LINGUA: Bridging the Grounding Gap in VideoQA via Typed Memory and Belief-State Reasoning

Saman Forouzandeh, Wei Peng, Xinghuo Yu and Mahdi Jalili
{saman.forouzandeh, wei.peng3, xinghuo.yu , mahdi.jalili}@rmit.edu.au
School of Engineering, RMIT University, Melbourne, Australia

The Grounding Gap

- VideoQA models are accurate but often miss the right video segment.
- On NExT-QA: ~69% answers correct, but only 16% justified by evidence.
- This answer–evidence disconnect is the grounding gap.
- Prior work treats grounding as pre/post-processing or alignment.
- LINGUA enforces grounding at inference time against localized evidence.

Research Questions

Q1: Can grounding be coupled with reasoning at inference time, instead of relying on pre-processing or training-time alignment?

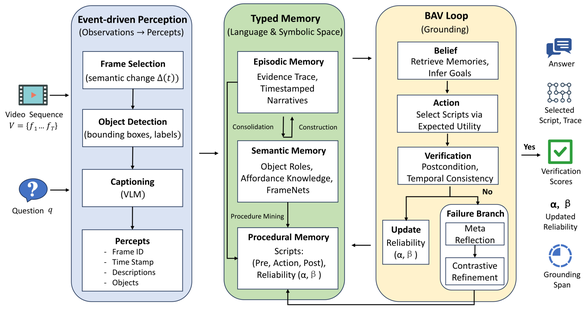
Q2: Can a compact model close the grounding gap through structure rather than parameter scale?

Q3: Can an agent learn continually from a video stream without gradient updates or catastrophic forgetting?

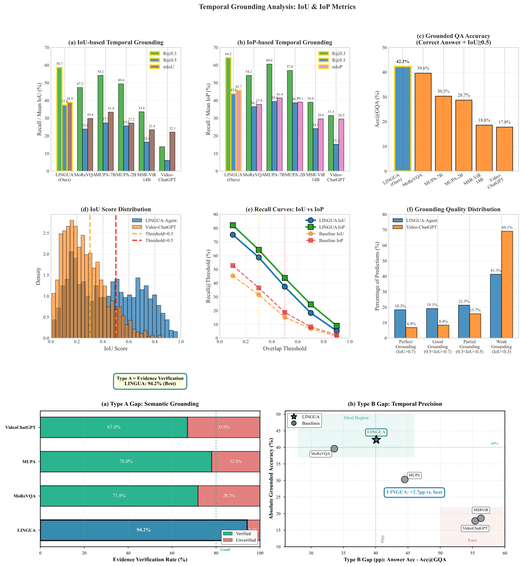
Contribution

1) Couples grounding with reasoning at inference time — not via implicit embeddings, pre/post-processing, or training-time alignment.
2) LINGUA: a typed-memory agent reasoning in linguistic belief states (episodic, semantic, procedural) on a compact Gemma3-4B backbone.
3) Event-driven perception keeps only 8–12% of frames while preserving 94% of question-relevant events.
4) A Belief–Action–Verification loop enforcing temporal and causal grounding with Bayesian reliability.
5) Gradient-free continual learning that improves over a video stream with no catastrophic forgetting.

Methodology

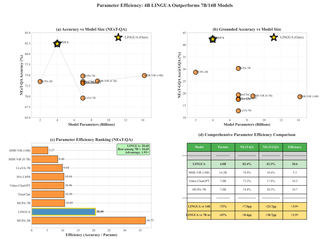


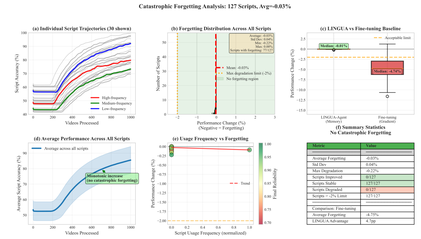
Experiments

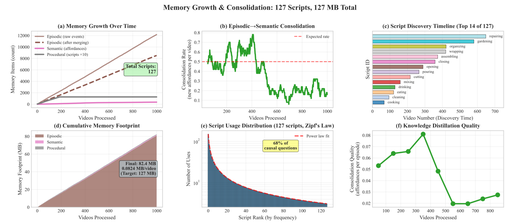


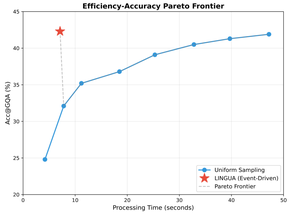
Performance Analysis

LINGUA sets a new accuracy–efficiency balance for grounded VideoQA. With a 4B backbone it reaches 82.4% on NExT-QA and 42.3% Acc@GQA — surpassing 7B–34B baselines while running 2.6× faster than dense-frame methods. It also attains 68.5% on Video-MME (69.4% long) and adapts continually from a video stream with no catastrophic forgetting.









Parameter efficiency: a 4B LINGUA outperforms 7B–14B systems. Upgrading the backbone from 4B to 11B adds only +0.5pp at 2.7× compute, indicating that structured reasoning — not parameter scale — drives grounded accuracy.

Continual learning without catastrophic forgetting: additive memory updates raise accuracy from 45.2% to 82.4% over 1,000 videos (84.2% at 2,000), with mean forgetting ≈ 0 — versus several points lost by gradient fine-tuning.

Saman Forouzandeh Wei Peng Xinghuo Yu Mahdi Jalili