WHICH MEMORY APPLIES HERE? ORGANIZING LONG-TERM MULTIMODAL AGENT MEMORY AS CONDITIONAL HIERARCHIES IN HYPERBOLIC SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory lets a multimodal agent carry what it has observed across sessions, and many observations are general facts accompanied by exceptions that hold under particular conditions. Existing memory systems treat such coexisting records as conflicts and resolve them by recency or by similarity, so the record that applies to the current query image is discarded at write time or outranked at read time. We propose ConeMem, which organizes observations as specificity-ordered conditional hierarchies in the Lorentz model of hyperbolic space, where a general fact sits near the origin and its exceptions branch outward inside entailment cones. The geometry is load-bearing, since the read and write rules need a region fixed by position alone, nested regions, and a constant shrinkage per level, and we prove that negative curvature supplies the three within a bounded radial range whereas zero curvature spends a radial budget that grows multiplicatively with depth. An observer built on frozen vision-language encoders maps the image region selected by the condition text of an exception into the same space and reads applicability off its position relative to the cone, so the choice among coexisting records is a geometric test with no language model call, and a local write procedure uses the same geometry to reinforce a fact, open a branch, or wait for recurrence. Under a leave-one-benchmark-out protocol, ConeMem raises macro-average judge accuracy on four public long-term multimodal memory benchmarks by 13 to 17 points over the strongest published baseline with two reader families at the lowest answer-stage cost, comes within three points of exhaustive verification on annotator-verified conditional-conflict questions, and loses its lead over Euclidean cones as the curvature is driven toward zero.

1 INTRODUCTION

Memory lets an agent built on a large vision-language model stay consistent across long-horizon multimodal interactions and absorb new information as it arrives (Wang & Chen, 2025; Feng et al., 2026; Bei et al., 2026). Because the same entity is observed again and again under changing circumstances, the memory comes to hold general facts together with exceptions that hold under particular conditions, such as how a pet usually behaves beside how it behaves when a stranger is present. The two records do not contradict each other, and which one applies is decided by what the current query image shows, which MemConflict calls a conditional conflict (Tao et al., 2026). Existing write policies settle the coexistence by recency or by similarity, so one record is lost (Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025), and existing read policies rank records by similarity to the question text, so the general fact outranks the exception, whose condition words the question does not mention (Chen et al., 2022; Yeo et al., 2026; Du et al., 2026; Wang et al.). The agent then answers from an inapplicable general fact (Figure 1), and benchmarks with evolving multimodal histories report conflict handling among the weakest capabilities of current systems (Bei et al., 2026; Guo et al., 2026; Zhu et al., 2026; Hu et al., 2026).

A memory that decides which record applies has to represent, for each fact, its specificity and the condition under which it holds. ConeMem organizes memory as specificity-ordered conditional hierarchies in the Lorentz model of hyperbolic space; each entity owns a subtree whose general fact sits near the origin and whose exceptions branch outward inside entailment cones, the cone of

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.

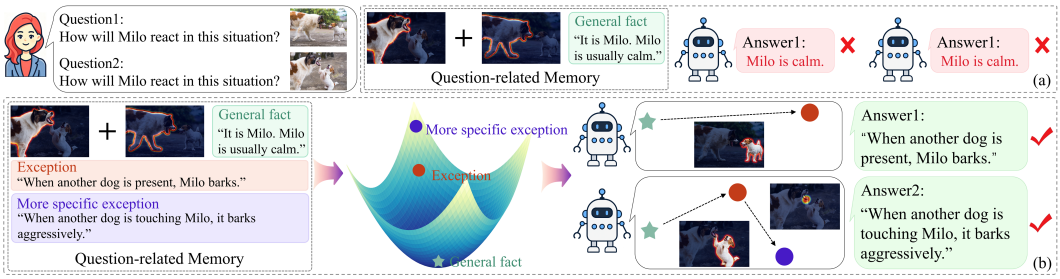

Figure 1: Motivation illustration. (a) A flat memory ranks the general fact above the exception because the question does not mention the condition, and answers from the inapplicable fact. (b) ConeMem stores the general fact near the origin and exceptions inside nested cones; the region selected by the condition text falls inside the cone of the first exception, and the reasoner answers from the resulting path.

an exception being the region in which it applies. The geometry follows from what the read and write rules consume, a region fixed by position alone, nested regions, and a constant shrinkage per level within the bounded radial range that a projection from frozen encoders provides. Negative curvature meets the three requirements at once, because the half-aperture of a Lorentz cone decays exponentially with radius, so a level of nesting costs a fixed radial increment, whereas the Euclidean cone family of the same axioms decays with the reciprocal of the radius, so the same shrinkage costs a multiplicative radius growth (Proposition 1). Realizing the design raises three challenges, and ConeMem answers each. Applicability has to be tested without turning the image into text, because captions discard the cues that decide whether a condition holds, so a cross-attention observer scores every exception of a subtree from the image region its condition text selects, and a frozen reasoner answers from the resulting path. Writes have to keep the hierarchy consistent without re-embedding the store, because global re-embedding drifts under continual updates (Liu et al., 2026a), so a local write procedure reinforces a fact, opens a branch, or holds an observation as pending, and audits each change with a bounded replay. The cones have to be trained from the memory itself, because cones inherited from image-text pretraining sit at the saturation edge where the geometry is effectively flat (Kim et al., 2026), so the training signal is the parent-versus-exception order that writes supply. Our contributions are as follows.

- We characterize how conditional conflict arises from the write and read policies of prior multimodal memory systems and show on annotator-verified questions of four public benchmarks that their errors follow the predicted signature.
- ConeMem is, to our knowledge, the first memory organization that represents the condition under which a fact applies as geometry, with specificity as radius, conditions as entailment cones, and applicability as a continuous margin inside a cone, so that a read scans one subtree and issues no verification call. We prove that the radial budget of nested cones is additive under negative curvature and multiplicative under zero curvature, and show that the lead over Euclidean cones grows with depth and branching and vanishes as the curvature approaches zero.
- Under a leave-one-benchmark-out protocol, ConeMem achieves the highest accuracy on ATM-Bench, Mem-Gallery, MemEye, and H2HMem with two reader families, 13 to 17 macro-average points over the strongest published baseline at the lowest answer-stage cost of the compared systems.

2 Related Work

Long-term memory for multimodal agents. Table 5 in Appendix C tabulates the systems below by whether they keep two records that both hold, test a condition against the query image, and update locally. Memory systems for language agents write observations into a store and retrieve by similarity (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2026), with the conflict policy made explicit as add, update, and delete operations (Chhikara et al., 2025) or validity intervals (Rasmussen et al., 2025). Multimodal extensions organize sources into typed records (Wang & Chen, 2025; Mei et al., 2026; Feng et al., 2026; Jain et al., 2025), retrieve natively over images

and text (Chen et al., 2022; Yeo et al., 2026; Jiang et al., 2025), merge, compress, filter, or rerank records (Bo et al., 2026; Fang et al., 2026; Du et al., 2026; Li et al., 2026), and navigate graph or event layers (Wang et al.; Cao et al., 2026). In each case the policy for two facts that both hold decides between newer and older or between more and less credible, and no representation records the condition under which each fact applies (Tao et al., 2026). Unlike latent memories passed between cooperating agents (Yu et al., 2026; Fu et al., 2026), ConeMem passes readable nodes with a numeric applicability, so the reasoner stays frozen and the training signal comes from writes; the preference for the more specific of two conflicting defaults is classical in nonmonotonic reasoning (Reiter, 1980; McCarthy, 1980; Touretzky, 1986; Horty et al., 1990), and ConeMem keeps it with a continuous geometric score over image regions in place of symbolic matching.

Hyperbolic representations and entailment cones. Hyperbolic space embeds trees with low distortion because its volume grows exponentially with radius (Sarkar, 2011; Nickel & Kiela, 2017; 2018; Tifrea et al., 2018; Chami et al., 2019), and entailment cones turn a partial order into geometry by giving each point a region that contains its descendants (Ganea et al., 2018; Vendrov et al., 2015). MERU and HyCoCLIP bring cones to image-text learning (Desai et al., 2023; Pal et al., 2025), Alignment across Trees aligns hierarchical features across curvatures (Wu et al., 2026), HMCL keeps hyperbolic representations stable under continual updates (Liu et al., 2026a), and hyperbolic retrieval layers serve retrieval-augmented generation and agent memory (Madhu et al., 2026; Cao et al., 2025; Liu et al., 2026b); an audit of released hyperbolic vision-language checkpoints finds their operating points near the Euclidean regime, with cones saturated or misaligned (Kim et al., 2026). These systems define neither write operations nor a condition-dependent applicability, and their cones encode image-text specificity rather than the parent-exception order that memory operations produce, which is the order ConeMem trains from.

3 METHOD

3.1 PROBLEM, MEMORY, AND TWO-AGENT INTERFACE

An agent observes sessions, each a text paired with an image, and later receives questions with an optional query image; a memory system writes each session into a store and serves a question by handing evidence to a frozen answering model, the reader. This paper addresses the case in which the store holds, for one entity, a general fact and exceptions that hold under conditions visible in the query image, and the read has to return the record that applies rather than the one closest to the question.

ConeMem's memory is a set of subtrees $\{T_o\}$, one per entity o; a node u carries a fact text f_u, a condition text c_u that is empty at the root, a node point z_u, the set O_u of observation points that support it, and their session stamps, the root holds the general fact about o, and a child of u refines the condition of u. ConeMem runs two cooperating agents, an observer Ω that sees images, built on frozen vision and text encoders with a trainable cross-attention block and projection, and a reasoner R, a frozen language model that sees text alone and is the reader of the compared systems. The two agents exchange two messages, an entity name at read time answered by a path of nodes given as level, condition text, fact text, and applicability, and a fact text, condition text, target node, and label at write time answered by a placement (Figure 2, Appendix B).

3.2 HYPERBOLIC MEMORY SPACE

We use the Lorentz model of d-dimensional hyperbolic space with curvature $-K$, the set $\mathbb{H}^d_K = \{x \in \mathbb{R}^{d+1} \mid \langle x, x\rangle_{\mathcal{L}} = -1/K,\ x_0 > 0\}$ with Lorentz inner product $\langle x, y\rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i$, geodesic distance $d_{\mathcal{L}}(x, y) = K^{-1/2} \operatorname{arcosh}(-K\langle x, y\rangle_{\mathcal{L}})$, and origin $o = (K^{-1/2}, 0, \ldots, 0)$ (Nickel & Kiela, 2018), with x_s the space-like part of x, $r(x) = d_{\mathcal{L}}(o, x)$ its radius, and $\exp_o$ the exponential map at the origin (Appendix A). The entailment cone of a point z has half-aperture $\omega(z)$, and the exterior angle of a point x with respect to z is $\operatorname{ext}(z, x)$ (Ganea et al., 2018; Desai et al., 2023),

$$\omega(z) = \arcsin\Big(\frac{2\kappa}{\sqrt{K}\,\|z_s\|}\Big), \qquad \operatorname{ext}(z, x) = \arccos\left(\frac{x_0 + K\, z_0\, \langle z, x\rangle_{\mathcal{L}}}{\|z_s\|\sqrt{(K\langle z, x\rangle_{\mathcal{L}})^2 - 1}}\right), \tag{1}$$

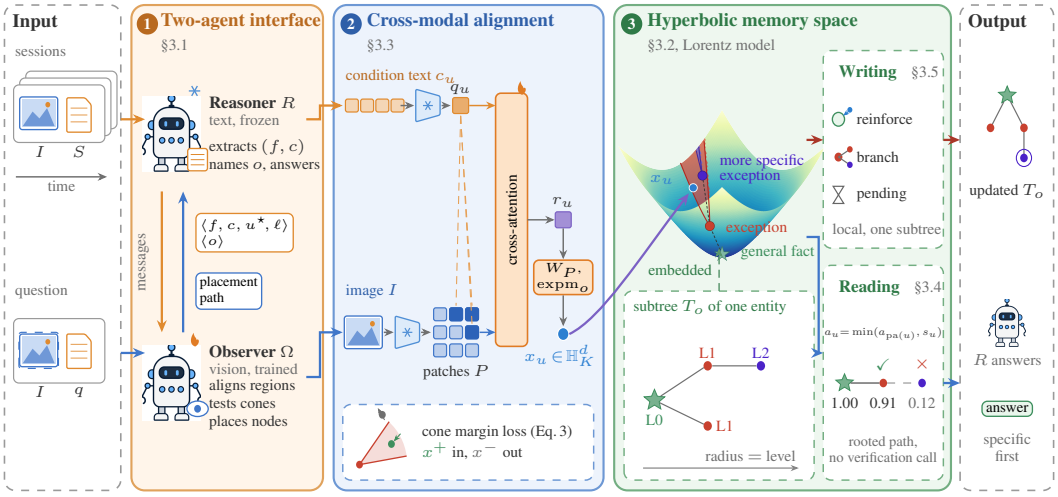

Figure 2: ConeMem framework. Sessions and questions enter on the left and split by modality between two cooperating agents (1). Inside the observer (2), the condition text attends over the image patches through a trained cross-attention block, and the attended region is projected by W_P and $\exp_o$ to a region point x_u in the Lorentz model, trained by the cone margin loss to fall inside the cone of its node. The hyperbolic memory space (3) embeds the subtree of each entity as nested entailment cones.

with the aperture clipped at $\pi/2$, and x lies inside the cone of z when $\mathrm{ext}(z, x) \le \omega(z)$. Points inside the cone of z lie at least as far from the origin as z, so depth is realized by radius, and cones are transitive, so the cone of an exception is contained in its parent's cone (Proposition 3 in Appendix A). The applicability of an observation x to a node z is the margin by which x sits inside the cone, $s(z, x) = \mathrm{clip}\left(1 - \mathrm{ext}(z, x)/\omega(z),\ 0,\ 1\right)$, which is one on the axis and zero outside.

Why the curvature has to be negative. Equation 1 gives every node its cone from its position alone, so a branch created from a single observation receives an applicable region without a parameter of its own, and the write procedure of Section 3.5 places a deeper exception at a larger radius. Nesting therefore consumes radius, and the radius available is bounded by the clipped tangent norm of the projection from the frozen encoders. Ganea et al. derive the cone family that satisfies transitivity and axial symmetry in each geometry, Equation 1 in the Lorentz model and $\omega_{\mathrm{E}}(x) = \arcsin(\kappa_{\mathrm{E}}/\|x\|)$ in Euclidean space (Ganea et al., 2018); since $\|z_s\| = K^{-1/2}\sinh(\sqrt{K}\, r(z))$, the sine of the Lorentz half-aperture decays as $e^{-\sqrt{K}r}$ and the Euclidean one as $1/r$.

Proposition 1 (Radial budget of nested cones, proof in Appendix A). *Fix a shrinkage factor $\varepsilon \in (0, 1)$ and call a level of nesting an ε-level when the sine of the child's half-aperture is at most ε times the parent's. In the Lorentz model of curvature $-K$, an ε-level costs a radial increment of at most $K^{-1/2}\log(1/\varepsilon)$, independent of the radius at which it starts, so a radial band of length L above the saturation radius hosts at least $\sqrt{K}L/\log(1/\varepsilon)$ nested ε-levels. In Euclidean space an ε-level starting at radius r costs the increment $r(1/\varepsilon - 1)$, so the band $[r_0, r_0 + L]$ hosts at most $\log(1 + L/r_0)/\log(1/\varepsilon)$ nested ε-levels.*

The depth capacity of the band is therefore linear in L under negative curvature and logarithmic under zero curvature, and branching follows the same accounting, since sibling cones must narrow as their number b grows (Appendix D). With $K = 1$ and $\kappa = 1$ the saturation edge of Equation 1, where the arcsine argument reaches one and the cone degenerates into a half-space (Kim et al., 2026), lies at radius 1.44; the root sits below it and the exception nodes above it.

3.3 Cross-Modal Alignment in the Observer

The observer aligns an image with a condition text inside the memory space and generates no caption. Given an image I, the frozen vision encoder produces patch features $P \in \mathbb{R}^{n \times d_v}$ once, and given the

condition text c_u of an exception, the frozen text encoder produces a query $q_u \in \mathbb{R}^{d_t}$; one trainable cross-attention block, as in grounded vision-language models (Li et al., 2022; Liu et al., 2024), lets the condition select the patches that concern it, and one trainable projection W_P carries the result into the tangent space at the origin,

$$r_u = \mathrm{softmax}\Big(\frac{q_u W_Q \,(PW_K)^\top}{\sqrt{d}}\Big) PW_V, \qquad x_u = \exp\!\mathrm{m}_o\big([0;\, W_P\, r_u]\big). \tag{2}$$

The region point x_u encodes whether the condition is present rather than describing the image, and since P is shared, the queries of a subtree stack into one matrix and Equation 2 runs once per image.

The training signal comes from writes. When an exception u is written or reinforced, the region points of the sessions that produced it are positives, and the region points computed with the same condition text on sessions that reinforced the parent of u, together with region points from other entities, are negatives; the cone margin loss pulls positives inside the cone and pushes negatives outside,

$$\mathcal{L} = \sum_{(u,x^+)} \big[\,\mathrm{ext}(z_u, x^+) - \omega(z_u) + m\big]_+ + \sum_{(u,x^-)} \big[\omega(z_u) - \mathrm{ext}(z_u, x^-) + m\big]_+, \tag{3}$$

with margin m and z_u detached within a step. Training follows a leave-one-benchmark-out protocol; the write procedure first runs over the histories of the other three benchmarks with a warmed-up observer, those subtrees stay fixed while Equation 3 is minimized, and the trained observer then builds the subtrees of the evaluated histories at test time with no further parameter update.

3.4 Reading

A question arrives as text q with an optional image I. The reasoner names the entity o, and the observer computes a region point x_u for each exception u of T_o with Equation 2 and scores it with $s(z_u, x_u)$. Transitivity keeps one region point consistent along a chain, and the minimum rule extends this across the region points of different condition texts, since an exception applies when its parent applies and not otherwise, $a_{\mathrm{root}} = 1$ and $a_u = \min\big(a_{\mathrm{pa}(u)},\, s(z_u, x_u)\big)$. With a threshold θ, the applicable set $A_\theta = \{u \in T_o \mid a_u > \theta\}$ contains the root and is closed under taking parents for every $\theta \in [0, 1)$ (Proposition 2 in Appendix A). The observer returns the nodes of A_θ, in the common case a single path, followed by the children that fell below θ, marked as not observed with their condition texts and scores. The reasoner answers by the specific-first rule, taking the deepest node of A_θ, and selects a marked child instead when the question states the child's condition in words, so conditions visible in the query image are decided by the observer and conditions stated in text by the reasoner; a question without an image receives the whole subtree marked as not observed.

3.5 Writing

A session arrives as text S with an image I. The reasoner extracts from S the facts it states about entities, each as a fact text f and a condition text c, and for each affected entity marks a target node $u^\star$ in T_o with a label ℓ that is *same* when the new fact restates $u^\star$ and *contrary* when it contradicts $u^\star$. The observer computes the region point x of I under c, or under f when c is empty, and lets u^+ be the deepest node among $u^\star$ and its ancestors whose cone contains x. Each node keeps the Lorentz centroid of its observations as its node point,

$$z_u = \mu(O_u), \qquad \mu(O) = \frac{\sum_{x \in O} x}{\sqrt{K}\,\big\|\sum_{x \in O} x\big\|_{\mathcal{L}}}, \qquad \|v\|_{\mathcal{L}} = \sqrt{|\langle v, v\rangle_{\mathcal{L}}|}, \tag{4}$$

which minimizes the sum of squared Lorentzian distances to the observations (Law et al., 2019). The root sits at a fixed radius ρ_0 below the saturation edge, which keeps its cone a half-space. A new point x is distinguishable from the observations of u when $d_{\mathcal{L}}(x, z_u) > \tau\,\sigma_u$, where σ_u is the mean distance of O_u to z_u, with a default σ_0 when O_u holds fewer than two points, and the geometry then decides among three outcomes (Algorithm 1 in Appendix B). ***Reinforce***, if ℓ is *same*, x joins $O_{u^\star}$ and $z_{u^\star}$ is recomputed with Equation 4. ***Branch***, if ℓ is *contrary* and x is distinguishable from the observations of u^+, a new child of u^+ is created with texts f, c and $O = \{x\}$. ***Pending***, if ℓ is *contrary* and x is not distinguishable, the fact is stored as a pending record on u^+, so a single anomalous observation does not open a noise branch; when a later session repeats the contrary fact

		ATM-Bench		Mem-Gallery		MemEye		H2HMem		Avg. J.Acc		Answer cost	
Method	Venue	J.Acc	EM	J.Acc	EM	J.Acc	M.EM	J.Acc	Recall	Qwen3-VL	Gemma 4	Hours	Tokens
A-MEM	NeurIPS'25	41.86	40.80	52.89	32.61	37.09	40.57	39.94	40.55	42.94	40.27	1.11	2.9k
UniversalRAG	ACL'26	43.10	41.19	63.76	32.50	47.98	49.33	44.36	45.04	49.80	55.09	1.88	4.8k
MemGuide	AAAI'26	48.47	47.80	48.10	32.73	41.40	44.61	39.94	35.81	44.48	45.43	0.78	3.1k
LightMem	ICLR'26	22.70	24.52	38.28	23.20	36.23	39.76	31.04	27.82	32.06	27.89	1.14	2.4k
VimRAG	ICML'26	33.81	32.76	48.33	16.13	28.25	28.23	17.89	17.01	32.07	35.08	9.57	8.2k
Hybrid retrieval	–	44.90	43.60	61.80	31.70	45.20	46.90	41.30	42.00	48.30	53.88	0.71	3.6k
ConeMem	–	**59.83** ↑14.93	**58.06** ↑14.46	**80.92** ↑19.12	**40.18** ↑8.48	**60.71** ↑15.51	**61.02** ↑14.12	**65.34** ↑24.04	**58.87** ↑16.87	**66.70** ↑16.90	**68.57** ↑13.48	**0.62** ↓12.7%	**1.1k** ↓69.4%

Table 1: End-to-end results with the Qwen3-VL-8B-Instruct reader and a GPT-5-mini judge under the leave-one-benchmark-out protocol, with the macro-average under the Gemma-4-12B reader and the answer-stage cost on ATM-Bench (wall-clock hours for 1,044 questions and prompt tokens per question). J.Acc is judge accuracy in percent with the half-width of a 95 percent question-level bootstrap confidence interval, EM normalized exact match, M.EM multiple-choice exact match, and Recall lexical recall; Appendix D gives per-benchmark Gemma results and the full lifecycle cost.

at an indistinguishable region point, the pending record replaces the text of u^+ and its children are retired, which covers the dynamic case in which the fact itself changed. Because the centroid is an extrinsic mean, a reinforced node point can leave its parent's cone, and the observer then moves it along the geodesic toward the parent's axis until it re-enters, on 2.3 percent of reinforce writes (Appendix D.10). A write modifies T_o alone, and after it the agent replays at most ρ stored past questions about o whose path met a modified node; if a replayed answer changes, the write is demoted to a pending record. By Proposition 3, a branch created inside the cone of u^+ lies farther from the origin with a narrower cone, so a deeper exception has a smaller applicable region; a joint condition is stored as a child of the first exception, and the minimum rule makes its applicability the smaller of the two condition scores (costs and scaling in Appendices B and D.12).

4 Experiments

4.1 Setup

We evaluate on ATM-Bench (Mei et al., 2026), Mem-Gallery (Bei et al., 2026), MemEye (Guo et al., 2026), and H2HMem (Zhu et al., 2026), with 1,044, 1,711, 1,855, and 1,982 questions, and report judge accuracy under a fixed GPT-5-mini judge (Zheng et al., 2023) with the half-width of a 95 percent question-level bootstrap interval, each benchmark's native metric, evidence recall, the share of gold records handed to the reader, and gold-first rate, the share of questions whose first evidence item is gold. Since most questions involve no exception, two annotators verified the 1,621 questions whose released annotations indicate coexisting records about one entity, labeling the general fact, the exception, whether the deciding condition is visible in the query image, stated in the text, or ambiguous, and whether a later record supersedes an earlier one; the verified stratum holds 1,014 questions (683 visible, 214 text-stated, 117 ambiguous) and the temporal stratum 289. We compare with A-MEM (Xu et al., 2026), UniversalRAG (Yeo et al., 2026), MemGuide (Du et al., 2026), LightMem (Fang et al., 2026), and VimRAG (Wang et al.) under one protocol, with a hybrid control that fuses BM25 (Robertson & Zaragoza, 2009) with dense multimodal retrieval and its equal-budget variant with a 6.1-million-parameter reranker trained on the observer's histories, and with controls that each change one component of ConeMem, three write policies (recency overwrite, temporal invalidation, similarity merge), hyperbolic retrieval without conditional structure, three replacements of the observer (caption plus reasoner, a single reader over a flat store, and pairwise verification by the reader), and three Euclidean geometries (cosine neighborhoods, entailment cones with the aperture fixed by the norm, and learnable cones) that keep the tree, the observer, and the training (Appendix C). The observer uses frozen SigLIP 2 So400m encoders (Tschannen et al., 2025) with d=256, K=1, κ=1, θ=0.5, τ=1.5, and ρ=20 on the four benchmarks, and the reasoner is the baselines' reader, Qwen3-VL-8B-Instruct (Bai et al., 2025) or Gemma 4 12B, invoked with text alone.

4.2 Main Results

Table 1 reports end-to-end results. With Qwen3-VL-8B-Instruct as the reader, ConeMem leads on each benchmark in judge accuracy and in the native metric, and its macro-average of 66.70 exceeds UniversalRAG, the strongest published method, by 16.90 points and the hybrid control by 18.40; with Gemma-4-12B it reaches 68.57 and exceeds UniversalRAG by 13.48, so the gain transfers

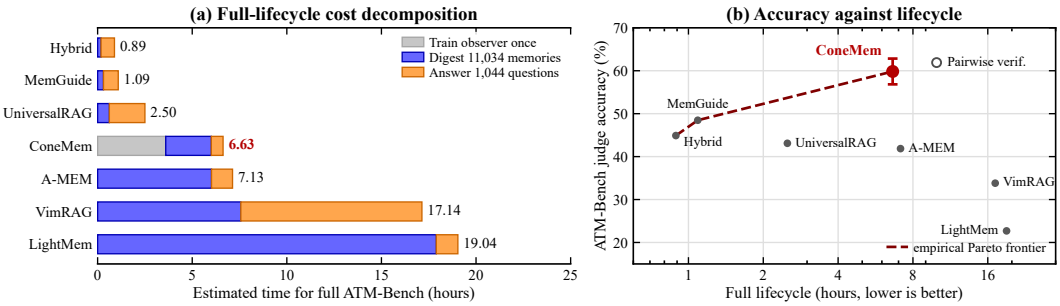

Figure 3: Lifecycle cost and accuracy on ATM-Bench with the Qwen3-VL reader. (a) Estimated time to train the observer once, digest the 11,034 memories once, and answer the 1,044 questions once. (b) Judge accuracy against the same lifecycle time, with the bootstrap interval of ConeMem, pairwise verification as the exhaustive language model route over the same tree, and the empirical Pareto frontier dashed; Appendix D repeats the analysis on the other benchmarks.

Method	Write policy	Full Avg.	Verified conditional stratum, J.Acc				Temporal	Evidence, complete benchmarks	
			Visible n=683	Text n=214	Ambig. n=117	Whole n=1,014	J.Acc n=289	Recall ↑	Gold-first ↑
Hybrid retrieval	rank by similarity	48.30	29.0	40.2	25.6	31.0	44.7	62.0	45.1
UniversalRAG	rank by similarity	49.80	30.8	42.5	27.4	32.9	46.3	63.7	47.8
Recency overwrite	replace older	51.9	26.1	38.9	23.5	28.5	55.8	58.3	41.2
Temporal invalidation	expire older	53.4	31.5	43.0	27.9	33.5	60.2	61.0	45.9
Similarity merge	merge into nearest	52.6	28.0	40.4	25.1	30.3	52.9	59.4	43.0
Hyperbolic retrieval	append	55.8	36.9	44.8	31.2	37.9	49.6	66.2	50.7
Caption plus reasoner	ConeMem tree	62.3	49.4	54.1	39.8	49.3	60.5	78.5	64.1
Single reader	append	58.9	46.2	51.3	37.5	46.3	57.1	70.8	57.3
Pairwise verification	ConeMem tree	67.4	68.1	55.9	44.0	62.7	62.8	88.0	82.6
ConeMem	ConeMem tree	**66.70**	**65.6**	**54.8**	**41.2**	**60.5**	**62.1**	**86.1**	**80.2**

Table 2: Judge accuracy on the annotator-verified conditional stratum by visibility of the deciding condition, on the temporal stratum, and evidence recall and gold-first rate on the complete benchmarks, with the Qwen3-VL reader; Full Avg. is the macro-average of Table 1. Pairwise verification is the exhaustive route over the same tree, listed as a reference; bold marks the best system that issues no verification call.

across readers and, under the held-out protocol, to unseen entities, images, and conditions. ConeMem also has the lowest answer-stage cost of the compared systems, 0.62 hours and 1.1k prompt tokens per question. Figure 3 places the systems in the accuracy-cost plane over a complete lifecycle. ConeMem's digest stage costs 2.41 hours of reasoner extraction and labeling, so its lifecycle of 6.63 hours sits between the retrieval systems and the structured memories; the systems that run faster trail it by at least 11 points, and A-MEM, VimRAG, and LightMem take longer and trail it by 18 to 37 points, so ConeMem lies on the empirical Pareto frontier. Pairwise verification over the same tree costs 6.4 times the answer-stage time and 5.7 times the tokens for a gain of two points.

4.3 Evidence Behavior on the Verified Conditional-Conflict Questions

Table 2 shows the failure and its resolution. The two similarity-ranked systems and the three write policies of prior systems land within five points of each other on the verified stratum; on the temporal stratum the order reverses, temporal invalidation reaches 60.2, and ConeMem matches it at 62.1 because promotion of a pending record replaces a superseded fact. Hyperbolic retrieval without conditional structure gains four to nine points, which bounds what geometry alone contributes. ConeMem reaches 60.5 on the verified stratum and 65.6 where the condition is visible, covers 86.1 percent of the gold records with its path, and hands a gold record over first for 80.2 percent of the questions, 32.4 points more than UniversalRAG; on the text-stated questions it leads the write-policy variants by 12 points, and outside the strata it still leads UniversalRAG by 11.6 points. The three observer comparisons decide the interface. Caption plus reasoner trails ConeMem by 16.2 points where the condition is visible, so captions lose the conditions that geometry keeps; single reader trails by 19.4 points, so the path suffices and the reasoner gains nothing from seeing the image; and

Variant	Full Avg.	Stratum Avg.	Gold-first
ConeMem ($K=1$)	**66.70**	**60.5**	**80.2**
Flat embedding top-K, no hierarchy, no observer	49.20	31.4	38.4
Hybrid retrieval with an equal-size trained reranker	52.6	35.8	44.9
Same observer, flat store, no tree	57.1	44.6	55.2
Euclidean space, same tree, cosine neighborhoods	62.1	51.0	66.0
Euclidean space, same tree, entailment cones (aperture from the norm)	63.0	52.4	68.1
Euclidean space, same tree, learnable cones	63.4	53.3	69.3
Lorentz space at near-flat curvature ($K=0.05$)	63.8	53.9	69.8
Depth capped at one exception	64.3	54.4	72.4
Without the minimum rule	65.4	57.9	76.1
Without pending records	63.9	55.6	74.0
Pretrained MERU cones, no write training	60.4	43.5	51.2

Table 3: Ablations with the Qwen3-VL reader; each row removes or replaces one element and keeps the rest fixed, and Stratum Avg. is judge accuracy on the verified stratum (further rows in Appendix D).

pairwise verification exceeds ConeMem by 2.5 points on the visible questions and 2.2 on the stratum at 6.4 times the answer-stage time, so the geometric test recovers most of what exhaustive verification recovers.

4.4 The Geometry Is Load-Bearing

Table 3 separates the sources of the gain. The flat embedding top-K variant sets the lower bound at 49.20 on the full average and 31.4 on the stratum, and a trained reranker of the observer's size on the hybrid control recovers 3.4 points, so the training budget by itself explains little; the trained observer over a flat store reaches 57.1 and 44.6 (Appendix D), so the remaining 16 stratum points come from the hierarchy with its minimum rule and from the geometry. Capping depth at one exception loses 6.1 points, concentrated on nested exceptions, removing the minimum rule lets a child apply while its parent does not, removing pending records fills the tree with noise branches, and pretrained MERU cones without write training drop the gold-first rate to 51.2, so the training signal that matters is the parent-versus-exception order that writes supply.

The three Euclidean rows and the near-flat row isolate the curvature. Under identical training, cosine neighborhoods lose 9.5 stratum points, the Euclidean entailment cones 8.1, and learnable cones with an axis and an aperture per node 7.2, so the cone parameterization recovers little, and driving the curvature of ConeMem toward zero reproduces the loss. Figure 4 traces the mechanism. Panels (a) and (b) evaluate the applicability test on subtrees whose depth and branching are set by construction (Appendix D.11); the Lorentz cones lose 7.8 points from depth one to depth four and the Euclidean cones lose 19.1 to 24.6, and the gap opens in the same way with branching, as Proposition 1 predicts. Panel (c) sweeps the curvature with the constants fixed; accuracy rises from the near-flat regime, where ConeMem is within the confidence interval of the Euclidean learnable cones, to a plateau around $K = 1$, where the exception nodes operate at $\sqrt{K}\,r$ of 1.8, 2.6, and 3.4 by level, above the saturation edge where the audit locates released hyperbolic checkpoints (Kim et al., 2026). Panel (d) confirms the prediction end to end on the verified stratum, where the lead over the Euclidean learnable cones grows from 5.1 points at depth one to 13.2 at depth three or more and from 4.6 to 13.7 points from two to six siblings.

4.5 Analysis

The diagnosis makes four predictions beyond aggregate numbers, and Figure 4e to 4h confirm each. First, the error of a similarity-ranked or overwrite-based system on questions with two or more gold records should rise with the gap between the question's similarity to the most and to the least similar gold record, which measures how far the deciding record sits below the ranking cutoff; panel (e) shows the error of the five prior policies climbing by 21 to 25 points from the low to the high tertile, while ConeMem stays within three points because applicability does not depend on question wording. Second, the gain should grow with exception depth, and the gain over the Euclidean cones follows this pattern in Figure 4d. Third, the two observers should fail on different questions; of one hundred sampled errors, 63 caption-plus-reasoner errors involve a visible condition absent from the caption, whereas the ConeMem errors split between text-stated or ambiguous conditions, region points outside

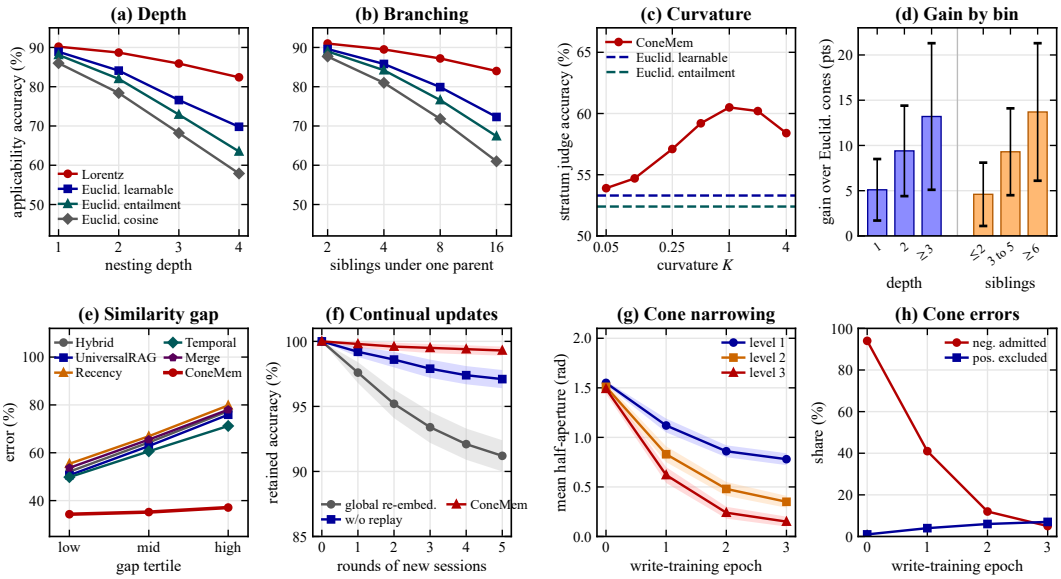

Figure 4: The geometry is load-bearing, and the diagnosis predicts behavior. (a, b) Applicability-test accuracy on constructed subtrees for the Lorentz and three Euclidean geometries. (c) Stratum judge accuracy as the curvature is swept; dashed lines are Euclidean controls. (d) Gain over Euclidean learnable cones by annotated depth (blue) and siblings (orange), 95% bootstrap. (e) Error by similarity gap (n=806 per tertile). (f) Retained accuracy after rounds of new sessions, mean over four benchmarks. (g) Mean half-aperture by level across write-training epochs. (h) Share of negatives admitted and positives excluded.

the cone, wrong parents at write time, and entity naming, which an oracle decomposition prices at 4.4, 4.4, and 1.7 points (Appendix D). Fourth, a local write should survive continual updates that a global re-embedding does not; panel (f) shows a global re-embedding baseline losing 8.8 points over five rounds of new sessions, ConeMem without replay 2.9, and ConeMem 0.7. Panels (g) and (h) show where the cones come from; write training narrows the exception cones within three epochs from a saturated mean half-aperture to level-dependent apertures of 0.78, 0.35, and 0.15 radians while the share of admitted negatives falls from 94 to 5 percent, so the geometry the read relies on is a product of the memory itself rather than of image-text pretraining.

5 Conclusion

We studied long-term multimodal agent memory as a problem of organizing conditional evidence, on which prior systems fail because they write by overwriting and read by similarity. ConeMem orders facts as radius in Lorentz space, turns conditions into entailment cones, and scores applicability geometrically, so that a write touches one subtree and the read-time choice among coexisting records is a geometric test, while extraction, naming, and labeling stay with a reasoner that answers from a readable path. The curvature is part of the argument, since nested cones spend an additive radial budget under negative curvature and a multiplicative one under zero curvature, and the lead over Euclidean cones grows with depth and branching and vanishes as the curvature approaches zero. With the evaluated histories held out of training, the result is the highest accuracy on four public benchmarks with two reader families at the lowest answer-stage cost, a lead of 16 points over captioning where the condition is visible, and accuracy within three points of exhaustive verification at a fraction of its cost. The evidence also bounds the claims; a condition that is neither visible nor stated falls back to the general fact, and entity naming and target-node labeling remain reasoner decisions priced at about four points. Representing the condition as geometry offers agents that keep defaults and exceptions a route whose cost grows with the hierarchy rather than with the number of pairwise comparisons.

AI USE STATEMENT

In this work, We have not used generative AI tools.

ETHICS STATEMENT

The benchmarks are public releases used under their licenses, and the verification annotation adds labels to their questions without collecting new data. A deployed system that retains facts about people requires user consent and a deletion mechanism; the subtree design makes deleting the records about one entity a single local operation.

REPRODUCIBILITY STATEMENT

Section 3 defines each operation; Appendices A to D give geometric preliminaries, proofs, architecture, constants, prompts, pseudocode, the baseline settings, the annotation protocol, and the stratum definitions. Anonymous source code, trained weights, the split files of the leave-one-benchmark-out protocol, and the annotation labels are provided as supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. pp. 40750–40784, 2026.

Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, et al. Vilomem: Agentic learner with grow-and-refine multimodal semantic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5476–5486, 2026.

Linxiao Cao, Ruitao Wang, Jindong Li, Zhipeng Zhou, and Menglin Yang. Hyperbolicrag: Enhancing retrieval-augmented generation with hyperbolic representations. *arXiv preprint arXiv:2511.18808*, 2025.

Shuqi Cao, Jingyi He, and Fei Tan. Higmem: A hierarchical and llm-guided memory system for long-term conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33853–33862, 2026.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. volume 32, 2019.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z Pan, Ruifeng Xu, and Kam-Fai Wong. Memguide: Intent-driven memory selection for goal-oriented multi-session llm agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30584–30592, 2026.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

Muxin Fu, Xiangyuan Xue, Yafu Li, Zefeng He, Siyuan Huang, Xiaoye Qu, Yu Cheng, and Yang Yang. Latentmem: Customizing latent memory for multi-agent systems. *arXiv preprint arXiv:2602.03036*, 2026.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pp. 1646–1655. PMLR, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026.

John F Horty, Richmond H Thomason, and David S Touretzky. A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial intelligence*, 42(2-3):311–348, 1990.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. 2026:156259–156291, 2026.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. Augustus: An llm-driven multimodal agent system with contextualized user memory. *arXiv preprint arXiv:2510.15261*, 2025.

Hongda Jiang, Xinyuan Zhang, Siddhant Garg, Rishab Arora, Shiun-Zu Kuo, Jiayang Xu, Aaron Colak, and Xin Luna Dong. Memory-qa: Answering recall questions based on multimodal memories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24255–24277, 2025.

Jaeyoung Kim, Eunseok Kim, and Dongsuk Jang. Is the geometry doing the work? an operating-point audit of hierarchy in hyperbolic vision-language models. *arXiv preprint arXiv:2607.05268*, 2026.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *International conference on machine learning*, pp. 3672–3681. PMLR, 2019.

Chunyu Li, Mengyuan Zhang, Jingyi Kang, Ding Chen, Jiajun Shen, Bo Tang, Xuanhe Zhou, Feiyu Xiong, and Zhiyu Li. Memreranker: Reasoning-aware reranking for agent memory retrieval. *arXiv preprint arXiv:2605.06132*, 2026.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Jiahong Liu, Ming Shen, Xiaohao Liu, Rex Ying, Menglin Yang, Tat-Seng Chua, and Irwin King. Hyperbolic multimodal continual learning. 2026a.

Jiahong Liu, Menglin Yang, and Irwin King. Hyperbolic learning for structured data, knowledge, and memory: A tutorial. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 13351–13355, 2026b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Hiren Madhu, Ngoc Bui, Ali Maatouk, Leandros Tassiulas, Smita Krishnaswamy, Menglin Yang, Sukanta Ganguly, Kiran Srinivasan, and Rex Ying. Hyprag: Hyperbolic dense retrieval for retrieval augmented generation. 2026.

John McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial intelligence*, 13 (1-2):27–39, 1980.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to me: Long-term personalized referential memory qa. *arXiv preprint arXiv:2603.01990*, 2026.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. volume 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pp. 3779–3788. PMLR, 2018.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Raymond Reiter. A logic for default reasoning. *Artif. Intell.*, 13(1-2):81–132, 1980. doi: 10.1016/0004-3702(80)90014-4. URL https://doi.org/10.1016/0004-3702(80)90014-4.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International symposium on graph drawing*, pp. 355–366. Springer, 2011.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. Memconflict: Evaluating long-term memory systems under memory conflicts. *arXiv preprint arXiv:2605.20926*, 2026.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. 2018.

David S Touretzky. *The mathematics of inheritance systems*, volume 8. Morgan Kaufmann, 1986.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. 2015.

Qiuchen Wang, Shihang Wang, Yu Zeng, Qiang Zhang, Fanrui Zhang, Zhuoning Guo, Bosi Zhang, Wenxuan Huang, Lin Chen, Zehui Chen, et al. Navigating massive visual context in retrieval-augmented generation via multimodal memory graph. In *Forty-third International Conference on Machine Learning*.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Wei Wu, Xiaomeng Fan, Yuwei Wu, Zhi Gao, Pengxiang Li, Yunde Jia, and Mehrtash Harandi. Modality alignment across trees on heterogeneous hyperbolic manifolds. In *International Conference on Learning Representations*, volume 2026, pp. 11519–11554, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. volume 38, pp. 17577–17604, 2026.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, 2026.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, et al. Dual latent memory for visual multi-agent system. 2026.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. volume 36, pp. 46595–46623, 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Shiping Zhu, Yibo Yang, Zhengyang Wang, Tiancheng Shen, Dandan Guo, and Ming-Hsuan Yang. H2hmem: A multimodal memory benchmark for agents in human-human interactions. *arXiv preprint arXiv:2606.09461*, 2026.

A Geometric Preliminaries and Proofs

Exponential map. A tangent vector v at the origin o is carried onto the manifold by

$$\exp_o(v) = \cosh(\sqrt{K}\|v\|)\, o + \sinh(\sqrt{K}\|v\|)\, \frac{v}{\sqrt{K}\|v\|}, \tag{5}$$

and the result satisfies $\langle x, x\rangle_{\mathcal{L}} = -1/K$ and $r(x) = \|v\|$. Tangent vectors are clipped to norm 6 before the map, so the radial range available to region points is $[0, 6]$, and arcosh and $\arccos$ receive arguments clamped to their domains, which keeps each operation stable at the depths that occur in the experiments.

Proposition 2 (The applicable set is a rooted subtree)**.** *For every $\theta \in [0, 1)$ the set $A_\theta = \{u \in T_o \mid a_u > \theta\}$ contains the root and is closed under taking parents.*

Proof. Since $a_{\text{root}} = 1 > \theta$, the root lies in A_θ. For a non-root $u \in A_\theta$, the minimum rule $a_u = \min(a_{\text{pa}(u)}, s(z_u, x_u))$ gives $a_{\text{pa}(u)} \geq a_u > \theta$, so $\text{pa}(u) \in A_\theta$. Inducting along the path to the root shows that the whole path belongs to A_θ, so A_θ is connected and contains the root. A_θ is therefore a rooted subtree. □

Proposition 3 (Radial order and transitivity)**.** *If x lies inside the cone of z, then $\|x_s\| \geq \|z_s\|$, and the cone of x is contained in the cone of z.*

Proof of Proposition 3. In the Lorentz model with curvature $-K$, $\cosh(\sqrt{K}\, d(o,y)) = \sqrt{K}\, y_0$. By the hyperbolic law of cosines for the triangle o, z, x with angle α at z,

$$\cosh(\sqrt{K}\, d(o,x)) = \cosh(\sqrt{K}\, d(o,z)) \cosh(\sqrt{K}\, d(z,x)) - \sinh(\sqrt{K}\, d(o,z)) \sinh(\sqrt{K}\, d(z,x)) \cos\alpha.$$

Inside the cone of z the angle α is at least $\pi - \omega(z) \geq \pi/2$, so the subtracted term is nonnegative, hence $\cosh(\sqrt{K} d(o,x)) \geq \cosh(\sqrt{K} d(o,z))$ and $d(o,x) \geq d(o,z)$. In the Lorentz model, $\sqrt{K}\, d(o,y) = \operatorname{arsinh}(\sqrt{K}\|y_s\|)$, and the function is increasing, so $\|x_s\| \geq \|z_s\|$. Transitivity of the cones follows from the construction of Ganea et al. (2018), whose apertures shrink with radius so that the cone of a point inside the cone of z is contained in the cone of z. □

Proof of Proposition 1. In the Lorentz model, $\|z_s\| = K^{-1/2} \sinh(\sqrt{K}\, r)$ at radius r, so below saturation Equation 1 gives $\sin\omega(r) = 2\kappa / \sinh(\sqrt{K}\, r)$. For $a > 0$ and $b \geq 0$, $\sinh(a+b) = \sinh a \cosh b + \cosh a \sinh b \geq \sinh a\,(\cosh b + \sinh b) = e^b \sinh a$, because $\cosh a \geq \sinh a$. With $a = \sqrt{K} r$ and $b = \sqrt{K}\Delta$,

$$\frac{\sin\omega(r+\Delta)}{\sin\omega(r)} = \frac{\sinh(\sqrt{K} r)}{\sinh(\sqrt{K}(r+\Delta))} \leq e^{-\sqrt{K}\Delta},$$

so the ratio falls to ε or below once $\Delta \geq K^{-1/2} \log(1/\varepsilon)$, whatever the starting radius. A band of length L above the saturation radius therefore hosts at least $\lfloor \sqrt{K} L / \log(1/\varepsilon) \rfloor$ consecutive ε-levels. In Euclidean space the cone family with the same axioms has $\sin\omega_{\mathrm{E}}(r) = \kappa_{\mathrm{E}}/r$ for $r \geq \kappa_{\mathrm{E}}$ (Ganea et al., 2018), so $\sin\omega_{\mathrm{E}}(r+\Delta)/\sin\omega_{\mathrm{E}}(r) = r/(r+\Delta)$, which equals ε exactly when $\Delta = r(1/\varepsilon - 1)$. Starting at r_0, the n-th ε-level therefore ends at radius $r_0 \varepsilon^{-n}$, and $r_0 \varepsilon^{-n} \leq r_0 + L$ holds if and only if $n \leq \log(1 + L/r_0)/\log(1/\varepsilon)$. □

Operating point. Following the audit of Kim et al. (2026), we report the dimensionless operating point $\sqrt{K}\, r$ of the nodes. Equation 1 saturates when $\sqrt{K}\|z_s\| \leq 2\kappa$, that is when $\sinh(\sqrt{K} r) \leq 2\kappa$; with $K = 1$ and $\kappa = 1$ the saturation radius is $\operatorname{arsinh}(2) = 1.44$. The root sits at $\rho_0 = 0.5$ below the edge, so its cone is a half-space and admits the region points of its descendants, and the exception nodes of the four evaluated benchmarks sit at operating points whose 5th and 95th percentiles are 1.5 and 3.7, with level means 1.8, 2.6, and 3.4, above the edge. Table 19 lists the level statistics.

Lorentz centroid. For a set O of points on $\mathbb{H}^d_K$, the squared Lorentzian distance $\|x - z\|^2_{\mathcal{L}} = -2/K - 2\langle x, z\rangle_{\mathcal{L}}$ is linear in z for fixed x, so the minimizer over the manifold of $\sum_{x \in O} \|x - z\|^2_{\mathcal{L}}$ is the rescaled sum of Equation 4 (Law et al., 2019). The Fréchet mean of the geodesic distance minimizes $\sum_{x \in O} d_{\mathcal{L}}(x,z)^2$ instead and has no closed form on the manifold; for three points on one geodesic at parameters 0, 0, and 2 the Fréchet mean lies at parameter 0.667 and the Lorentz centroid at 0.740, which is why Section 3.5 treats the centroid as an extrinsic mean and repairs containment after each update.

Containment repair. Let p be the parent of u and let g be the geodesic ray from the origin through z_p, the axis of the cone of p. When $\operatorname{ext}(z_p, z_u) > \omega(z_p)$ after a reinforce, the repair replaces z_u by the point of the geodesic from z_u to the closest point of g at the same radius as z_u that first satisfies $\operatorname{ext}(z_p, z_u) \leq \omega(z_p)$, found by bisection on the geodesic parameter. The radius of z_u is not decreased below that of z_p, so the radial order holds, and the children of u are then checked against the new cone of u in the same way. Each repair touches the subtree below the reinforced node and no other subtree.

B Implementation Details

Cost accounting. A read costs one image encoding, one batched geometric pass of Equation 2 over the exceptions of the named subtree, and two short language model calls, one to name the entity and one to answer from a path that spans a few lines of text; the applicability scores involve no language model. A write costs one extraction call per session, one labeling call per affected entity, one observer pass, and a replay bounded by ρ, and neither procedure calls a language model on pairs of memories.

The size of the store enters a read through entity resolution, which matches the name against the stored subtree names by text embedding, and through the fallback for a question that names no stored entity, which returns the root facts of the five closest subtrees; Appendix D.12 shows that read latency grows by 0.05 seconds per question and accuracy drops by 1.1 points when the store grows from eleven thousand to one hundred thousand records.

Notation. Table 4 lists the symbols used in Section 3.

T_o	subtree of entity o	x_u	region point of an image under c_u
f_u, c_u	fact text and condition text of node u	z_u	node point, the Lorentz centroid of O_u
O_u	observation points supporting u	$\omega(z)$, $\mathrm{ext}(z, x)$	half-aperture of the cone of z, exterior angle of x
$u^\star$, ℓ	target node and label chosen by the reasoner	$s(z, x)$, a_u	cone margin of x in the cone of z, applicability of u
u^+	deepest ancestor of $u^\star$ whose cone contains x	θ, τ, ρ_0, ρ	read threshold, distinguishability factor, root radius, replay bound
$r(x)$	geodesic radius of x	$\sqrt{K}\, r$	operating point of a node

Table 4: Notation. The reasoner produces f_u, c_u, $u^\star$, and ℓ from text; the observer produces x_u, z_u, $s(z, x)$, and a_u from images and geometry.

Dataset details. ATM-Bench contains 6,742 emails, 3,759 images, and 533 videos in one personal archive, 11,034 memories in total, and 1,044 questions. Mem-Gallery contains 20 persona-driven contexts, 240 sessions, 7,944 memory rows, 1,711 questions over nine task types, and 487 questions with a query image. MemEye spans 16 contexts, 3,392 memory rows, and 1,855 evaluation rows in open-answer and multiple-choice form. H2HMem contains 25 dialogue contexts, 7,078 memory turns, 1,300 images, and 1,982 questions. Image captions are retained from each benchmark's release when available; for H2HMem, which provides no captions, a frozen sidecar is generated with Qwen3-VL-8B-Instruct at temperature 0 and shared across the compared methods. Videos in ATM-Bench enter the observer as eight uniformly sampled frames per video, and the baselines receive the sampled frames and the released derived text of each video.

Model assignment. In the Qwen3-VL setting, the reader of the baselines and the reasoner of ConeMem are Qwen/Qwen3-VL-8B-Instruct; in the Gemma 4 setting, google/gemma-4-12B-it. The reasoner receives text alone. The generation stages use a 32,768-token context limit and temperature 0. The judge is gpt-5-mini with a fixed rubric that receives the question, the reference answer, and the prediction, and no other content.

Observer. The frozen vision encoder returns $n = 729$ patch features of dimension $d_v = 1152$ for a 384×384 input, and the frozen text encoder returns a pooled embedding of dimension $d_t = 1152$. The cross-attention block has eight heads of width 64, and the projection W_P is a two-layer perceptron with hidden width 512 and output width $d = 256$, followed by Equation 5. The trainable parameters together number 6.1 million and occupy 24 MB.

Training under leave-one-benchmark-out. For each evaluated benchmark, warm-up uses the caption entailment objective of MERU on the image and caption pairs of the three training benchmarks for one epoch with batch size 256, learning rate 10^{-4}, and AdamW. Write training then proceeds in two stages. The write procedure runs over the histories of the three training benchmarks in session order with the warmed-up observer and the reasoner's labels, producing subtrees that are then frozen. Positives and negatives are collected from these frozen subtrees as described in Section 3.3, with two negatives per positive sampled from the parent's sessions and two from other entities, and Equation 3 is minimized for three epochs with batch size 128, learning rate 5×10^{-5}, and margin $m = 0.05$. The trained observer is then frozen, and the write procedure runs once over the histories of the evaluated benchmark to produce the subtrees used at test time; the reasoner's labels on these histories decide placements and update no parameter. Pending records contribute no training signal until promoted. Four observers are trained in total, one per evaluated benchmark, at 3.6 GPU hours each on one A800.

Cold start. Before the first exception exists, a subtree is a single root, reading returns the general fact, and the system behaves as a per-entity store whose only structure is the grouping of records by entity. The warm-up gives the projection a usable geometry from the first write, and the first branch of a subtree is decided by the distinguishability test against the root's observations with the default spread σ_0.

Reasoner prompts. The reasoner receives three prompt families. The extraction prompt takes a session transcript and returns records with entity name, fact text, and condition text. The labeling prompt takes a new record and the subtree texts and returns a target node identifier and a label (*same* or *contrary*). The answer prompt takes the question and the path as numbered lines of the form (level, condition text, fact text, applicability), with the marked children listed after the applicable nodes, and instructs the specific-first rule together with the rule that a marked child is chosen when the question states its condition. Temperature is zero in each call.

Write procedure. Algorithm 1 lists the write procedure for one record.

Algorithm 1 ConeMem write procedure for one record

Input. Fact f, condition c, target $u^\star$, label ℓ, image I, subtree T_o
Compute region point $x \leftarrow \exp_o([0;\ W_P\, r])$ via Equation 2
$u^+ \leftarrow$ deepest among $u^\star$ and its ancestors whose cone contains x
if ℓ is *same* **then**
  $O_{u^\star} \leftarrow O_{u^\star} \cup \{x\}$; update $z_{u^\star}$ via Equation 4
  Repair containment of $u^\star$ in its parent's cone and of the descendants of $u^\star$ (Appendix A)
else if ℓ is *contrary* and $d_{\mathcal{L}}(x, z_{u^+}) > \tau\, \sigma_{u^+}$ **then**
  Create child of u^+ with texts (f, c) and $O = \{x\}$
else if ℓ is *contrary* and x is not distinguishable **then**
  if a matching pending record exists on u^+ **then**
    Promote it; replace the text of u^+; keep the old text as history; retire the children of u^+
  else
    Create pending record (f, c, x) on u^+
  end if
end if
Replay at most ρ past questions about o whose path met a modified node
If any replayed answer changes, demote the write to pending

Aliases and multiple entities. Entity names are normalized by the reasoner at both write and read time against the existing subtree names, and a name that matches no stored name is resolved by text embedding over the subtree names. A question that names several entities receives one path per entity. A question that names no stored entity receives the root facts of the five closest subtrees by text embedding of their names; on the four benchmarks 4.1 percent of the questions take this route, and their accuracy is reported in Appendix D.12.

Fixed constants. Memory dimension $d = 256$, curvature $K = 1$, aperture constant $\kappa = 1$, root radius $\rho_0 = 0.5$, tangent norm clip 6, margin $m = 0.05$, threshold $\theta = 0.5$, distinguishability factor $\tau = 1.5$, default spread $\sigma_0 = 0.4$, replay bound $\rho = 20$. The same values are used on the four benchmarks.

C Baselines, Controls, Strata, and Annotation

Candidate questions for annotation. The 1,621 candidate questions are the ATM-Bench questions whose canonical evidence lists two or more memories, the Mem-Gallery knowledge resolution and conflict detection tasks, the MemEye rows at the evolutionary synthesis level of its reasoning-depth axis, and the H2HMem questions from multi-party conversations, which are the places where the released annotations indicate coexisting records about one entity.

Prior families. Table 5 answers the three questions of Section 2 for the families of prior work.

Published baselines. Each published baseline runs its official public implementation at the release used for the comparison, with the same reader, the same 32,768-token context limit, the same caption sidecar, the same judge, and top-ten retrieval where the method exposes a retrieval size. A-MEM builds and links notes with its default note schema. UniversalRAG uses its default modality router and granularity settings. MemGuide uses its default intent extraction and gap-based filtering. LightMem

Family	Both records kept	Condition tested on the image	Local update
Overwrite or expiry (Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025)	×	×	✓
Similarity merge (Bo et al., 2026)	×	×	✓
Similarity retrieval (Chen et al., 2022; Yeo et al., 2026)	✓	×	✓
Intent and constraint filters (Du et al., 2026; Li et al., 2026)	✓	×	✓
Graph and event memory (Xu et al., 2026; Wang et al.; Cao et al., 2026)	✓	language model on text	partial
Hyperbolic retrieval (Madhu et al., 2026; Cao et al., 2025)	✓	×	×
Latent memory (Yu et al., 2026; Fu et al., 2026)	compressed	×	×
Inheritance networks (Touretzky, 1986; Horty et al., 1990)	✓	symbolic literals	✓
ConeMem	✓	geometric, no model call	✓

Table 5: How prior families answer the three questions of the conditional case. A family keeps both records when a general fact and its exception survive a write, tests the condition when the read decides from the query image which record applies, and updates locally when a write touches the records of one entity alone.

runs its three compression stages with the default settings of its release; its digest stage dominates its lifecycle in Table 7. VimRAG builds its multimodal memory graph with the default node and edge settings and navigates with the default budget, which produces the 8.2k prompt tokens per question reported in Table 7.

Hybrid retrieval control and its equal-budget variant. The control indexes the canonical text of each memory with BM25 and embeds each memory and each question with the frozen gme-Qwen2-VL-2B-Instruct encoder (Zhang et al., 2024), fuses the two rankings by reciprocal rank fusion with the constant 60, and hands the top ten records to the reader with the same serialization as the other retrievers. The equal-budget variant adds a cross-attention reranker with 6.1 million trainable parameters over the frozen encoder outputs, trained with a contrastive objective on the session and fact pairs of the training benchmarks under the same leave-one-benchmark-out protocol, and reranks the top fifty fused candidates before the top ten are handed over.

Write-policy variants and hyperbolic retrieval. The three write-policy variants use ConeMem's extraction prompt and reasoner and keep a flat list of records per entity. On a *contrary* label, recency overwrite replaces the older record, temporal invalidation keeps it with a validity interval that ends at the new session and hides it from later reads, and similarity merge asks the reasoner to merge the new record into its nearest neighbor by text embedding when the cosine similarity exceeds 0.85 and appends it otherwise. Reading returns the ten most similar records of the named entity. Hyperbolic retrieval embeds records and questions in the same Lorentz space with the warmed-up projection, ranks by geodesic distance, and returns the ten closest records without a tree.

Observer variants. Caption plus reasoner keeps ConeMem's tree and replaces the observer at read time with a caption of the query image generated by the reader; the reasoner receives the caption and the subtree texts and selects the deepest node whose condition text it judges to hold in the caption. Single reader gives the reader the query image and the flat per-entity store directly, with the ten most similar records serialized as for the retrievers. Pairwise verification keeps ConeMem's tree, gives the reader the query image, and asks one question per exception of the named subtree, whether the condition text holds in the image; the deepest exception judged to hold, with its ancestors judged to hold, forms the path, and the answer call is the same as ConeMem's. Same observer with a flat store uses the trained observer to score the condition text of each record of the named entity against the query image and hands the ten highest-scoring records to the reader without hierarchy or minimum rule.

Rule library. The rule library stores each fact with an integer priority, zero for a general fact and one more for each nesting level, and stores the condition text beside it. At read time the reader captions the query image once and is asked, for each condition of the entity, whether the condition holds in the caption; the answer comes from the highest-priority fact whose condition holds, and from the general fact otherwise.

Geometric controls. The three Euclidean controls keep the tree, the cross-attention block, the training data, the margin loss, and the leave-one-benchmark-out protocol, and place region points and nodes in $\mathbb{R}^{256}$ with the same tangent norm clip. The cosine variant scores applicability by the cosine similarity between the region point and the node point rescaled to $[0, 1]$ through a learned threshold. The entailment-cone variant uses the Euclidean cone family of Ganea et al. (2018), whose half-aperture $\arcsin(\kappa_{\mathrm{E}}/\|x\|)$ is fixed by the norm with $\kappa_{\mathrm{E}} = 1$, and scores applicability by the margin s of Section 3.2 with the Euclidean exterior angle; it is the Euclidean member of the same axiomatic family as the Lorentz cones, so the two differ in curvature alone. The learnable-cone variant gives each node a learnable axis and a learnable half-aperture, initialized from the node point and the Euclidean cone family, and scores applicability in the same way. The near-flat row of Table 3 is ConeMem itself with $K = 0.05$ and the other constants unchanged.

Released strata and verification annotation. The released strata use released annotations alone. ATM-Bench lists canonical evidence memories per question, and the stratum consists of the 386 questions with two or more; Mem-Gallery labels each question with one of nine task types, and the stratum consists of the 162 questions of the knowledge resolution and conflict detection types; MemEye labels each row with a level on its reasoning-depth axis, and the stratum consists of the 302 rows at the evolutionary synthesis level; H2HMem labels each conversation as dyadic or multi-party, and the stratum consists of the 771 questions from multi-party conversations. Two annotators with the released gold evidence, the history, and the query image in view then filled one form per question with the entity, the general fact, the exception, whether the answer depends on the more specific record, the depth of the exception below the general fact, the number of sibling exceptions of the same entity present in the history, the visibility of the deciding condition (visible in the query image, stated in the question text, or ambiguous), and whether the later record supersedes the earlier one. Disagreements were resolved by discussion, and the reported κ values are computed on the labels before resolution. The similarity gap of Figure 4e is computed on the 2,418 questions with two or more gold records across the four benchmarks as the difference between the question's cosine similarity to its most similar and to its least similar gold record under the gme-Qwen2-VL-2B-Instruct encoder, and tertiles are formed per benchmark. The deciding depth and the sibling bins of Figure 4d and Table 9 come from the annotation.

D Additional Results

This appendix collects the results that the main text cites, in the order in which the main text reaches them, per-benchmark results under the second reader, the lifecycle analysis on the three benchmarks other than ATM-Bench, the further ablations, the depth and branching bins with the oracle decomposition, the training protocols, the second judge, the sensitivity sweeps, the second reasoner, label noise, the write invariants, the controlled depth and branching sweeps, scale, the Mem-Gallery task types, the cone diagnostics behind Section 4.4, and traced cases.

D.1 Per-Benchmark Results with the Gemma 4 Reader

Table 6 reports the per-benchmark results whose macro-average appears in Table 1. The ranking of the systems under the Gemma 4 reader matches the ranking under the Qwen3-VL reader on each benchmark, and the gain of ConeMem over the strongest published method ranges from 4.98 points on MemEye to 18.84 points on H2HMem.

D.2 Lifecycle Cost on the Four Benchmarks

Table 7 decomposes the ATM-Bench lifecycle whose answer-stage part appears in Table 1 and whose accuracy-cost plane appears in Figure 3, and Figure 5 repeats the analysis on Mem-Gallery, MemEye, and H2HMem under the same workload definition, one digest of the complete history and one answer

		ATM-Bench		Mem-Gallery		MemEye		H2HMem		Avg.
Method	Venue	J.Acc	EM	J.Acc	EM	J.Acc	M.EM	J.Acc	Recall	J.Acc
A-MEM	NeurIPS'25	42.43	41.95	50.03	28.40	24.42	30.93	44.20	39.97	40.27
UniversalRAG	ACL'26	49.04	48.18	64.82	32.14	58.17	38.07	48.34	41.31	55.09
MemGuide	AAAI'26	54.98	53.07	48.33	22.85	30.40	36.66	47.98	41.97	45.43
LightMem	ICLR'26	24.33	25.48	33.72	17.01	20.70	30.59	32.80	27.79	27.89
VimRAG	ICML'26	22.51	21.36	46.99	14.14	31.81	34.64	39.00	33.67	35.08
Hybrid retrieval	–	50.10	48.90	62.70	30.60	54.80	37.10	47.90	40.80	53.88
ConeMem	–	**61.52** ↑6.54	**59.78** ↑6.71	**82.41** ↑17.59	**41.30** ↑9.16	**63.15** ↑4.98	**62.84** ↑24.77	**67.18** ↑18.84	**60.23** ↑18.26	**68.57** ↑13.48

Table 6: End-to-end results with the Gemma-4-12B reader and the GPT-5-mini judge, with the same metrics as Table 1.

for each evaluation question, with the one-time observer training of 3.6 GPU hours charged in full to each benchmark. ConeMem's lifecycle on the three benchmarks is 6.4, 5.4, and 6.4 hours, its answer stage is the cheapest of the compared systems on each of them, and it lies on the empirical Pareto frontier of each benchmark together with the hybrid control and UniversalRAG. The systems that run faster than ConeMem trail it by 13 to 28 points, and LightMem and VimRAG take longer and trail it by 24 to 47 points. The digest stage of ConeMem is dominated by the extraction and labeling calls of the reasoner, at 1.9, 0.8, and 1.7 hours, and the observer pass adds under three minutes per benchmark.

	Hybrid	MemGuide	UniversalRAG	ConeMem	Pairwise verif.	A-MEM	VimRAG	LightMem
Train (GPU h)	0	0	0	3.6	3.6	0	0	0
Digest (h)	0.18	0.31	0.62	2.41	2.41	6.02	7.57	17.90
Answer (h)	0.71	0.78	1.88	**0.62**	3.95	1.11	9.57	1.14
Lifecycle (h)	0.89	1.09	2.50	6.63	9.96	7.13	17.14	19.04
Calls per question	1	2	2	2	7.4	1	4.2	1
Tokens per question	3.6k	3.1k	4.8k	**1.1k**	6.3k	2.9k	8.2k	2.4k
J.Acc (%)	44.90	48.47	43.10	**59.83**	61.9	41.86	33.81	22.70

Table 7: Lifecycle cost on ATM-Bench, digesting 11,034 memories once and answering 1,044 questions once with the Qwen3-VL reader. Train is one-time observer training in GPU hours, Digest and Answer are wall-clock hours, Calls and Tokens are language model calls and prompt tokens per question at the answer stage, and J.Acc is ATM-Bench judge accuracy.

D.3 Further Ablations

Table 8 adds the rows that Table 3 omits. The rule library stores integer priorities and asks the reader whether each condition text holds in a caption of the query image, so it inherits the losses of captioning and lands 14.4 stratum points below ConeMem. Replacing the cross-attention block by a pooled image embedding costs 13.2 stratum points, because a pooled embedding describes the image as a whole and the condition text no longer selects the region that decides applicability. Removing the marked children costs 2.9 stratum points, most of them on the text-stated questions, whose conditions the reasoner reads from the marked children. Removing containment repair or replay costs little on the stratum and more on retained accuracy under continual updates, which Figure 4f reports, and removing the warm-up slows training without changing the final accuracy by more than 1.6 points.

Variant	Full Avg.	Stratum Avg.	Gold-first
ConeMem	66.7	60.5	80.2
Rule library with priorities and a language model condition detector	58.7	46.1	55.9
Observer without cross-attention, pooled image embedding	60.9	47.3	58.0
Without the marked children on the path	66.1	57.6	79.6
Without containment repair	66.3	59.6	79.1
Without replay	65.8	59.7	79.0
Without warm-up	65.1	59.2	77.9

Table 8: Further ablations with the Qwen3-VL reader, in the format of Table 3.

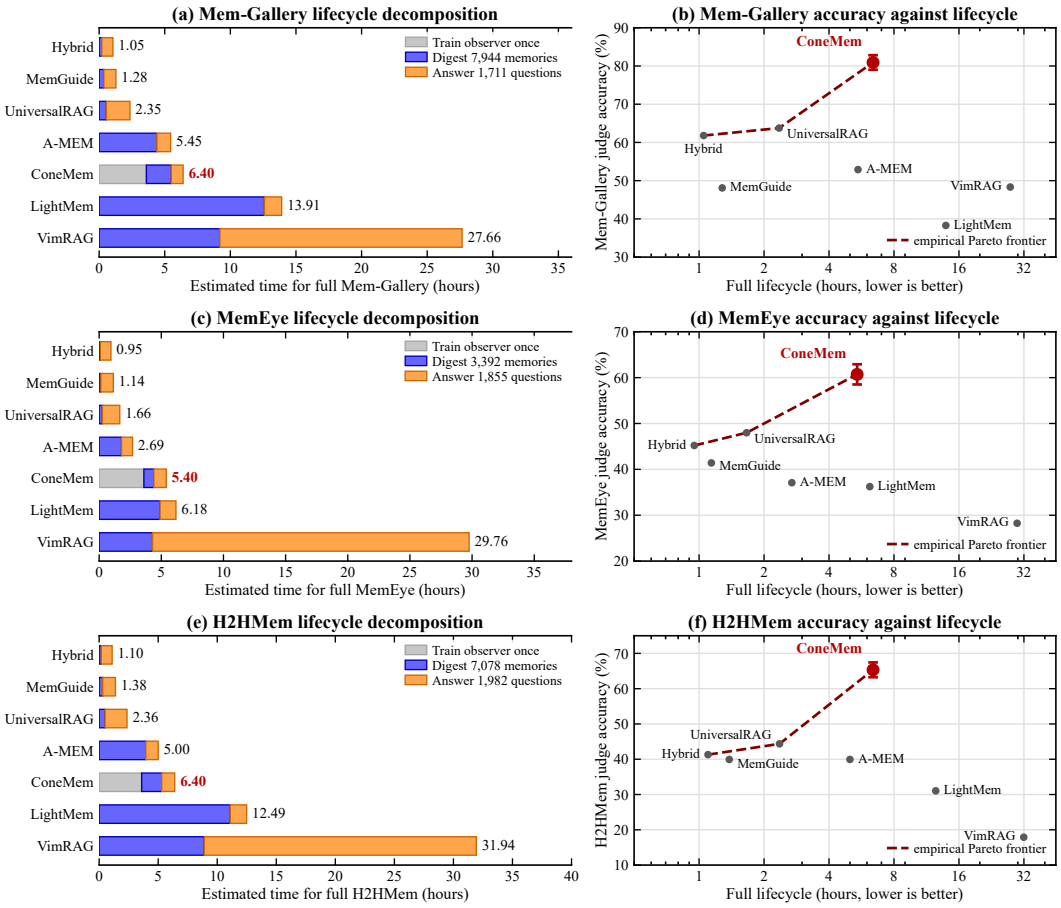

Figure 5: Lifecycle cost and accuracy on the three benchmarks not shown in Figure 3, with the Qwen3-VL reader. Left panels decompose one-time observer training, one digest of the complete history, and one answer for each evaluation question; right panels pair the same estimated wall clock with judge accuracy, with the bootstrap confidence interval of ConeMem as an error bar and the empirical Pareto frontier of the compared systems as a dashed line.

D.4 Depth and Branching Bins and the Oracle Decomposition

Table 9 reports the two analyses that Section 4.5 summarizes. The gain of ConeMem over the flat store grows from 25.2 to 38.7 points across the depths the annotators assigned to the deciding exception, and the gain over each Euclidean cone family grows in the same way with depth and with the number of sibling exceptions under one parent; the bins weight to the stratum gaps of Table 3, 7.2 points over the learnable cones and 8.1 over the entailment cones. The oracle decomposition replaces one reasoner or observer decision at a time by the annotation. Oracle target and label add 4.4 points, an oracle visual condition adds 4.4 more, and gold extraction adds 1.7, so the write-time decisions of the reasoner and the misses of the observer account for comparable shares of the remaining gap, and the extraction step accounts for the smallest.

D.5 Training Protocols

Table 10 compares three ways of separating training from evaluation. Leave-one-benchmark-out, used throughout the paper, trains the observer on the histories of the other three benchmarks. The in-benchmark context split trains on half of the contexts of each benchmark, or on the first half of the sessions of the single ATM-Bench archive, and evaluates the questions of the other half. Unlabeled adaptation starts from the leave-one-benchmark-out observer and minimizes Equation 3 on the subtrees written on the evaluated histories, without reading questions, answers, or gold evidence.

Deciding depth	1	2	≥ 3
n	612	289	113
Gain vs. flat top-K	+25.2±3.4	+33.6±5.1	+38.7±8.2
Gain vs. Euclidean learnable cones	+5.1±3.4	+9.4±5.0	+13.2±8.1
Gain vs. Euclidean entailment cones	+5.8±3.5	+10.5±5.1	+15.0±8.3
Sibling exceptions	≤ 2	3 to 5	≥ 6
n	571	318	125
Gain vs. Euclidean learnable cones	+4.6±3.5	+9.3±4.8	+13.7±7.6
Gain vs. Euclidean entailment cones	+5.3±3.6	+10.6±4.9	+15.3±7.8

(a) Gain in judge accuracy on the verified stratum by annotator-labeled depth of the deciding exception and by the number of sibling exceptions under its parent, with 95 percent bootstrap half-widths.

Setting	Stratum
ConeMem	60.5
+ oracle target and label	64.9
+ oracle visual condition	69.3
+ gold extraction	71.0

(b) Oracle decomposition on the verified stratum, replacing one reasoner or observer decision at a time by the annotation.

Table 9: Predictions of the diagnosis by depth and branching, and the remaining headroom.

The three protocols lie within 1.5 points of each other, so the gain does not depend on seeing the evaluated histories. The seed variation of ConeMem over three training seeds is 0.34, 0.41, 0.52, and 0.47 points on the four benchmarks under the leave-one-benchmark-out protocol.

Protocol	ATM-Bench	Mem-Gallery	MemEye	H2HMem	Avg.
Leave-one-benchmark-out (main)	59.8	80.9	60.7	65.3	66.7
In-benchmark context split	60.4	81.6	61.3	66.1	67.4
Unlabeled adaptation on evaluated histories	61.0	82.2	62.0	66.9	68.0

Table 10: Judge accuracy (%) under three training protocols with the Qwen3-VL reader.

D.6 Second Judge and Manual Audit

Table 11 repeats the macro-average of Table 1 with Qwen3-235B-A22B-Instruct as the judge under the same rubric. The ranking of the systems is unchanged and the lead of ConeMem over UniversalRAG is 16.7 points under the second judge. A stratified audit of 200 judgments against blind manual decisions agrees with GPT-5-mini on 193 (96.5 percent) and with the second judge on 190 (95.0 percent).

Judge	A-MEM	UniversalRAG	MemGuide	LightMem	VimRAG	Hybrid	ConeMem
GPT-5-mini	42.9	49.8	44.5	32.1	32.1	48.3	66.7
Qwen3-235B-A22B-Instruct	42.1	49.2	43.8	31.4	31.6	47.7	65.9

Table 11: Macro-average judge accuracy (%) under two judges with the Qwen3-VL reader.

D.7 Sensitivity

Table 12 varies the constants that shape the geometry and the read rule. The read threshold, the distinguishability factor, and the root radius move the full average by at most 1.7 points across their ranges. The curvature row reproduces the sweep of Figure 4c on the full average, and the aperture constant κ behaves as the saturation edge predicts, since a larger κ raises the edge to $\operatorname{arsinh}(2\kappa)$ and pushes more exception nodes into the saturated regime, whereas a smaller κ narrows every cone and excludes more positives. The tangent norm clip sets the length of the radial band that Proposition 1 counts; shortening it from 6 to 3 removes 5.4 stratum points, most of them on questions whose deciding exception lies at depth two or more, while widening it to 8 changes accuracy by 0.1 point, because the trained subtrees do not reach that radius.

D.8 A Second Reasoner

Table 13 replaces the reasoner by Llama-3.1-8B-Instruct (Grattafiori et al., 2024), a text-only model, with the same observer and the same memory. The lead over the write-policy variants under the same

Threshold θ	0.3	0.4	0.5	0.6	0.7		
Full Avg. / Stratum Avg.	65.9 / 58.3	66.4 / 59.9	66.7 / 60.5	66.5 / 60.0	65.6 / 57.8		
Factor τ	1.0	1.25	1.5	2.0	2.5		
Full Avg. / Stratum Avg.	65.2 / 57.9	66.3 / 59.8	66.7 / 60.5	66.2 / 59.6	65.0 / 57.3		
Root radius ρ_0	0.25	0.375	0.5	0.75	1.0		
Full Avg. / Stratum Avg.	66.1 / 59.5	66.5 / 60.1	66.7 / 60.5	66.3 / 59.8	65.4 / 58.2		
Curvature K	0.05	0.1	0.25	0.5	1	2	4
Full Avg. / Stratum Avg.	63.8 / 53.9	64.2 / 54.7	65.1 / 57.1	66.0 / 59.2	66.7 / 60.5	66.5 / 60.2	65.6 / 58.4
Aperture constant κ	0.5	0.75	1	1.5	2		
Full Avg. / Stratum Avg.	65.8 / 58.6	66.4 / 59.9	66.7 / 60.5	66.2 / 59.7	65.1 / 57.9		
Tangent norm clip	3	4	6	8			
Full Avg. / Stratum Avg.	64.4 / 55.1	65.7 / 58.2	66.7 / 60.5	66.7 / 60.6			

Table 12: Sensitivity of judge accuracy to the read threshold, the distinguishability factor, the root radius, the curvature, the aperture constant, and the tangent norm clip, each varied with the other constants at their defaults.

reasoner is preserved, which shows that the gain is carried by the memory and the observer rather than by the reasoner.

Reasoner	Method	ATM-Bench	Mem-Gallery	MemEye	H2HMem	Avg.
Llama-3.1-8B	Recency overwrite	47.9	66.1	45.8	48.7	52.1
Llama-3.1-8B	Temporal invalidation	49.2	68.0	47.3	50.1	53.7
Llama-3.1-8B	ConeMem	57.6	79.1	58.9	63.2	64.7
Qwen3-VL-8B	ConeMem	59.8	80.9	60.7	65.3	66.7

Table 13: Judge accuracy (%) with a second reasoner.

D.9 Robustness to Label Noise

The write procedure depends on the reasoner's *same* or *contrary* labels and target nodes. Table 14 flips a fixed share of the labels at random during both write training and test-time writes. Accuracy degrades gracefully, losing 2.1 points on the full average at 20 percent noise and 4.4 at 30 percent, and the stratum, where the labels matter most, loses 4.3 and 8.5 points. Across the four benchmarks the labels of Qwen3-VL-8B-Instruct and Llama-3.1-8B-Instruct agree on 94.7 percent of records and their target nodes on 91.4 percent, so the noise the reasoner introduces in practice lies below the levels at which the geometry stops compensating.

Share of flipped labels	0%	10%	20%	30%
Full Avg.	66.7	65.9	64.6	62.3
Stratum Avg.	60.5	58.9	56.2	52.0

Table 14: Judge accuracy under random flips of the reasoner's labels.

D.10 Write Invariants and the Choice of Mean

Table 15 measures, over the writes on the four evaluated histories, how often the recomputed node point of a reinforced node leaves the cone of its parent before repair, how far the repair moves it, and what three update rules cost and yield. The Lorentz centroid leaves the parent cone on 2.3 percent of reinforce writes, the repair moves the point by 0.04 geodesic units on average, the repaired subtrees violate neither containment nor the radial order on the recorded writes, and the repair changes 0.16 percent of the replayed answers. The Fréchet mean, computed by ten Riemannian gradient steps, reaches the same accuracy within 0.2 points at nine times the update time, and the tangent-space mean at the origin loses 0.4 points.

Update rule	Containment violations before repair	Mean repair distance	Replayed answers changed	Update time (ms)	Full Avg.
Lorentz centroid (Eq. 4)	2.3%	0.04	0.16%	0.6	66.7
Fréchet mean, ten Riemannian steps	1.9%	0.03	0.12%	5.4	66.9
Tangent-space mean at the origin	3.1%	0.05	0.21%	0.7	66.3

Table 15: Write invariants under three update rules. Violations are counted before the containment repair; after repair no violation remains on the recorded writes.

D.11 Controlled Depth and Branching

Figure 4a and 4b plot Table 16. To remove the dependence of the annotator bins on which questions the benchmarks happen to contain, we build nested and widened variants of the verified subtrees of each evaluated benchmark by pruning each verified exception chain to a fixed depth and by grafting verified exceptions of other entities of the same benchmark under one parent, so that depth and branching are set by construction, and we measure the accuracy of the applicability test against the annotators' visibility labels for the four geometries under identical training. The Lorentz cones lose 7.8 points from depth one to depth four, the Euclidean learnable cones 19.1, the Euclidean entailment cones 24.6, and cosine neighborhoods 28.1, and from two to sixteen siblings the four geometries lose 7.0, 17.3, 21.6, and 26.7 points. At depth one and with two siblings the four geometries lie within four points of each other, so the gap is a property of nesting and branching rather than of the observer.

	Depth				Siblings under one parent			
Geometry	1	2	3	4	2	4	8	16
Lorentz cones (ConeMem)	90.2	88.7	85.9	82.4	91.0	89.5	87.2	84.0
Euclidean learnable cones	88.9	84.1	76.6	69.8	89.6	85.8	79.9	72.3
Euclidean entailment cones	88.1	82.0	72.9	63.5	89.0	84.2	76.6	67.4
Euclidean cosine neighborhoods	86.0	78.4	68.2	57.9	87.7	81.0	71.8	61.0

Table 16: Accuracy (%) of the applicability test on subtrees whose depth and branching are set by construction, for the Lorentz cones and the three Euclidean controls.

D.12 Scale

Table 17 pads the ATM-Bench store with the memory histories of the other three benchmarks and, beyond that, with disjoint slices of those histories under distinct entity namespaces, so that the store grows from 11 thousand to 100 thousand records while the 1,044 ATM-Bench questions stay the same. Read latency grows by 0.05 seconds per question over this range, because a read scans the named subtree and the two language model calls dominate, and write latency grows by 0.13 seconds per session, because replay is bounded by ρ. Accuracy drops by 1.1 points at 100 thousand records, and the drop is concentrated in the entity resolution step, whose failure rate rises from 2.9 to 4.8 percent as more subtrees compete for a name; the questions that name no stored entity and take the five-subtree fallback reach 41.7 percent at 11 thousand records and 39.9 at 100 thousand.

Records in the store	11k	25k	50k	100k
Subtrees	1.3k	3.1k	6.4k	13.2k
Read latency (s per question)	0.38	0.39	0.41	0.43
Observer pass (ms per question)	31	31	33	34
Write latency (s per session)	1.92	1.94	1.98	2.05
Entity resolution failures (%)	2.9	3.4	4.1	4.8
Store size (MB)	38	86	171	340
ATM-Bench J.Acc	59.8	59.6	59.2	58.7

Table 17: Latency, storage, and accuracy on the ATM-Bench questions as the store grows.

D.13 PER-TASK RESULTS ON MEM-GALLERY

Table 18 breaks Mem-Gallery down by its nine native task types. The largest gains over UniversalRAG appear on conflict detection (+32.1), knowledge resolution (+32.1), and visual-centric reasoning (+29.3), the three types in which coexisting or updated records about one entity decide the answer, while the extraction types gain by 11 to 15 points and answer refusal gains by 10.3.

Method	FR	VS	TTL	TR	VR	MR	KR	CD	AR
UniversalRAG	74.0	78.1	71.2	55.3	48.9	60.7	44.4	34.6	58.2
ConeMem	88.6	89.3	84.1	74.8	78.2	79.6	76.5	66.7	68.5

Table 18: Judge accuracy (%) on the nine Mem-Gallery task types, factual retrieval (FR), visual-centric search (VS), test-time learning (TTL), temporal reasoning (TR), visual-centric reasoning (VR), multi-entity reasoning (MR), knowledge resolution (KR), conflict detection (CD), and answer refusal (AR).

D.14 CONE DIAGNOSTICS

Table 19 lists, per level, the operating point $\sqrt{K}\, r$, the mean half-aperture, and the cone error rates of the exception nodes of the four evaluated benchmarks after write training, and Figure 6 shows the distributions behind Section 4.4. Before write training the half-apertures of the exception nodes concentrate at the clip of $\pi/2$, the saturated regime that the audit of released hyperbolic vision-language checkpoints reports (Kim et al., 2026), with a mean of 1.52 radians and 94 percent of negative regions admitted. After write training the distribution separates into three groups that follow the levels, with means of 0.78, 0.35, and 0.15 radians, an overall mean of 0.63 radians, 5 percent of negatives admitted, and 7 percent of positives excluded. Panel (b) compares where the node points of the three cone families come to rest. The Lorentz nodes of successive levels sit at operating points 1.8, 2.6, and 3.4, an increment of about 0.8 per level, as the additive budget of Proposition 1 predicts; the nodes of the Euclidean entailment cones sit at radii 2.1, 3.9, and 5.4, and 38 percent of their level-three nodes rest against the tangent norm clip of 6, because a level of nesting in that family costs a multiplicative radius growth, and the Euclidean learnable cones sit between the two. Panel (c) reports the effective dimension of the region points, measured as the participation ratio of the eigenvalues of their covariance in the tangent space at the origin, which falls from 61 after warm-up to 18 after write training, far below the nominal dimension of 256; the branching accounting of Section 3.2 therefore applies at a rate set by an effective dimension near 18 rather than by 256. Panel (d) plots the bounds of Proposition 1 for a shrinkage factor of one half, the number of nested levels that a band of length L hosts, at least $L/\log 2$ under curvature -1 and at most $\log_2(1+L)$ under zero curvature with the innermost cone at radius one; the band above the saturation edge that the clip leaves, of length 4.56, hosts at least six levels in the Lorentz model and at most two and a half in Euclidean space.

Level	Nodes	Mean $\sqrt{K}\, r$	5th percentile	95th percentile	Mean half-aperture (rad)	Negatives admitted (%)	Positives excluded (%)
Root (general fact)	4,870	0.50	0.50	0.50	$\pi/2$		
1	5,860	1.8	1.5	2.2	0.78	6	5
2	2,140	2.6	2.1	3.1	0.35	4	9
≥ 3	610	3.4	2.8	3.7	0.15	3	13

Table 19: Operating points and cone statistics of the nodes of the four evaluated benchmarks after write training, with $K = 1$ and $\kappa = 1$. The saturation edge lies at $\operatorname{arsinh}(2) = 1.44$; the root sits below it by design and the exception nodes above it.

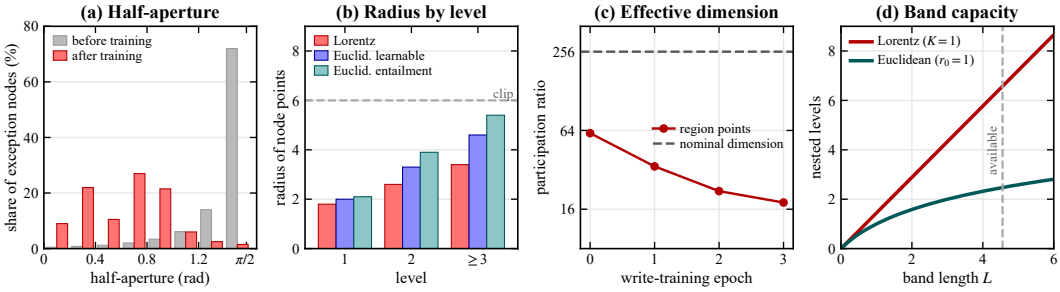

Figure 6: Cone diagnostics. (a) Distribution of the half-apertures of the exception nodes before and after write training. (b) Mean radius of the node points by level for the Lorentz cones and the two Euclidean cone controls, with the tangent norm clip marked. (c) Effective dimension of the region points across write-training epochs against the nominal dimension. (d) Bounds of Proposition 1 on the number of nested levels a radial band of length L hosts, for a shrinkage factor of one half; the dashed line marks the band above the saturation edge that the clip leaves.