Retrieve What's Missing: Coverage-Maximizing Retrieval for Consistent Long Video Generation

Minseok Joo[1] **Dogyun Park**[1] **Taehoon Lee**[2] **Kyujin Lee**[2] **Hyunwoo J. Kim**[2†]

[1]Korea University [2]KAIST [†]Corresponding author

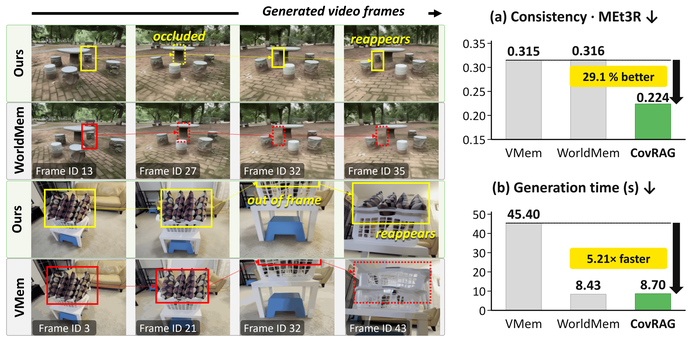
Machine Learning & Vision Lab · KAIST · Korea University

arXiv paper · Project page · Code (GitHub)

NEURAL INFORMATION PROCESSING SYSTEMS

Summary

CovRAG retrieves past frames by **what they add**: it maps which **target-view pixels** each memory frame supports and greedily picks frames that cover what the context **misses**.



- ✓ **Pixel-level 3D evidence** — **coverage maps** from cached VGGT depth; no global 3D scene to build or align.
- ✓ **Complementary selection** — greedy **residual coverage gain**: each frame covers only what is still missing.
- ✓ **Scalable** — **depth caching** keeps per-step geometry cost nearly constant; **training-free** plug-in for pretrained world models.

Background & Motivation

Autoregressive video diffusion sees only a bounded window



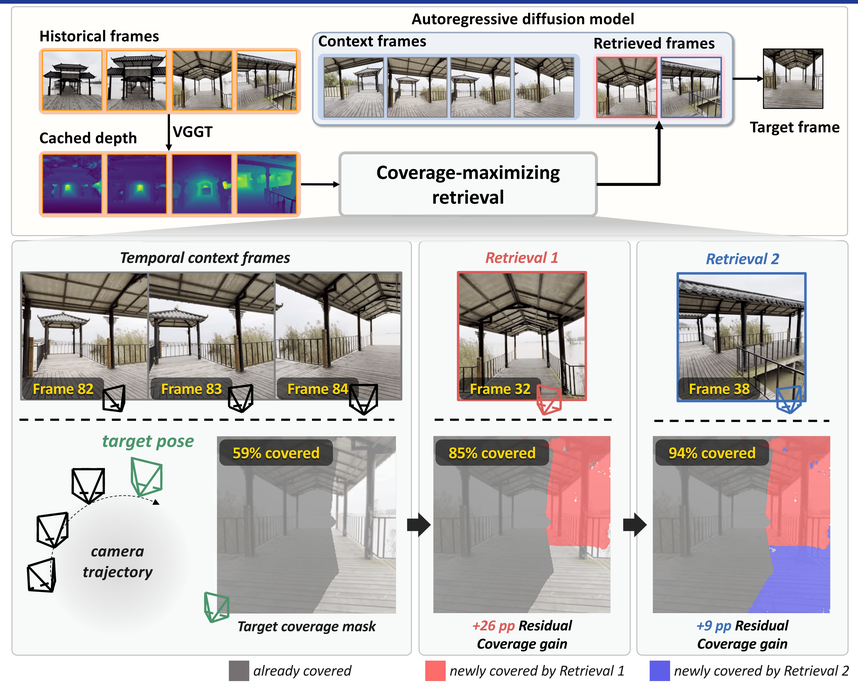
- Autoregressive video diffusion conditions each new frame only on the last **δ context frames**; everything generated earlier is forgotten.
- ⚠ **Spatial drift & inconsistent structure**: when the camera comes back to a place seen before the window, the model cannot reproduce the geometry it generated — rooms change, objects vanish.
- Memory retrieval helps only if it brings back the right frames — so two things matter: **what evidence** says a frame is relevant, and **how the frames are chosen**.

Memory retrieval: what evidence? how to select frames?

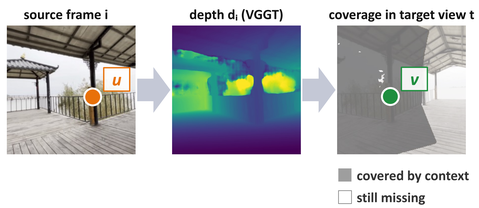
	FoV overlap *camera frustum*	Global 3D scene *reconstruct & update*	CovRAG (Ours) *cached depth per frame*
Per-pixel 3D evidence		✓	✓
Scalable over long rollouts	✓		✓
Complementary selection			✓

- **FoV overlap** (WorldMem) is cheap but coarse — a frustum hit ≠ a pixel actually seen, so occlusion stays ambiguous.
- **Global 3D scene memory** (VMem, Spatia) is fine-grained but must reconstruct and update the scene as the rollout grows.
- **CovRAG** gives per-pixel evidence at near-FoV cost, with a selection rule that **rewards only what is still uncovered**.

CovRAG: Coverage-Maximizing Retrieval-Augmented Generation



① Target-view coverage map as 3D memory evidence



Source-to-target warping

$$\mathbf{u} \xrightarrow{d_i(\mathbf{u}),\,K_i^{-1}} \mathbf{X}_i \xrightarrow{T_{t\leftarrow i}} \mathbf{X}_t \xrightarrow{K_t,\,\pi} \mathbf{v}$$

source pixel · 3D point (i) · 3D point (t) · target pixel

Coverage map $\mathbf{m}_{i\to t}(\mathbf{v}) = 1$

Context coverage $\mathbf{m}_t^{\mathrm{ctx}} = \bigvee_{i\in\mathcal{C}_t} \mathbf{m}_{i\to t}$

② Residual coverage gain → non-redundant retrieval

- **Score** a candidate only by the target pixels that are **still uncovered** by context and already-selected frames.
- **Greedy**: take the arg-max, OR its map into the coverage, repeat n times (Algorithm 1).
- Redundant frames score **0** → a compact, **complementary** memory set under a fixed retrieval budget.

Retrieval gain $g_j = \sum_{\mathbf{v}} \underbrace{\left(1-\mathbf{m}_t^{\mathrm{ctx}}(\mathbf{v})\right)}_{\text{not yet covered}} \underbrace{\mathbf{m}_{j\to t}(\mathbf{v})}_{\text{covered by } j}$

Select $j^\star = \arg\max_{j\in\mathcal{M}_t\setminus\mathcal{H}_t} g_j\,,\ \mathcal{H}_t \leftarrow \mathcal{H}_t \cup \{j^\star\}$

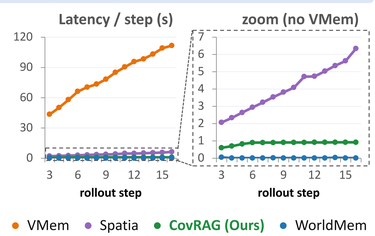
Update $\mathbf{m}_t^{\mathrm{ctx}} \leftarrow \mathbf{m}_t^{\mathrm{ctx}} \vee \mathbf{m}_{j^\star\to t}$

③ Sliding-window depth caching → scalable geometry

VGGT on the latest L frames only

$$\mathbf{d}_i = \mathrm{VGGT}\left(\mathbf{x}^0_{\mathcal{W}_t}\right), \qquad \mathcal{W}_t = \{t-L+1,\ldots,t\}$$

- Each new frame's **depth is cached** with its RGB + pose and reused for every later target view.
- **Per-step geometry cost is bounded by L** — no re-estimation over the whole memory bank as it grows.
- Retrieval is just **warp + sum** over cached depths: **no reconstruction, no alignment, no global scene**.



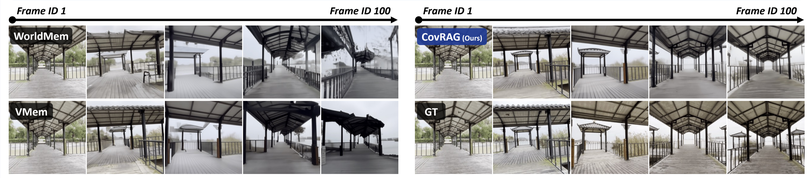
Experiments

Long-horizon consistency & quality

Dataset	Method	Consistency MEt3R ↓	LPIPS ↓	PSNR ↑	SSIM ↑	Quality FVD ↓	FID ↓
RealEstate10K (T=40)	DFoT (No memory-augmentation)	0.415	0.588	11.7	0.372	265.03	29.88
	+ WorldMem	0.144	0.197	18.2	0.555	242.36	24.77
	+ VMem	0.120	0.174	18.6	0.568	231.77	24.05
	+ Spatia	0.140	0.204	17.9	0.575	228.68	**22.60**
	+ CovRAG (Ours)	**0.102**	**0.154**	**18.9**	**0.576**	**226.40**	23.00
DL3DV10K (T=100)	DFoT (No memory-augmentation)	0.430	0.529	14.5	0.394	706.70	88.01
	+ WorldMem	0.316	0.401	16.2	0.436	428.49	58.24
	+ VMem	0.315	0.380	16.3	0.447	394.79	57.61
	+ Spatia	0.327	0.416	15.9	0.434	439.64	56.10
	+ CovRAG (Ours)	**0.224**	**0.308**	**17.9**	**0.504**	**321.81**	**46.92**

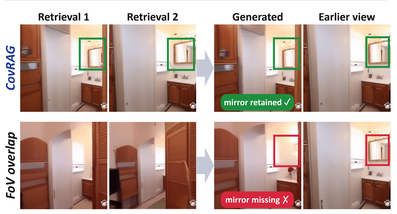
- **Only the retrieval rule differs** (same DFoT backbone & conditioning). Best MEt3R / LPIPS / PSNR / FVD on both datasets; on DL3DV10K (T = 100): MEt3R **−29 %**, FVD **−18 %** vs. the best baseline.

Qualitative — DL3DV10K 100-frame rollout



Analysis & Training-free Plug-in

What each method retrieves → generates (RE10K)



- **FoV overlap** picks frames whose view of the mirror is **occluded**; the **coverage map** picks frames that saw it.

Each retrieval covers what the previous one missed



- Grey = covered by context · **red** / **blue** = newly covered by Retrieval 1 / 2: each adds a different region.

Plug-in to pretrained world models

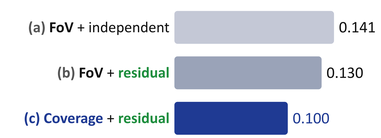
WorldPlay-8B dynamic scenes · SpatialVID

WorldPlay-8B	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
WorldPlay-8B	14.3	0.390	0.373	418.44
+ CovRAG (Ours)	**14.9**	**0.411**	**0.333**	**376.83**

GEN3C on Cosmos-7B 3D-conditioned · RE10K

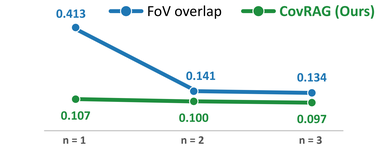
GEN3C	PSNR ↑	SSIM ↑	LPIPS ↓	MEt3R ↓
GEN3C	14.2	0.572	0.519	0.377
+ CovRAG (Ours)	**17.6**	**0.675**	**0.349**	**0.251**

Ablation (RE10K · MEt3R ↓)



- **Residual selection** helps FoV (a→b); combining it with **coverage maps** is best (c): −29 % vs (a).

Retrieval budget n (RE10K · MEt3R ↓)



- **One retrieved frame is enough**: CovRAG beats FoV overlap at every n; n = 2→3 adds little.