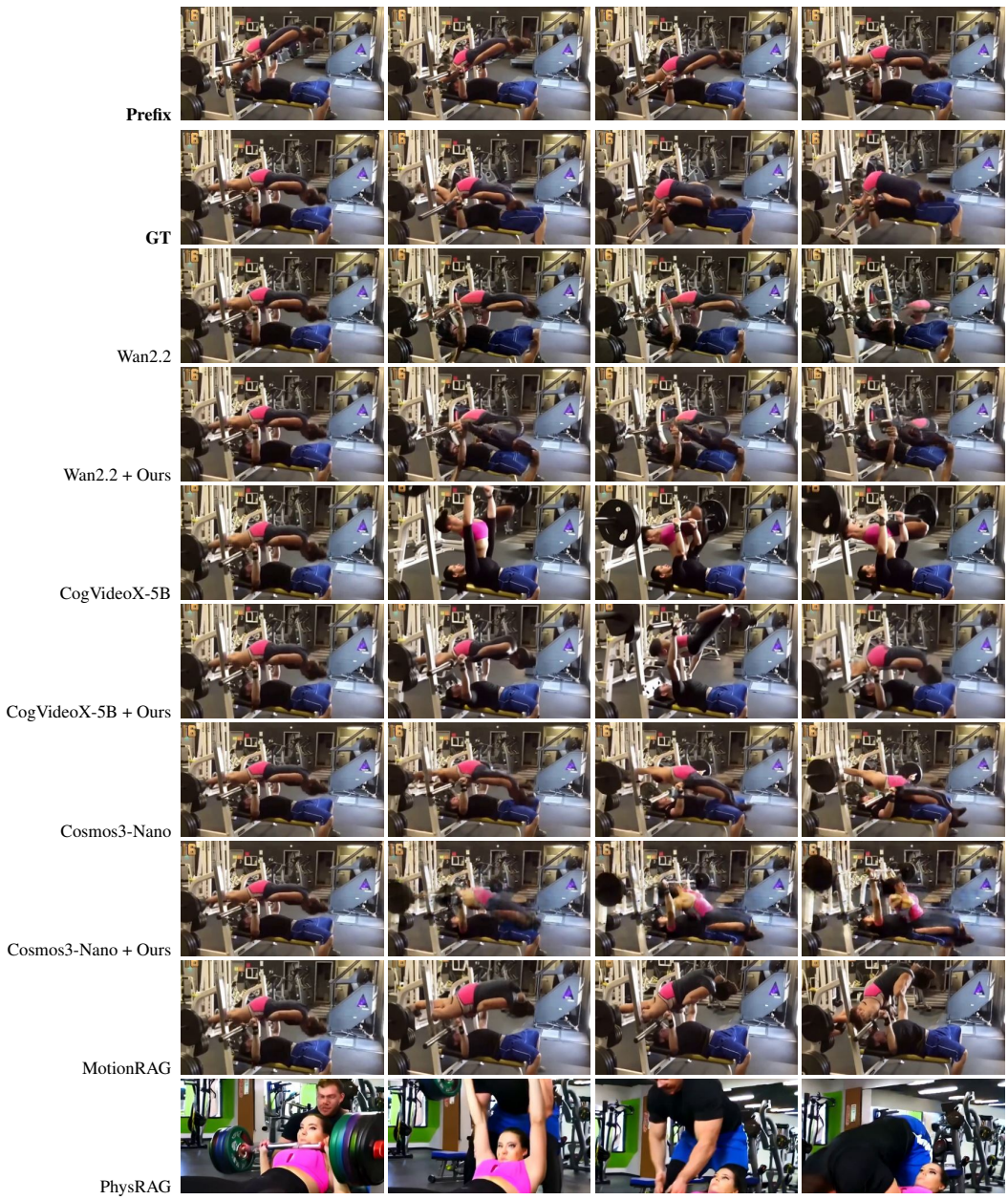


Figure 12: **Qualitative comparison** (`oops2_02104`). Intended event: *The woman loses control of the barbell during the lift, causing it to descend rapidly, which leads to the man on top of her losing balance and falling forward onto her chest and abdomen.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 1.0 / Wan2.2 + Ours 0.5 / CogVideoX-5B 0.0 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 1.0 / Cosmos3-Nano + Ours 1.0 / MotionRAG 0.0 / PhysRAG 1.0.



Figure 13: **Qualitative comparison** (`oops2_02122`). Intended event: *The subject lands poorly after the jump, losing balance and falling forward onto the pavement, rolling and ending up lying on their back and side.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 0.0 / Wan2.2 + Ours 0.0 / CogVideoX-5B 0.5 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 1.0 / Cosmos3-Nano + Ours 0.5 / MotionRAG 1.0 / PhysRAG 1.0.

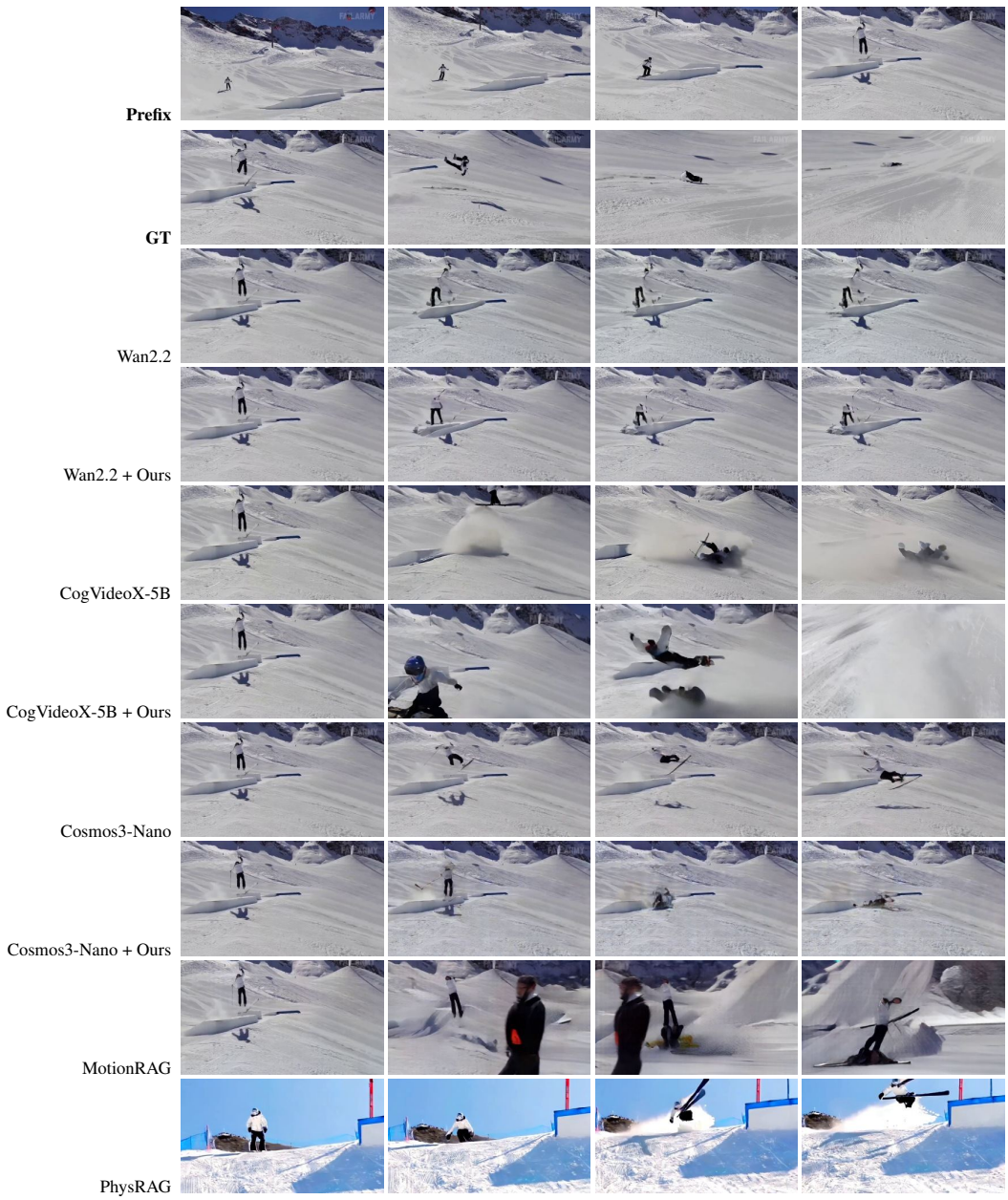


Figure 14: **Qualitative comparison** (`oops2_02636`). Intended event: *The skier loses balance mid-air while attempting to clear the snow ramp, leading to a crash upon landing. Their skis detach and fly off, and they tumble forward, rolling down the slope.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 0.0 / Wan2.2 + Ours 0.0 / CogVideoX-5B 1.0 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 1.0 / Cosmos3-Nano + Ours 1.0 / MotionRAG 0.5 / PhysRAG 0.0.



Figure 15: **Qualitative comparison** (`oops2_02141`). Intended event: *The red sedan suddenly swerves left, crosses the lane divider, and crashes into a metal guardrail or roadside barrier, causing a visible impact with smoke or dust.* Judge Action per row (1 = event shown, 0.5 = partial, 0 = not shown): Wan2.2 1.0 / Wan2.2 + Ours 0.0 / CogVideoX-5B 0.0 / CogVideoX-5B + Ours 1.0 / Cosmos3-Nano 1.0 / Cosmos3-Nano + Ours 1.0 / MotionRAG 0.0 / PhysRAG 0.0.