BankCodec: Compact Memory Banks for Multi-Turn Novel View Synthesis

Gabriele Serussi [1 2] **Chaim Baskin** [1 2]

Abstract

Multi-turn novel view synthesis generates a scene one target view at a time, making it a useful testbed for reliable long-horizon video generation. Current memory-augmented systems maintain persistent state by retrieving past views and feeding their dense tokens to the denoiser, which makes the conditioning interface grow linearly with retrieval size. We ask how small that interface can be when compression is optimized for the behavior that retrieved memory induces in the generator. We show this with **BankCodec**, which jointly reads retrieved views with learned queries and emits a fixed-rate compact bank trained directly through the generator's flow-matching loss. On the Memory-V2V benchmark, BankCodec exposes only 42 denoiser-facing memory tokens in place of the 16,380 retrieved-memory tokens used by uncompressed Memory-V2V, while preserving camera, subject, image quality, temporal flicker, and motion-smoothness metrics. It remains competitive with a 126-token local codec, while local streams retain better cross-view consistency. BankCodec thus turns a growing memory stream into a compact scene-level state and identifies the regimes where local correspondence cues remain useful.

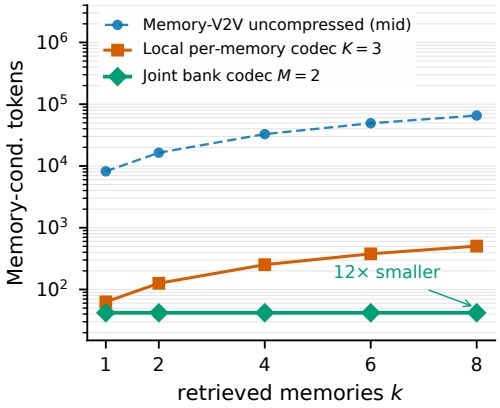

Figure 1. Retrieved-memory conditioning-token cost versus the number of retrieved memories k. BankCodec uses a fixed-rate compact bank: with bank rate $M{=}2$ and 21 latent frames in our setup, it emits 42 memory-conditioning tokens regardless of retrieval size. In contrast, uncompressed Memory-V2V and local per-memory compression both grow with the number of retrieved views.

1. Introduction

Multi-turn novel view synthesis (NVS) generates a camera path one target view at a time. Each target view is denoised independently, yet the sequence should preserve subject identity, scene layout, and camera behavior across turns. Memory-V2V (Lee et al., 2026) caches generations, retrieves views overlapping the next target camera, and feeds their tokens back as conditioning.

Retrieval answers *which* past views to use. We study the complementary question: *how small can the retrieved-memory conditioning interface be?* Dense retrieved-view tokens are expensive and scale with retrieval size, but they matter only through the behavior they induce in the generator. We therefore treat compression as a behavioral problem: the compact bank should make the generator respond as if it had seen the retrieved memories, without carrying their dense token sequences.

This makes multi-turn NVS a controlled testbed for a broader long-horizon video question: what persistent state should a generator carry forward under a fixed memory budget? The camera trajectory gives an explicit control signal, the retrieved views provide the persistent state, and failures appear as camera drift, identity drift, or cross-view inconsistency. We introduce BankCodec, a compressor that maps retrieved views into a fixed-rate compact bank whose rate directly controls how many memory-conditioning tokens the generator sees. For each latent frame, learned queries jointly read all retrieved memories; the resulting bank is trained through the generator's rectified flow-matching objective (Liu et al., 2023). This joint read exploits redundancy across memories that observe the same subject, appearance,

[1]INSIGHT Lab, Ben-Gurion University of the Negev, Israel [2]Decart.ai, Israel. Correspondence to: Gabriele Serussi <serussig@post.bgu.ac.il>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

and scene layout, representing shared state once rather than repeating it in separate memory streams. We evaluate how far this interface can be compressed, where it matches a learned local per-memory codec, and where preserving locality remains valuable.

Our contributions can be summarized as follows:

- **Behavioral compression.** We study retrieved-memory compression as preserving the denoising behavior induced by retrieved views.
- **Compact banks.** We introduce BankCodec, which jointly reads all retrieved views into a fixed-rate bank and trains directly through the generator's flow-matching loss.
- **Aggressive token reduction.** We show that 42 denoiser-facing BankCodec tokens retain the camera, subject, image quality, temporal flicker, and motion-smoothness behavior of the 16,380 retrieved-memory tokens used by uncompressed Memory-V2V, while remaining competitive with a 126-token learned local codec.
- **Joint-local tradeoff.** We identify where compact scene-level banks are sufficient and where local correspondence cues still matter: local per-memory compression retains an advantage in pairwise cross-view consistency.

2. BankCodec

At one generation turn, Memory-V2V retrieves k previously generated views whose cameras overlap the target view, tokenizes them, appends their tokens to the conditioning sequence, denoises the new target view, and stores it back in memory. BankCodec encodes the retrieved-memory tokens into a compact bank and appends that bank through Memory-V2V's token-conditioning interface.

Let $X_i \in \mathbb{R}^{F\cdot S\times D}$ be the tokenized i-th memory, with F latent frames, S spatial tokens per frame, and hidden width D; $X_{i,f}$ denotes frame f. BankCodec has a learned query bank $Q^B \in \mathbb{R}^{M\times D}$, shared across samples and applied independently at each latent frame. For frame f, it concatenates the k retrieved memories and uses Q^B to read them through cross-attention:

$$\tilde{X}_f = [X_{1,f}; \ldots; X_{k,f}], \qquad \tilde{X}_f \in \mathbb{R}^{(kS)\times D}, \tag{1}$$

$$B_f = \mathrm{Resampler}(\tilde{X}_f; Q^B), \qquad B_f \in \mathbb{R}^{M\times D}, \tag{2}$$

$$B = [B_1, \ldots, B_F] \in \mathbb{R}^{F\cdot M\times D}. \tag{3}$$

Here Q supplies the cross-attention queries, while X supplies keys and values:

$$\mathrm{Resampler}(X; Q) = \mathrm{FFN}(Q + \mathrm{CrossAttn}(Q, X, X)).$$

The bank rate is M tokens per latent frame, so memory-conditioning cost is $F \cdot M$, independent of k. The joint read represents subject and scene structure shared across retrieved views while preserving the temporal frame layout expected downstream (Figure 2). Because the learned queries read all retrieved memories for a frame at once, they must allocate a fixed token budget across redundant and complementary evidence. Shared appearance and layout can be represented once in the bank, while view-specific details compete for the same limited interface.

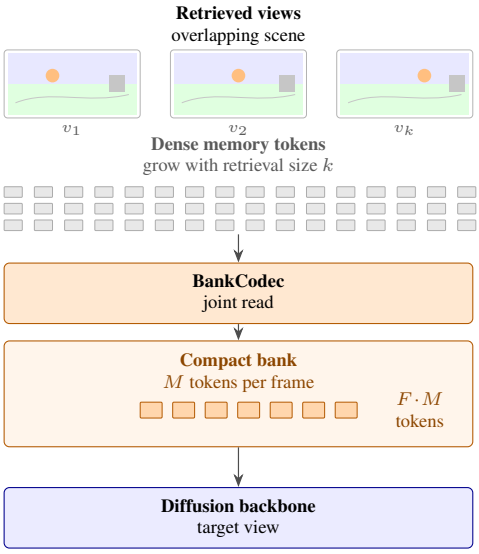

Figure 2. BankCodec reads dense tokens from retrieved views and exposes a fixed-rate compact bank to the diffusion backbone. The generator-facing memory size is $F \cdot M$, independent of retrieval size.

BankCodec returns B in the layout expected by the existing token-conditioning interface. We optimize the bank encoder end-to-end with the base generator's rectified flow-matching objective (Liu et al., 2023). With clean target latent x_0, noise $\varepsilon \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(0, 1)$, and $x_t = (1-t)\varepsilon + tx_0$, the Memory-V2V/Wan convention predicts $u^\star = \varepsilon - x_0$. For non-memory conditioning c and bank tokens $B_\phi(\tilde{X})$,

$$\mathcal{L}_{\text{flow}}(\theta, \phi) = \mathbb{E}_{x_0,\varepsilon,t}\|u_\theta(x_t, t; c, B_\phi(\tilde{X})) - (\varepsilon - x_0)\|_2^2,$$

so the compact bank is optimized through the same denoising objective that consumes its memory state at generation time.

3. Experiments

We evaluate on the Memory-V2V (Lee et al., 2026) multi-turn NVS benchmark: 40 held-out clips, three target trajectories per clip, and k=2 retrieved memories unless stated otherwise. Metrics are VGGT camera error (Wang et al., 2025), MEt3R multi-view consistency (Asim et al., 2025), and VBench-style subject, image, flicker, and motion scores (Huang et al., 2024). Token counts in the tables refer to retrieved-memory conditioning tokens consumed by the denoiser under each pipeline; they exclude target-view, text, camera, and any pre-compression tokens read

Table 1. Main Memory-V2V benchmark, k=2. BankCodec M=2 and the local codec K=1 form the matched-budget comparison at 42 denoiser-side memory-conditioning tokens; local K=3 is the stronger learned local baseline at 126 tokens. Numerical bolds mark column-best means. Method-level bootstrap CIs are reported in Appendix C.

Method	Tok.	MEt3R↑	Rot.↓	Trans.↓	Subj.↑	Img.↑	Flick.↑	Smooth.↑
Memory-V2V	16,380	0.374	1.76	0.064	0.924	0.645	**0.977**	0.992
Local K=3	126	0.376	1.60	0.063	**0.927**	0.646	**0.977**	**0.993**
Local K=1	**42**	**0.379**	1.62	0.065	**0.927**	0.646	**0.977**	0.992
BankCodec M=2	**42**	0.362	**1.58**	**0.062**	0.926	**0.647**	**0.977**	**0.993**
BankCodec M=4	84	0.362	1.60	**0.062**	0.926	0.646	**0.977**	**0.993**

by BankCodec or local-codec encoders. The main learned comparator is a local per-memory codec that applies the same learned-query compression independently to each retrieved memory, emitting $F \cdot K \cdot k$ denoiser-side tokens. We report method-level bootstrap confidence intervals over the 40 clips (N=10,000 resamples) as an uncertainty audit in Appendix C. The benchmark, retrieval, training, and metric protocols are specified in Appendix D.7, Appendix D.5, and Appendix D.6.

This study is designed to isolate four decisions about compact memory: how far the memory interface can shrink, what is lost by making memory joint rather than local, which bank rate is sufficient, and whether the measured behavior remains stable as retrieval grows.

- **RQ1: Compression limit.** How much can retrieved-memory conditioning shrink while preserving downstream NVS behavior?
- **RQ2: Joint vs. local memory.** What is gained and lost by replacing local memory streams with one joint bank?
- **RQ3: Bank rate.** What is the smallest reliable bank rate?
- **RQ4: Retrieval scaling.** Do BankCodec scores remain stable as retrieval grows?

3.1. RQ1: How Much Can We Shrink Retrieved-Memory Conditioning?

Table 1 shows that, under the same retrieval and evaluation protocol, BankCodec M=2 reduces generator-facing retrieved-memory conditioning from the 16,380 tokens used by uncompressed Memory-V2V to 42 bank tokens while retaining the measured camera, subject, image quality, flicker, and motion-smoothness behavior. The metric that changes most clearly is MEt3R, where the compact bank is below both uncompressed Memory-V2V and the local codecs.

Retrieved-memory conditioning is highly compressible when it is learned for the behavior induced in the generator. In this benchmark, the interface shrinks to 42 denoiser-side tokens without losing the measured camera and perceptual/temporal behavior; pairwise cross-view consistency is the visible boundary. Method-level uncertainty is reported in Appendix C; the token layout and evaluation protocol are in Appendix D.2 and Appendix D.7.

> **Key insight 1**
>
> Retrieved-memory tokens are highly compressible as *generator state*: 42 bank tokens preserve the measured camera and perceptual/temporal behavior, while MEt3R reveals the remaining cost in pairwise cross-view consistency.

3.2. RQ2: What Changes When the Bank Is Joint Instead of Local?

Table 2 shows that, at a matched 42-token budget, BankCodec and the local per-memory codec remain close on rotation, translation, subject consistency, image quality, temporal flicker, and motion smoothness. The consistent difference is MEt3R: the local codec is higher in both the end-to-end and frozen-backbone regimes. Table 1 shows the same pattern when the local codec is given a larger 126-token budget.

Table 2. Matched-budget comparison at 42 memory-conditioning tokens. The frozen regime keeps the Memory-V2V backbone fixed and trains only the compressor. Method-level uncertainty for the end-to-end benchmark rows is reported in Appendix C.

Regime	Method	Tokens	MEt3R↑	RotErr° ↓	TransErr↓
End-to-end	Local K=1	42	**0.379**	1.62	0.065
End-to-end	BankCodec M=2	42	0.362	**1.58**	**0.062**
Frozen	Local K=1	42	**0.378**	1.70	0.067
Frozen	BankCodec M=2	42	0.356	**1.66**	**0.061**

The joint bank is a compact scene-level state: it represents shared information once and keeps generator-facing cost fixed. The local codec keeps separate streams for each retrieved memory, preserving more pairwise correspondence and improving the consistency measured by MEt3R. Compact banks are attractive when fixed generator-side memory is the priority, while local streams remain valuable when correspondence is the bottleneck. The frozen-backbone control is detailed in Appendix E; the local codec is specified in Appendix D.4.

> **Key insight 2**
>
> Joint banks and local streams preserve different information: the joint bank stores shared scene state at fixed cost, while local streams keep pairwise correspondence cues that improve MEt3R, even with the backbone frozen.

Table 3. Bank-rate selection by coarse-teacher drift. $M{=}2$ is the smallest point on the saturated part of the curve: it matches $M{=}4$ while using half as many denoiser-side memory-conditioning tokens. The full sweep protocol is in Appendix F.

M	Tokens	Cosine↑	Rel. error↓
1	21	0.992	0.127
2	42	**0.994**	**0.108**
4	84	**0.994**	**0.108**
8	168	0.993	0.118
12	252	0.993	0.116
16	336	0.982	0.180

3.3. RQ3: What Is the Smallest Reliable Bank Rate?

Table 3 shows the rate-selection evidence for choosing the smallest bank that preserves the downstream behavior from RQ1 and lies on the saturated part of the coarse-teacher drift curve. $M{=}1$ is below that point, while $M{=}2$ and $M{=}4$ reach the best teacher-relative drift values; larger banks spend additional tokens without improving teacher tracking under this training recipe.

$M{=}2$ is the smallest reliable operating point. $M{=}1$ is useful as a lower-rate stress test, but it loses broader behavioral signal; larger banks mostly spend additional tokens without improving the measured NVS behavior under this training recipe. Two supporting diagnostics test whether this 42-token state is merely under-optimized or actually near a useful rate. Per-sample test-time optimization asks whether direct optimization of the bank tokens can find a better state for each example; it reduces flow-matching loss by only 0.07%. Linear probes ask whether the compact state still carries scene information; scene identity, focal group, and camera position remain recoverable from the bank. The rate sweep, test-time optimization, probing, and frozen lower-bound protocols are detailed in Appendix F, Appendix G, Appendix H, and Appendix E.

> **Key insight 3**
>
> $M{=}2$ is the smallest reliable bank rate: $M{=}1$ loses behavioral signal, while larger banks match the coarse teacher no better and do not improve the measured NVS metrics enough to justify extra tokens.

3.4. RQ4: Do BankCodec Scores Stay Stable as Retrieval Grows?

Table 4 asks whether BankCodec's scores change as the retriever supplies more memories. Across $k \in \{2, 4, 6, 8\}$, including $k{=}8$ beyond the $k \in [2, 6]$ training range, BankCodec stays nearly flat: MEt3R remains 0.361–0.362, RotErr remains 1.58–1.60°, and TransErr remains 0.062–0.065. The local codec is also stable and remains higher on MEt3R and slightly lower on the TransErr range, consistent with the correspondence advantage from RQ2. BankCodec is better on the RotErr range while exposing the same 42 generator-facing memory tokens throughout the sweep.

Table 4. Retrieval-scaling summary over $k \in \{2, 4, 6, 8\}$. Brackets show the observed metric range across the sweep. BankCodec's scores stay nearly flat as retrieval grows; local remains higher on MEt3R and slightly lower on TransErr. Full per-k rows are in Appendix I.

Method	Tokens, $k{=}2{\to}8$	MEt3R↑	RotErr° ↓	TransErr↓
Local $K{=}3$	$126 \to 504$	$[\mathbf{0.376}, \mathbf{0.377}]$	$[1.60, 1.63]$	$[\mathbf{0.061}, \mathbf{0.063}]$
BankCodec $M{=}2$	$\mathbf{42} \to \mathbf{42}$	$[0.361, 0.362]$	$[\mathbf{1.58}, \mathbf{1.60}]$	$[0.062, 0.065]$

The result is not that more retrieved memories improve the compact bank, but that they do not destabilize it. The same local-vs-joint tradeoff remains visible throughout the sweep: local streams preserve stronger pairwise consistency, while the bank preserves stable camera behavior with a fixed denoiser-side interface. Full per-k rows and the scaling protocol are in Appendix I.

> **Key insight 4**
>
> BankCodec's scores are stable as retrieval grows: increasing k from 2 to 8 leaves MEt3R, RotErr, and TransErr nearly flat, while the local codec preserves its MEt3R/TransErr advantage at linearly growing token cost.

4. Conclusion

BankCodec shows how small retrieved-memory conditioning can become in this setting: 42 denoiser-side bank tokens preserve the measured camera, subject, image quality, flicker, and motion-smoothness behavior of Memory-V2V's 16,380-token retrieved-memory branch, while remaining competitive with a 126-token learned local baseline and keeping cost fixed as retrieval grows. The result supports a simple design principle: retrieved memories should be compressed for the behavior they induce in the generator.

The remaining boundary is pairwise cross-view consistency: the joint bank trades some local correspondence for fixed-rate scene state, which is precisely the pattern suggested by the MEt3R gap. BankCodec is therefore most attractive when the priority is a compact persistent state with fixed generator-facing cost. Tasks that depend on explicit view-to-view matching may still need local streams, or hybrid local-global interfaces that keep a small scene-level bank while preserving selected correspondence cues.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2204.14198.

Asim, M., Wewer, C., Wimmer, T., Schiele, B., and Lenssen, J. E. MEt3R: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2501.06336.

Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., and Zhang, D. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., and Liu, Z. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2311.17982.

Lee, D., Huang, C.-H. P., Chen, X., Ye, J. C., Ceylan, D., and Jeong, H. Memory-v2v: Memory-augmented video-to-video diffusion for consistent multi-turn editing. *arXiv preprint arXiv:2601.16296*, 2026.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.03003.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2212.09748.

Wan Team, Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.11651.

A. Qualitative Examples

This section provides qualitative support for the over-provisioning thesis through paired comparisons drawn from the same Memory-V2V multi-turn NVS benchmark used for every quantitative result in the paper. Examples are selected by explicit criteria, not visual cherry-picking. The selection script and the chosen clip ids are saved alongside the benchmark outputs in `results/bankcodec/qualitative_selection.json`.

Method comparison (Figure 3). Three scenes are shown across the four interfaces compared in the paper: uncompressed Memory-V2V (16,380 tokens), local per-memory $K{=}1$ (42 tokens, matched budget), BankCodec $M{=}2$ (42 tokens, ours), and local per-memory $K{=}3$ (126 tokens, $3\times$ budget). The leftmost column is the ground-truth target view at the same trajectory and frame indices, providing a direct reference. The three scenes are sampled at the cohort median MEt3R (averaged across the four methods) so that each row is a representative scene rather than a hand-picked best case. Each cell shows two frames from the same trajectory (early, frame 10; late, frame 70) so the reader can also assess temporal persistence. The visual takeaway matches the bootstrap evidence: at 42 tokens, BankCodec preserves subject identity, object layout, and scene structure as well as the much larger uncompressed Memory-V2V interface and the $3\times$-larger local codec.

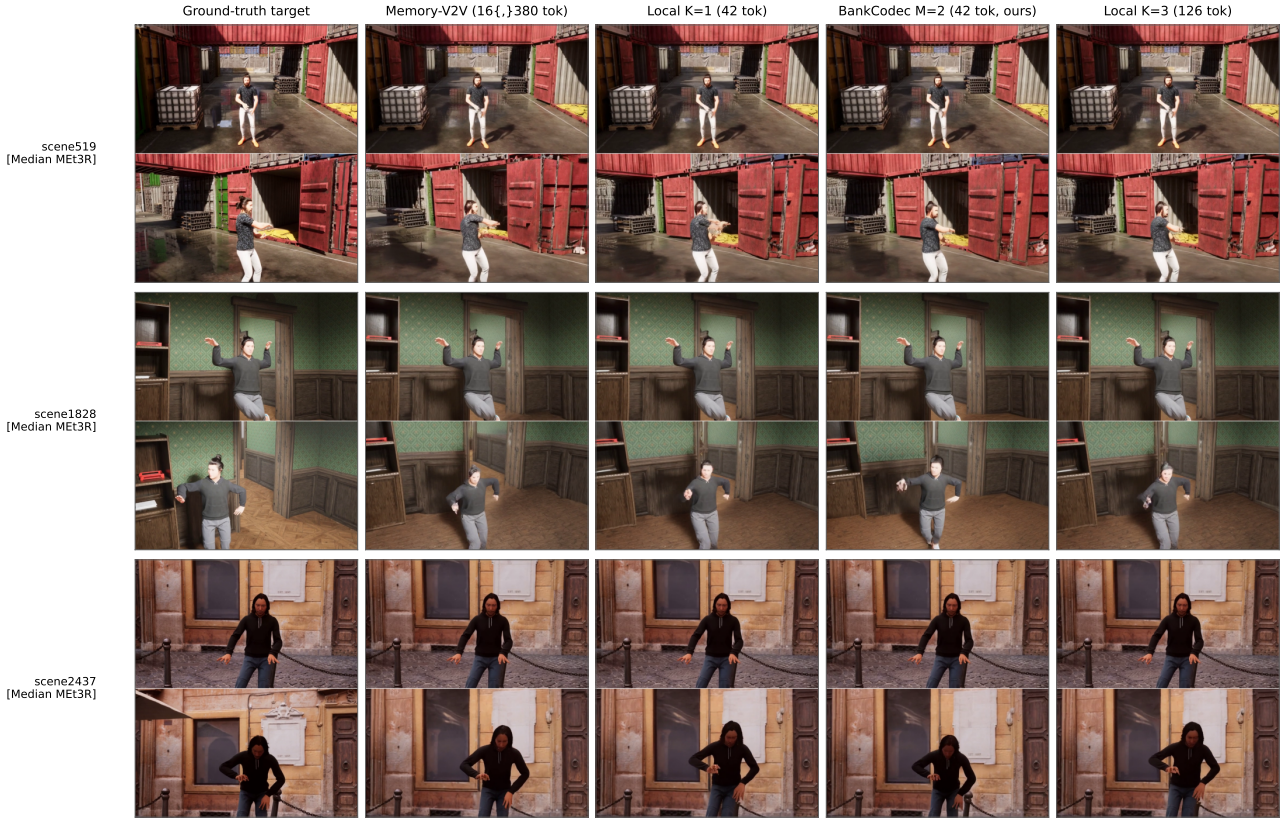

Figure 3. Method comparison at $k{=}2$ on the Memory-V2V multi-turn NVS benchmark. Each row is one held-out clip selected at the cohort median MEt3R (representative scene); columns are the ground-truth target view at the same trajectory and frame indices, then the four methods. Each cell stacks an early frame (10) above a late frame (70) from the same generated trajectory. The reader is invited to inspect: subject identity preservation across rows, object layout (truck and shipping containers in row 1; person and stage in row 2; alley structure in row 3), and temporal persistence between the early and late frame within each cell.

Bank-rate sweep (Figure 4). A second figure visualizes the M-sweep that brackets the operating point in Table 3A. We pick two scenes where BankCodec $M{=}2$ leads $M{=}1$ on the VBench-style behavioral metrics by the largest amount in the benchmark, so any visual signature of the sub-saturation $M{=}1$ regime should be visible. The columns are $M{=}1$ (21 tokens, below saturation), $M{=}2$ (42 tokens, ours), and $M{=}4$ (84 tokens, in the saturation plateau). The visual finding is consistent with the quantitative one: $M{=}1$ is plausible but a little less clean on the foreground subject than $M{=}2$, while $M{=}2$ and $M{=}4$ are visually indistinguishable, matching the drift saturation in Appendix F and the per-metric bootstrap CIs

in Appendix C. The figure supports the choice of $M{=}2$ as the smallest joint bank rate at which both behavioral metrics and visual outputs stabilize.

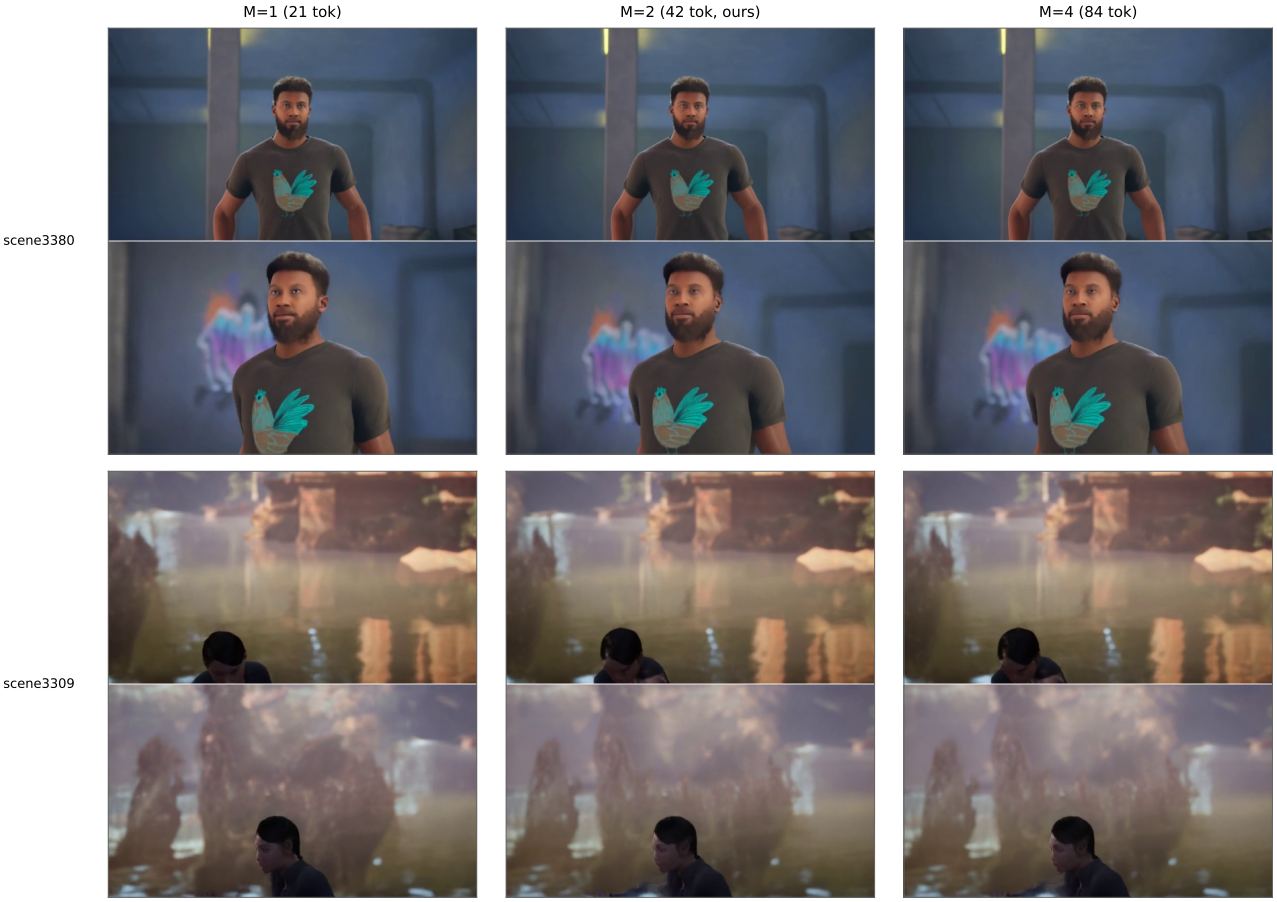

Figure 4. Bank-rate sweep: BankCodec $M{=}1$, $M{=}2$, $M{=}4$ on two rate-sensitivity clips selected by largest $M{=}2$-over-$M{=}1$ gap on the VBench-style behavioral metrics. Each cell stacks an early frame (10) above a late frame (70) from the same trajectory. The reader is invited to inspect whether identity and scene layout degrade visibly as the bank rate drops below saturation ($M{=}1$, 21 tokens) or improve as the rate doubles past saturation ($M{=}4$, 84 tokens). The visual signature matches the drift bracket in Table 3A: $M{=}1$ is plausible but cleaner at $M{=}2$, while $M{=}2$ and $M{=}4$ are visually indistinguishable from each other.

Out-of-distribution generalization (Figure 5). The benchmark in the main paper uses the synthetic Memory-V2V multi-turn NVS protocol. To check that the over-provisioning behavior also holds on real-world video, we generate target views on the OpenVid-1M proxy benchmark with the same checkpoints used throughout the paper. OpenVid clips are single-camera — no second-camera ground-truth view exists — so the comparison is method-vs-method on the source video. *The source-video column and the generated columns are at different camera trajectories by construction:* the source column shows the input clip at its own original camera, while each generated column is the corresponding method's output at a novel target trajectory chosen by the benchmark protocol. The framing and subject position therefore differ between the source and the generated columns. The reader should compare *across the three generated columns* (which share the same target trajectory and frame indices) for the over-provisioning argument, and treat the source column as a content reference rather than a per-pixel target. We pick three clips spanning content categories (human / scene / object) at the highest aesthetic-score within each category to ensure visual quality, with no selection by output appearance. The columns are the source video, then the three interfaces directly relevant to the over-provisioning thesis: uncompressed Memory-V2V (16,380 tokens), BankCodec $M{=}2$ (42 tokens, ours), and local per-memory $K{=}3$ (126 tokens). The visual finding extends the on-distribution result: BankCodec at 42 tokens preserves subject identity, scene structure, and object geometry as well as the much larger uncompressed Memory-V2V interface and the $3\times$ larger local codec on real-world video drawn from a distribution disjoint from training.

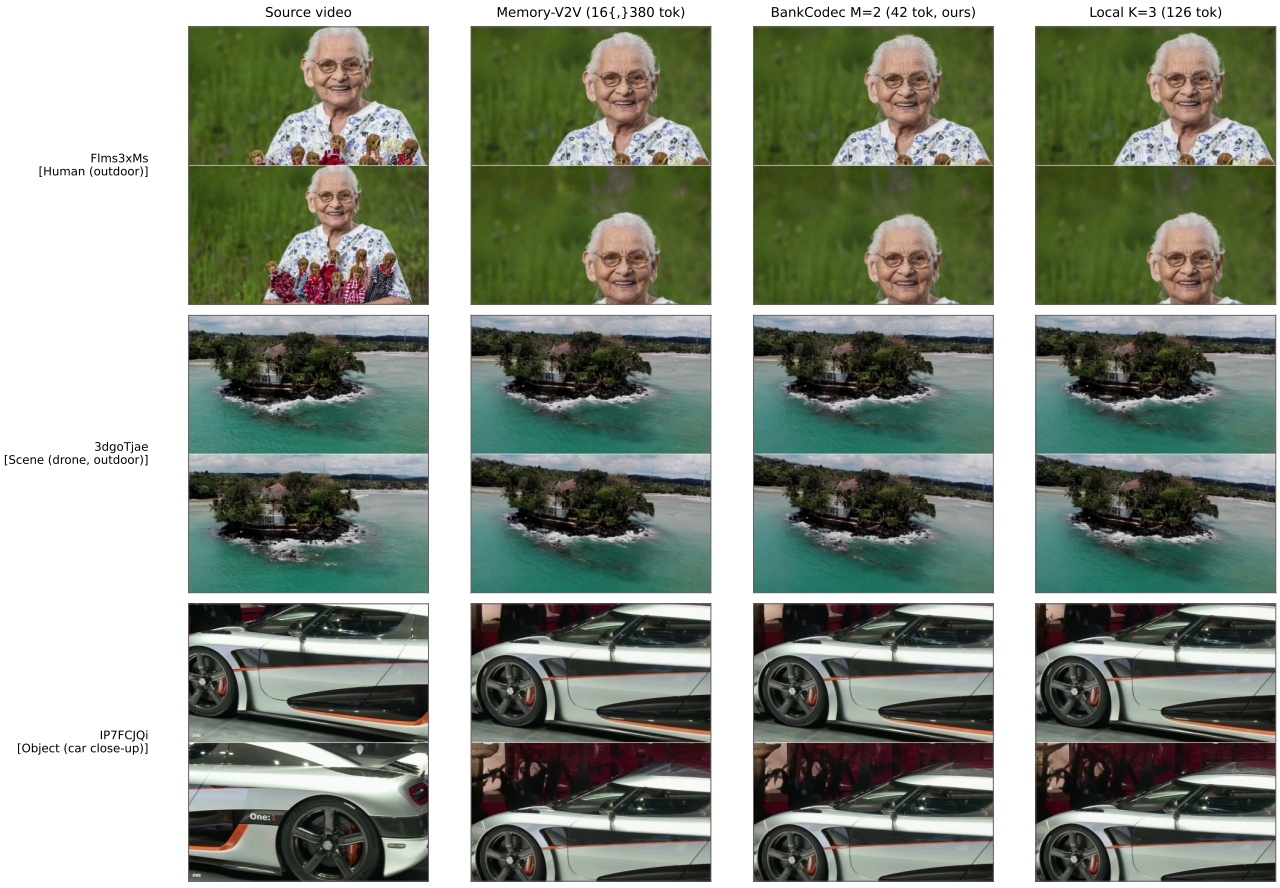

Figure 5. Out-of-distribution generalization on the OpenVid-1M proxy benchmark. Three clips were selected by content category (human, scene, object) and aesthetic score. Columns are the source video, uncompressed Memory-V2V (16,380 tokens), BankCodec $M{=}2$ (42 tokens, ours), and local per-memory $K{=}3$ (126 tokens). Each cell stacks an early frame (10) above a late frame (70) from the same generated novel-view trajectory. **The source column is at the input clip's original camera; the three generated columns share a different, common target trajectory — the framing and subject position therefore differ between source and generated columns by design.** OpenVid clips have no second-camera ground-truth view, so the per-pixel comparison is across the three generated columns; the source column is a content reference. The reader is invited to inspect: identity preservation in row 1 (face, expression), scene structure in row 2 (island, water, vegetation), and object geometry in row 3 (car body, wheel, panel detail).

Multi-turn coherence (Figure 6). The previous figures isolate behavior *within a single turn*. The multi-turn protocol generates three sequential target trajectories per clip, where each later target retrieves the earlier ones as memory; the MEt3R measurement explicitly compares the three target views against each other. Figure 6 therefore unfolds those three turns visually: each row is a single (clip, method) pair, and the three columns are turn 1, turn 2, and turn 3 from the multi-turn generation. Two clips are each shown twice — once for BankCodec $M{=}2$ (42 tokens, ours) and once for local per-memory $K{=}3$ (126 tokens, $3\times$ budget) — so paired rows enable a direct cross-turn comparison of the two interfaces on the same scene. The first clip (*scene519*, truck and shipping containers) is one of the median clips already used in Figure 3; the second clip (*scene2399*, cart on a city street) is the correspondence-stress clip with the largest local-$K{=}3$-over-BankCodec MEt3R lead in the cohort, included here so the structural MEt3R cost has a visible home in the figure. Each cell additionally stacks an early frame (10) above a late frame (70) of that turn, so within-turn temporal behavior is also visible. The multi-turn read of the figure is consistent with the bootstrap evidence of Appendix C: BankCodec maintains coherent subject identity, scene layout, and camera behavior across the three sequential turns; the local codec is similarly coherent and retains a small, statistically real lead on pairwise cross-view consistency that is difficult to see at this resolution but is what the $\sim$0.02 matched-budget MEt3R gap measures.

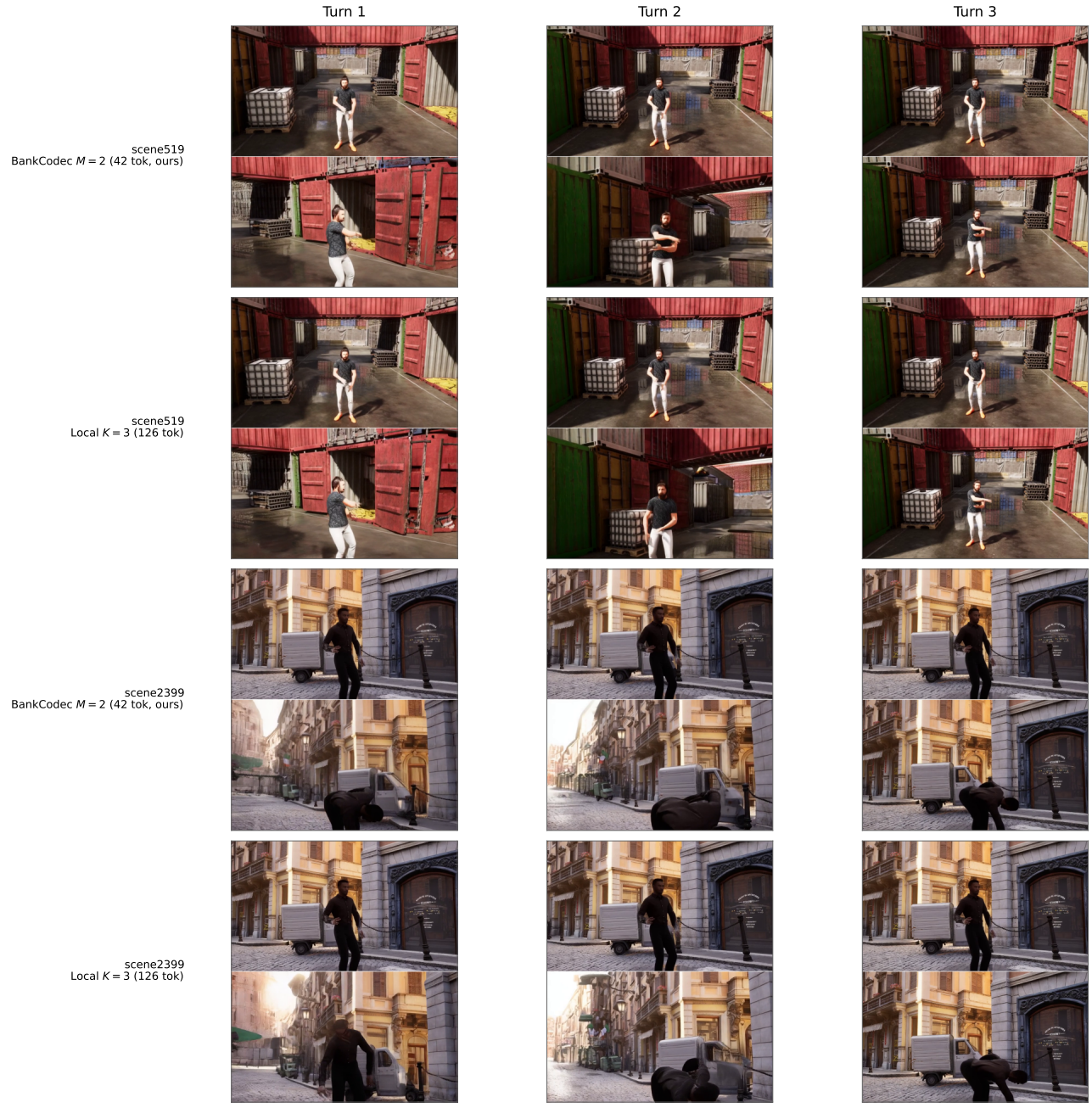

Figure 6. Multi-turn coherence. Each row is a single (clip, method) pair; columns are the three sequential target trajectories produced by the multi-turn protocol (turn 1, turn 2, turn 3). Rows 1 and 2 pair BankCodec $M{=}2$ and local $K{=}3$ on the median-MEt3R clip *scene519* (truck and shipping containers); rows 3 and 4 pair the same two interfaces on the correspondence-stress clip *scene2399* (cart on a city street, the clip with the largest local-$K{=}3$-over-BankCodec MEt3R lead in the cohort). Each cell stacks an early frame (10) above a late frame (70) from that turn so within-turn temporal behavior is also visible. The reader is invited to compare *across columns within one row* for the multi-turn-coherence reading (does subject identity and scene layout persist as the camera trajectory changes between turns?) and *between paired rows* for the BankCodec-vs-local reading (does either interface maintain better cross-turn consistency on this clip?). Both methods maintain coherent multi-turn outputs; the small MEt3R lead the local codec retains is the structural cost named in the main text.

B. Roadmap

Table 5. Evidence-to-protocol map. Each appendix link points to the specific section needed to reproduce or audit the corresponding main-text result.

Main evidence	Role in the paper	Protocol and supporting details
Table 1	Main k=2 benchmark	Appendix D.7; retrieval in Appendix D.5; training in Appendix D.6
Table 2	Matched-budget local-vs-joint tradeoff	Frozen protocol in Appendix E; local codec in Appendix D.4
Table 3	Bank-rate saturation	Drift protocol and full sweep in Appendix F
RQ3 prose	Test-time optimization slack	TTO protocol and plots in Appendix G
RQ3 prose	Recoverable scene structure	Probe construction and controls in Appendix H
Table 4	Retrieval scaling	Full per-k rows in Appendix I
Table 1 rows	Uncertainty audit	Method-level bootstrap CIs in Appendix C
Figure 4	Qualitative support	Real-world proxy examples in Appendix A

C. Bootstrap Confidence Intervals for Method Means

For the Table 1 rows and the BankCodec M=1 lower-rate boundary used in RQ3, each interval in Table 6 resamples the 40 benchmark clips with replacement (10,000 resamples) and reports the mean with its percentile 95% CI.

Table 6. Method-level bootstrap uncertainty for the main k=2 benchmark. Cells report mean [95% CI] over the same 40 clips used in Table 1; lower is better for rotation and translation, higher is better otherwise.

Method	N	MEt3R↑	Rot.↓	Trans.↓	Subj.↑	Img.↑	Flick.↑	Smooth.↑
Memory-V2V	40	0.374 [0.349, 0.400]	1.76 [1.53, 2.01]	0.064 [0.049, 0.081]	0.924 [0.914, 0.934]	0.645 [0.603, 0.681]	**0.977** [0.973, 0.980]	0.992 [0.991, 0.993]
Local K=3	40	0.376 [0.351, 0.402]	1.60 [1.42, 1.80]	**0.062** [0.050, 0.076]	**0.927** [0.918, 0.935]	0.646 [0.604, 0.682]	**0.977** [0.973, 0.980]	**0.993** [0.992, 0.993]
Local K=1	40	**0.379** [0.353, 0.405]	1.62 [1.44, 1.81]	0.064 [0.051, 0.079]	**0.927** [0.918, 0.936]	0.646 [0.603, 0.684]	**0.977** [0.973, 0.980]	0.992 [0.991, 0.993]
BankCodec M=1	40	0.365 [0.339, 0.391]	**1.56** [1.34, 1.78]	**0.062** [0.048, 0.079]	0.922 [0.912, 0.932]	0.643 [0.601, 0.681]	0.976 [0.972, 0.980]	**0.993** [0.991, 0.993]
BankCodec M=2	40	0.362 [0.337, 0.386]	1.58 [1.39, 1.77]	**0.062** [0.049, 0.076]	0.926 [0.916, 0.935]	**0.647** [0.606, 0.683]	**0.977** [0.973, 0.980]	**0.993** [0.992, 0.994]
BankCodec M=4	40	0.362 [0.337, 0.388]	1.60 [1.41, 1.80]	**0.062** [0.049, 0.076]	0.926 [0.916, 0.935]	0.646 [0.606, 0.683]	**0.977** [0.973, 0.980]	**0.993** [0.992, 0.994]

D. Reproducibility Details

D.1. Base Generator

The base generator is Wan2.1-T2V-1.3B (Wan Team et al., 2025), a 1.3B-parameter diffusion transformer (Peebles & Xie, 2023) with 30 blocks and hidden dimension 1,536. We use 12 attention heads with head dimension 128, patch size $(1, 2, 2)$, and 16 input channels from the pre-encoded VAE latent space. The backbone is extended with ReCamMaster-style camera conditioning (Bai et al., 2025): per-frame camera poses are projected to a per-frame embedding and injected inside each block at a dedicated cross-attention, rather than being added to the token stream once at the input. We inherit this injection scheme unchanged from the pretrained checkpoint.

We initialize BankCodec and the local per-memory codec from a shared Memory-V2V (Lee et al., 2026) checkpoint trained with a per-frame Perceiver Resampler over each retrieved memory, rather than from the vanilla Wan2.1-T2V weights, so that both learned codecs start training from the same point. The initialization checkpoint is the same across all BankCodec variants reported in the main paper and in this appendix.

D.2. Tokenizer and Token Layout

Memory videos enter BankCodec through the Memory-V2V dynamic tokenizer (Lee et al., 2026) at the `coarse` level, which is a 3D convolution with kernel $(1, 8, 8)$ applied to the pre-encoded VAE latents. For an 81-frame 480×832 input, the VAE produces $ppf = 21$ latent frames and the coarse tokenizer produces a $(pph, ppw) = (8, 13)$ spatial grid per latent frame, for $S = pph \cdot ppw = 104$ tokens per memory per latent frame. Each memory therefore contributes $21 \times 104 = 2{,}184$ coarse tokens to the bank.

BankCodec does not touch the user-conditioning branch: user-frame tokens, camera embeddings for the target trajectory, and text embeddings are produced by the base generator exactly as in the pretrained checkpoint. The intervention is confined to the retrieved memory branch.

Table 7. BankCodec architectural hyperparameters. All settings are held fixed across the bank-rate studies unless M is explicitly varied.

Hyperparameter	Value
Base generator	Wan2.1-T2V-1.3B (30 blocks, $d = 1{,}536$)
Attention heads (base)	12 (head dim 128)
Camera conditioning	ReCamMaster (in-block)
BankCodec encoder layers L	2
BankCodec query count M per frame	2 (main); varied in rate studies
BankCodec attention heads	12
BankCodec FFN expansion	4 (GELU)
Bank-query init	$\mathcal{N}(0, 0.02^2)$
Tokenizer level	`coarse`, kernel $(1, 8, 8)$
Latent-frame count ppf	21
Spatial tokens per frame S	104 (pph=8, ppw=13)

D.3. BankCodec Architecture

For each latent frame f, BankCodec receives the concatenated spatial tokens from the k retrieved memories, $\tilde{X}_f \in \mathbb{R}^{(kS)\times D}$ with $D = 1{,}536$, and cross-attends M learned queries into them with a pre-norm residual block. Each of the L encoder layers applies a LayerNorm on the queries and keys/values, a multi-head cross-attention (h heads, head dimension D/h, `batch_first`), a second LayerNorm, and a two-layer FFN with expansion factor 4 and GELU activation. A final LayerNorm is applied to the output before reshaping to $(ppf \cdot M) \times D$.

We set $L = 2$ encoder layers, $h = 12$ heads (inherited from the base model), FFN expansion 4, and vary only the bank rate M in the reported capacity studies. Bank queries are initialized from $\mathcal{N}(0, 0.02^2)$. The encoder contains no time- or space-specific positional embedding; temporal and spatial position are re-introduced downstream through the base generator's RoPE when the bank tokens enter the denoiser's token-only conditioning interface.

Table 7 summarizes the BankCodec architectural hyperparameters.

D.4. Local Per-Memory Codec

The local per-memory codec is a per-frame Perceiver Resampler (Alayrac et al., 2022) applied to the coarse tokens of each retrieved memory independently. For each memory and latent frame, K learned queries cross-attend into the $S = 104$ spatial tokens and produce K tokens per frame per memory. Stacked across frames, one memory therefore contributes $ppf \cdot K = 21K$ tokens.

We report two settings:

- K=3 **(primary)**: 63 tokens per memory, 126 at k=2. This matches the baseline used in Memory-V2V (Lee et al., 2026) and is the strongest local per-memory compressor reported here.
- K=1: 21 tokens per memory, 42 at k=2. This matches BankCodec M=2 at identical total budget and is added in this submission as a strict token-matched comparison.

The resampler uses the same 2-layer cross-attention+FFN design as BankCodec (h=12, FFN expansion 4).

D.5. Retrieval

We retrieve memory videos by field-of-view (FOV) overlap on a discretized sphere (Lee et al., 2026). For each past clip, we rasterize the per-frame camera FOV onto a $(n_\theta, n_\phi) = (180, 360)$ grid of the unit sphere, then score candidate memories by a weighted combination of overlap area with the target trajectory ($\lambda = 0.5$) and coverage distance. We retrieve top-k memories per generation call. Evaluation uses $k = 2$ in the main benchmark unless stated otherwise.

D.6. Training

BankCodec and the local per-memory codec are trained end-to-end on the same data with rectified flow matching (Liu et al., 2023). We use the Memory-V2V/Wan convention where t=0 is noise and t=1 is clean: $x_t = (1-t)\varepsilon + tx_0$ with $t \sim \mathcal{U}(0, 1)$.

Table 8. BankCodec training-time hyperparameters. All BankCodec variants in the main paper and in this appendix share this recipe; only the bank rate M changes where stated.

Hyperparameter	Value
Objective	Rectified flow matching (Liu et al., 2023)
Optimizer	AdamW, weight decay 0.01
LR (base unfrozen)	1×10^{-5}
LR (new modules)	1×10^{-4}
LR schedule	linear warmup 200 + cosine to 1×10^{-7}
Gradient clipping	1.0 (global norm)
Micro-batch	1
Gradient accumulation	4
GPUs	8
Effective batch	32
Precision	bf16 mixed
Gradient checkpointing	On (base generator)
Total steps	2,000
RoPE dropout	0.1
Memory token noise σ	0.05

The model prediction target is the Wan flow output $u^\star = \varepsilon - x_0$, which is the negative of the path derivative under this time parameterization. We use AdamW with two parameter groups: unfrozen base-model parameters (self-attention, FFN, norms) at learning rate 1×10^{-5} and new modules (BankCodec encoder, tokenizer heads, camera embedder) at learning rate 1×10^{-4}, both with weight decay 0.01 and gradient-norm clipping at 1.0. The schedule is a linear warmup for 200 steps followed by cosine annealing to $\eta_{\min} = 1\times10^{-7}$ over the remaining 1,800 steps.

The effective batch size is 32: micro-batch 1, gradient accumulation 4, 8 GPUs per node. We train in bf16 mixed precision with gradient checkpointing enabled on the base generator. Each run is 2,000 optimizer steps ($\sim$64,000 scene-trajectory pairs), matching the Memory-V2V NVS recipe. RoPE dropout 0.1 and memory noise σ=0.05 on the retrieved token stream are active throughout training; the adaptive token merger is disabled for BankCodec (`merger.enable_prob` = 0) because the bank encoder already compresses the memory branch.

Table 8 summarizes the training-time hyperparameters.

D.7. Evaluation

We evaluate on the Memory-V2V multi-turn NVS benchmark: 40 held-out clips, each with three target trajectories, generated with k=2 retrieved memories per call. We report the following metrics:

- **Camera pose accuracy.** We run VGGT on the generated clip to estimate per-frame camera rotation and translation, average over frames, and compute (i) angular rotation error in degrees against the target trajectory (RotErr) and (ii) scale- normalised translation error in scene units (TransErr).

- **Multi-view consistency.** We compute MEt3R (Asim et al., 2025) over the three pairings of target trajectories per clip (views 1↔2, 1↔3, 2↔3), then average. MEt3R is a reference-free pairwise consistency score.

- **VBench-style behavioral metrics.** From VBench (Huang et al., 2024) we use subject consistency, image quality, temporal flicker, and motion smoothness, each averaged over the generated clip.

For the capacity sweep in the main paper and in Section F we additionally report *drift* of the joint-bank-conditioned denoiser output against the independently encoded coarse-token teacher on held-out training scenes, measured as cosine similarity and relative ℓ_2 error on velocity predictions. For qualitative comparisons (Appendix A) we use a separate 40-clip proxy benchmark drawn from OpenVid-1M; the proxy benchmark is used only for qualitative inspection and not for the quantitative numbers in Table 1.

D.8. Compute

Each 2,000-step training run occupies 8 GPUs for approximately 24 hours wall-clock. Evaluation of a single method on the full 40×3 benchmark runs in $\sim$4 hours on 8 GPUs including VBench, VGGT, and MEt3R.

E. Controlled Local-vs-Joint Comparison

The main comparison in Table 1 uses end-to-end-trained wrapper checkpoints: BankCodec is trained jointly with the Memory-V2V backbone's self-attention, and the local per-memory codec $K{=}3$ is trained jointly with its own self-attention. The frozen-backbone control isolates the memory interface from this possible source of co-adaptation.

To remove this confound, we train two additional local per-memory variants and a matched BankCodec variant that share an identical *frozen* Memory-V2V backbone. Only the compressor (a $K{=}1$ or $K{=}3$ Perceiver Resampler for the local codec, the compact bank encoder for BankCodec) is trainable; all $\sim$1.6B wrapper parameters (self-attention, FFN, layer-norms, tokenizer, camera encoder, ReCamMaster injection layers) are frozen at their Memory-V2V values. The recipe is otherwise identical to the main runs (Appendix D.6): rectified flow matching, 2,000 steps, micro-batch 1 with gradient accumulation 4 on 8 GPUs, bf16. We pin the local codec's stochastic route probability to 1.0 and the minimum memory count to 2 so every training step has a non-empty trainable graph; this distribution change is necessary because under a frozen backbone any batch that bypasses the resampler produces a loss with no gradient path.

Table 9. Frozen-wrapper ablation. The frozen rows are the controlled cells: the Memory-V2V backbone is fixed and only the compressor is trained. The matched-budget MEt3R gap persists in both end-to-end and frozen regimes, while camera means remain close. The $M{=}1$ rows provide the lower-rate boundary used in RQ3.

Method	Backbone	Tokens	MEt3R↑	RotErr° ↓	TransErr↓	SubjCons↑
End-to-end (own backbone fine-tune):						
Local codec $K{=}3$	MV2V (e2e)	126	0.376	1.60	0.063	**0.927**
Local codec $K{=}1$	MV2V (e2e)	42	**0.379**	1.62	0.065	**0.927**
BankCodec $M{=}1$	bankcodec_m1 (e2e)	**21**	0.365	**1.56**	**0.062**	0.922
BankCodec $M{=}2$	bankcodec_m2 (e2e)	42	0.362	1.58	**0.062**	0.926
Frozen Memory-V2V backbone, compressor-only training:						
Local codec $K{=}3$	MV2V (frozen)	126	0.377	1.75	0.066	0.924
Local codec $K{=}1$	MV2V (frozen)	42	**0.378**	1.70	0.067	**0.928**
BankCodec $M{=}1$	MV2V (frozen)	**21**	0.364	1.68	**0.061**	0.924
BankCodec $M{=}2$	MV2V (frozen)	42	0.356	**1.66**	**0.061**	0.926

Table 2 uses these end-to-end and frozen rows for the matched-budget local-vs-joint comparison.

F. Extended Capacity Analysis

$M \in \{1, 2, 4, 8, 12, 16\}$ brackets the rate selected in Table 3: lowering the rate to $M{=}1$ worsens teacher tracking, and raising the rate above $M{=}2$ does not improve it under this training recipe (with the highest rate tested, $M{=}16$, becoming less stable).

Full M-sweep. Drift is measured on held-out training scenes as cosine similarity and relative ℓ_2 error on velocity predictions of the joint-bank-conditioned denoiser output against the independently encoded coarse-token teacher (Table 10).

Below saturation ($M{=}1$, 21 tokens): drift cosine drops to 0.992 and relative error rises to 0.127; the bank no longer carries enough capacity to track the uncompressed teacher. *Saturation ($M \in \{2, 4\}$):* drift stalls at cosine 0.994 / relative error 0.108, and doubling the bank does not improve teacher tracking. *Plateau with mild degradation ($M \in \{8, 12\}$):* cosine drops marginally to 0.993 and relative error rises to $\sim$0.117; useful shared structure is captured, but added capacity is spent on degrees of freedom the downstream denoiser does not use. *High-rate degradation ($M{=}16$):* relative error rises to 0.180, showing that higher bank capacity does not by itself improve teacher tracking.

The saturation between $M{=}2$ and $M{=}4$ is also visible in the downstream benchmark (Table 1): BankCodec $M{=}4$ matches $M{=}2$ on every reported metric, including MEt3R (both 0.362); the method-level uncertainty is reported in Appendix C. The $M{=}1$ point reported in Table 9 carries this sub-saturation drift through to behavior: relative to $M{=}2$, it has lower

Table 10. Drift of the BankCodec-conditioned denoiser velocity output against the same denoiser fed with uncompressed coarse-level tokens, on held-out training scenes. The bank-rate sweep is bracketed: $M{=}1$ sits below the saturation knee (cosine 0.992); $M{=}2$ and $M{=}4$ are at saturation (cosine 0.994); $M \in \{8, 12\}$ plateaus with mild degradation; and $M{=}16$ underperforms, showing that extra query capacity is not automatically useful under this training recipe.

M	Tokens	Cosine (↑)	Rel. error (↓)
1	21	0.992	0.127
2	42	**0.994**	**0.108**
4	84	**0.994**	**0.108**
8	168	0.993	0.118
12	252	0.993	0.116
16	336	0.982	0.180

subject consistency, image quality, and temporal flicker without a compensating MEt3R gain. The two saturations therefore agree: $M{=}2$ is the smallest bank rate at which both the velocity-level drift and the downstream behavioral metrics stabilize. Going below it costs behavioral quality without buying camera or MEt3R; going above it costs tokens without changing anything measurable. $M{=}1$ is not the controlled local-vs-joint comparison: at $k{=}2$, one local query per retrieved memory already costs $F \cdot k = 42$ tokens, so the matched-budget local codec is $K{=}1$ at 42 tokens, not $42/k = 21$. $M{=}1$ is therefore a lower-rate boundary ablation on the joint-read interface, not a budget-matched contrast.

G. Test-Time Optimization at the $M{=}2$ Bank Rate

The bank-rate sweep in Appendix F establishes that the *amortized* compressor has saturated at $M{=}2$ relative to the uncompressed coarse teacher. A remaining question is whether a better 42-token representation exists for each individual sample but is not recovered by the feedforward encoder. We estimate that residual slack by running per-sample test-time optimization (TTO) with 42 tokens as the only free variable.

Protocol. For each of the first 10 clips in the benchmark, we evaluate all 3 target trajectories, giving $N = 30$ tasks. For each task we retrieve $k = 2$ memories — the two *ground-truth* novel-view videos of the other target trajectories of the same clip. Using ground-truth memories isolates the amortization question from any compounding error introduced by sequential generation. We run BankCodec $M{=}2$ on the two memories to obtain $T_{\mathrm{bc}} \in \mathbb{R}^{1\times 42\times 1536}$, initialize $T = T_{\mathrm{bc}}.\texttt{clone}().\texttt{detach}()$ with `requires_grad=True`, freeze the wrapper and the bank encoder, and optimize T with AdamW (lr 10^{-3}, weight decay 0) for 100 steps. Each step averages the flow-matching MSE over 2 fresh random timesteps $t \sim \mathcal{U}(0, 1)$ and noises ε. We measure L_{init} (with T_{bc}) and L_{opt} (with the optimized $T^\star$) over 16 *fixed held-out seeds*. We report $\Delta = (L_{\mathrm{init}} - L_{\mathrm{opt}})/L_{\mathrm{init}}$ and $\|T^\star - T_{\mathrm{bc}}\|_2/\|T_{\mathrm{bc}}\|_2$. A single-sample convergence trace places the 100-step budget past the optimization knee (Figure 7).

Loss-level slack. Across the $N = 30$ tasks, the 16-seed deterministic protocol reduces flow-matching loss from $L_{\mathrm{init}} = 0.1601$ to $L_{\mathrm{opt}} = 0.1600$, a mean relative reduction of 0.07% (median 0.04%, std 0.07%). TTO produced a positive reduction on $30/30$ tasks; the best was 0.3% and the worst 0.01%. Figure 8 shows the per-task data. The optimized tokens stay close to the feedforward prediction: mean relative ℓ_2 displacement $\|T^\star - T_{\mathrm{bc}}\|_2/\|T_{\mathrm{bc}}\|_2 = 0.020$. TTO did not flee the BankCodec region; it refined it.

Metric confirmation. To verify that the small loss reduction translates into a small downstream metric change rather than a cheaper route to a visually different sample, we re-ran inference on 3 clips × 3 target trajectories ($N{=}9$ video pairs) with both T_{bc} and $T^\star$ at identical integrator seeds. The small loss reduction does not translate into a meaningful downstream gain (Table 11): MEt3R decreases from 0.358 to 0.351, TransErr worsens slightly (0.0395 → 0.0433), and RotErr is essentially flat (1.36 → 1.40°). The deltas are within the run-to-run noise band of the 40-clip benchmark, consistent with a near-flat local minimum of the conditioning loss.

Interpretation. Given 42 tokens and the flow-matching training objective, BankCodec is close to the local solution reached by direct per-sample optimization of the bank tokens. Combined with the teacher-drift analysis (Appendix F), this gives two complementary diagnostics: BankCodec tracks the uncompressed teacher on *teacher-aligned* velocities

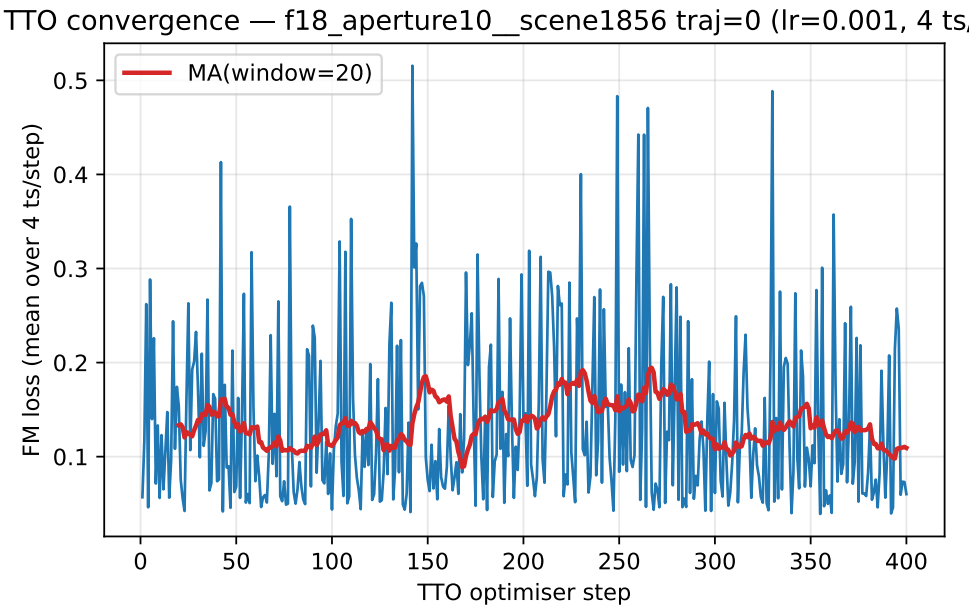

Figure 7. Convergence of per-sample TTO on one task (clip `f18_aperture10__scene1856`, trajectory 0). The 100-step budget used in the main sweep (red arrow) sits well past the knee; extending to 400 steps adds little further reduction.

Table 11. Metric confirmation on 3 clips × 3 target trajectories (N=9 video pairs), identical integrator seeds. Per-sample TTO marginally reduces the training loss but does not improve the downstream metrics in this small-N confirmation, consistent with a near-flat local minimum of the conditioning loss.

Variant	MEt3R↑	RotErr° ↓	TransErr↓
T_{bc} (BankCodec feedforward)	**0.358**	**1.36**	**0.0395**
$T^{\star}$ (per-sample TTO)	0.351	1.40	0.0433

(cosine 0.994), and direct optimization of the same 42-token interface recovers only 0.07% additional flow-matching loss on average. This is a bound on slack under a specific intervention, not a claim that no stronger test-time method could help. The optimized objective is flow-matching MSE with 2 timesteps per step; losses defined over full denoising trajectories or decoded-frame perceptual quality could expose different directions. The search is also restricted to the 42-token BankCodec interface, so interventions deeper in the conditioning stack would answer a different question. The downstream metric check is a small-N confirmation; the headline measurement is the N=30 loss-level TTO result.

H. The Compact Bank Encodes Recoverable Scene Structure

To test whether the M=2 BankCodec tokens retain recoverable scene information, we train linear probes on 50 scenes. For each scene we form 3 k=2 memory configurations (two for probe train, one for test), and extract three feature sources on the same samples.

The feature sources are matched across the same samples. *BankCodec* M=2: $B \in \mathbb{R}^{1\times 42\times D}$ (D=1536), flattened to a 64,512-d vector. *MV2V uncompressed* uses the same Memory-V2V wrapper that produces the Table 1 "Memory-V2V (uncompressed)" row: at k=2 the wrapper's mixed tokenizer schedule assigns both memory videos to the *mid* level, yielding 16,380 conditioning tokens per sample ($2 \times 21 \times 15 \times 26$). The 25M-d full flattening is too wide for a meaningful linear probe with 100 training points, so we compare against a per-frame pooled control: for each retrieved memory and latent frame, we average the 390 spatial tokens into one D-dimensional vector, preserving memory identity and temporal layout. Flattening this control gives a 64,512-d feature, the same dimensionality as BankCodec's flattened bank. A random Gaussian matching the bank's feature dimension serves as a chance-level reference. Features are standardised with train statistics; logistic / ridge regularisation is fixed at C=1, α=1.

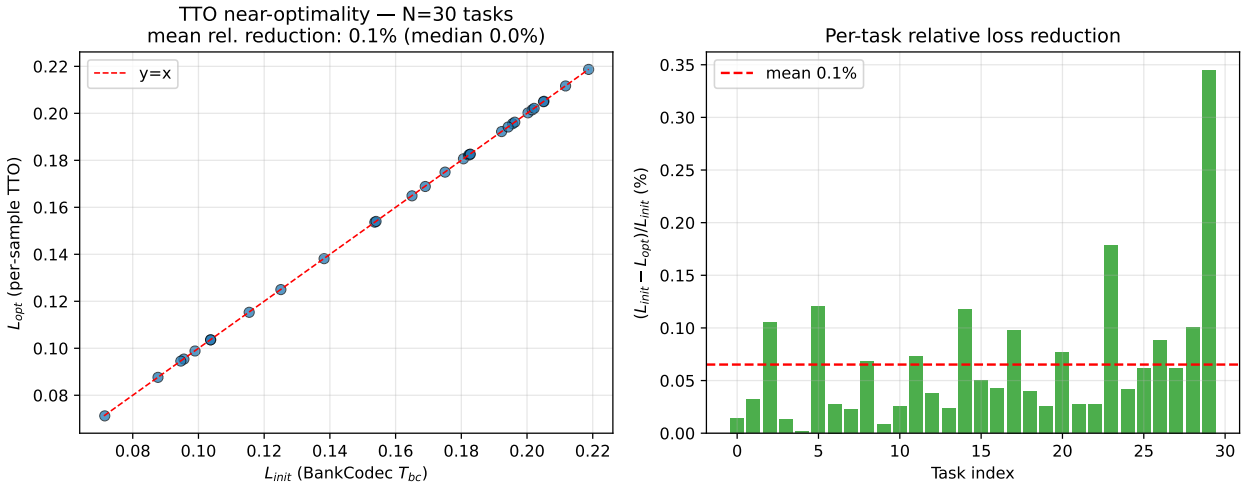

Figure 8. (Left) per-task L_{init} vs. L_{opt} on $N = 30$ tasks (10 clips $\times$ 3 target trajectories). Points on the $y = x$ line indicate that TTO produced no improvement; points below the line indicate a loss reduction. (Right) per-task relative reduction $(L_{\text{init}} - L_{\text{opt}})/L_{\text{init}}$, with the mean drawn as a dashed line.

Figure 9 reports the probe scores. The 64,512-d bank features recover scene identity at 94.0% top-1 (chance 2.0%), versus 90.0% for the per-frame pooled MV2V control. Camera position recovery is R^2=0.90 for BankCodec and 0.97 for the control. Focal-group classification is 98.0% for BankCodec and 98.0% for the control (chance 25.0%).

The probes are representation diagnostics: they measure what scene and camera information is linearly recoverable from the compact bank, not whether the probe itself is a downstream NVS metric.

At an even lower rate (M=1, 21 tokens), categorical identity recovery drops measurably (scene identity 92% vs. 94% at M=2; focal-group 94% vs. 98%) while the camera-position R^2 does not (0.93 at M=1 vs. 0.90 at M=2, both well below the per-frame uncompressed control's 0.97). Categorical identity is the first signal lost when the bank shrinks below the saturation knee; continuous geometric structure is more robust. This pattern is consistent with the M=2 choice from drift saturation (Appendix F) and downstream behavioral metrics (Appendix C).

These probes should be read as diagnostics, not as a complete account of the representations. The held-out set has only 50 scenes, so the absolute scores have wide uncertainty even though the qualitative ordering is stable across the three targets. The MV2V control is also a tractable per-frame pooled summary rather than the full uncompressed token set: it preserves memory identity and temporal layout, but removes spatial detail. All columns use the same held-out scene IDs, samples, and train/test split, so the comparison is paired across feature sources.

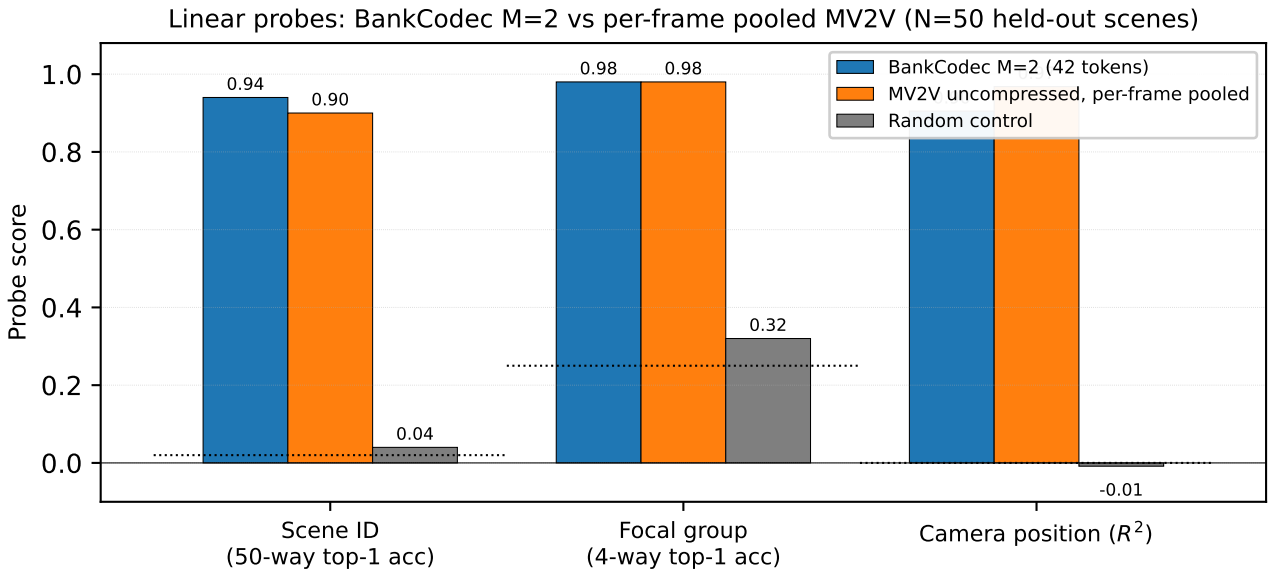

Figure 9. Linear-probe accuracies on the $M{=}2$ BankCodec tokens (**blue**) versus a per-frame pooled summary of Memory-V2V's uncompressed conditioning tokens (**orange**) and a random Gaussian control (**grey**). Dotted lines mark chance level. $N{=}50$ held-out scenes.

I. Benchmark k-Scaling at $k \in \{2, 4, 6, 8\}$

Table 12. Benchmark-level k-scaling at $k \in \{2, 4, 6, 8\}$. BankCodec stays at 42 tokens; the local codec grows from 126 to 504 tokens. The observed metric ranges are stable across retrieval sizes. BankCodec $M{=}2$ is trained with $k \in [2, 6]$.

Method	k	Tokens	MEt3R↑	RotErr$^\circ$ ↓	TransErr↓	SubjCons↑
Local codec $K{=}3$	2	126	**0.376**	1.60	0.063	**0.927**
Local codec $K{=}3$	4	252	**0.377**	1.61	**0.062**	**0.927**
Local codec $K{=}3$	6	378	**0.377**	1.61	**0.063**	**0.927**
Local codec $K{=}3$	8	504	**0.376**	1.63	**0.061**	**0.927**
BankCodec $M{=}2$	2	**42**	0.362	**1.58**	**0.062**	0.926
BankCodec $M{=}2$	4	**42**	0.362	**1.60**	0.063	0.926
BankCodec $M{=}2$	6	**42**	0.361	**1.60**	**0.063**	0.926
BankCodec $M{=}2$	8	**42**	0.361	**1.60**	0.065	0.926