

HIMMEL

Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding

Haopeng Jin¹, Wenlong Zhao¹, Tao Yu², Tiankun Yang²
Zhenyu Guan², ShiQuan Dong³, Yujia Yang²
Chubin Chen⁴†, Hongzhu Yi²†, Jinwen Luo¹†

¹ Tencent ² University of Chinese Academy of Sciences

³ Beijing Forestry University ⁴ Tsinghua University

NeurIPS 2026

github.com/Sunbeam23333/HIMMEL

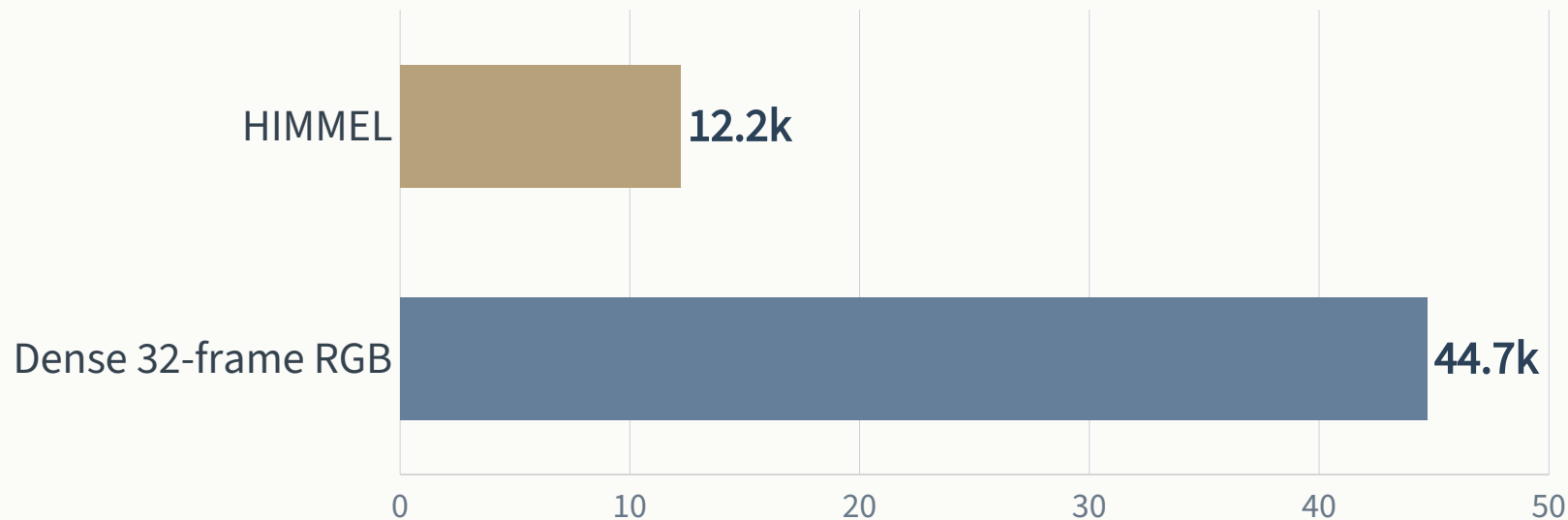


† Corresponding authors

Long videos need appearance and motion

Sparse frames preserve scene semantics.
Events between them can still matter.

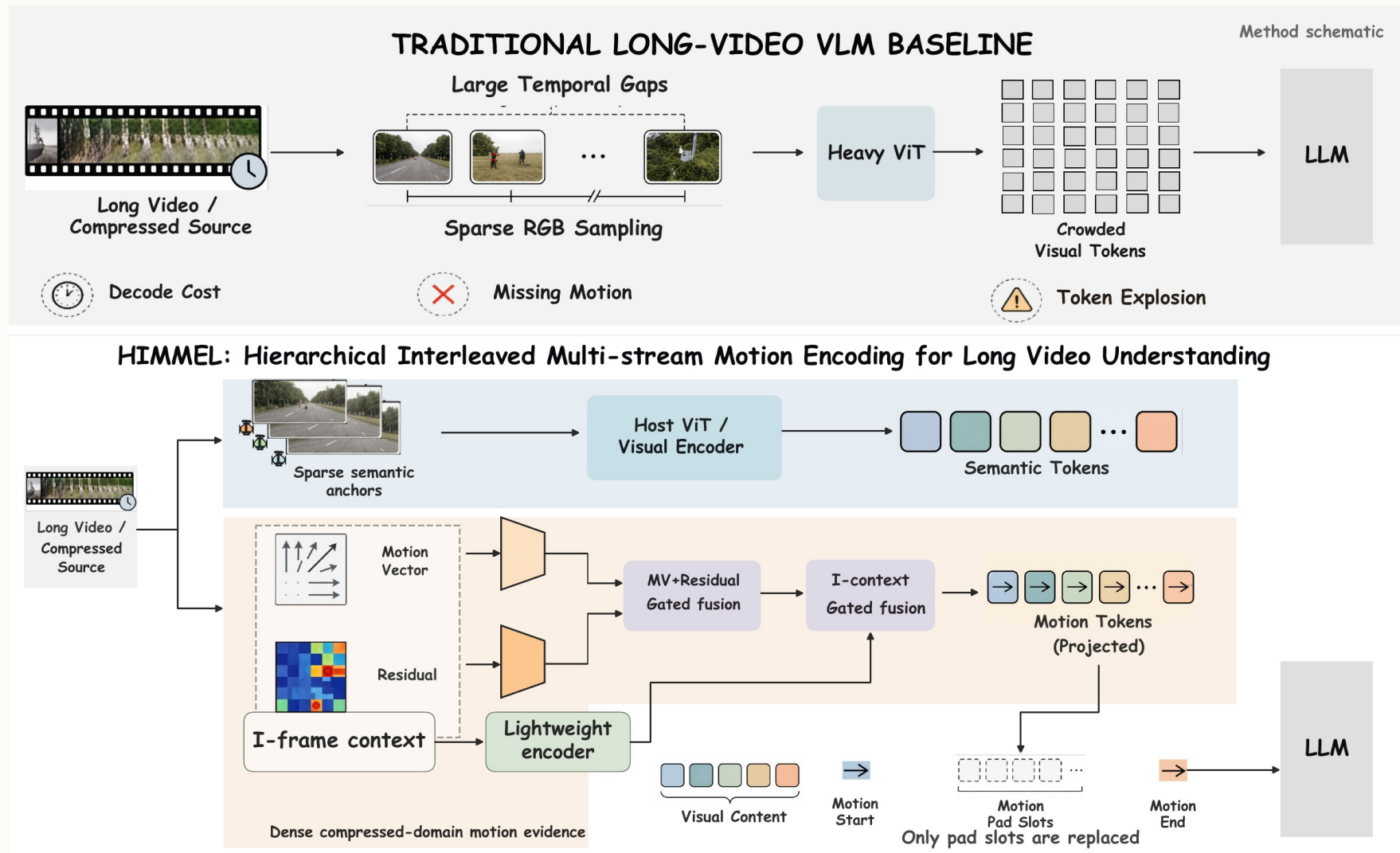
HIMMEL uses 8 semantic anchors and 64 motion tokens per interval.



3.6×

fewer visual-context tokens
at the reported operating
point

Two paths, one interleaved sequence



Only motion-pad slots receive motion tokens. The host visual pathway remains frozen.

Complementary interval signals



Appearance context

Objects and scene layout

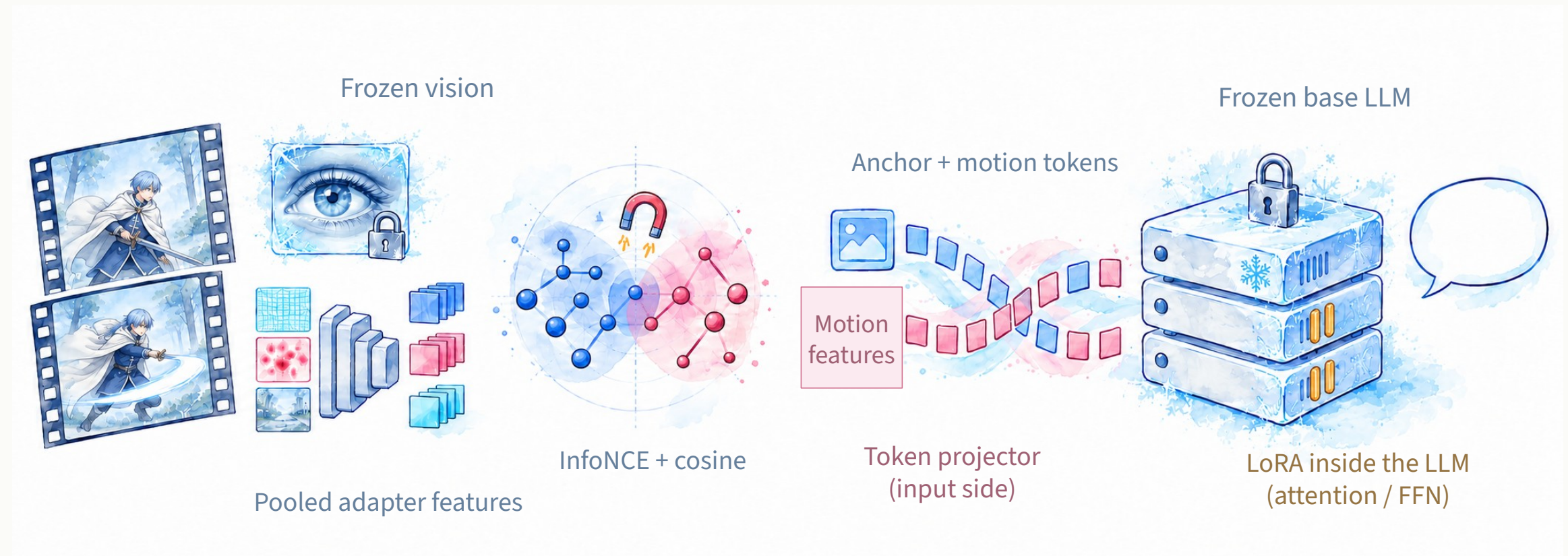
Motion-vector maps

Horizontal and vertical
displacement, plus magnitude

Residual maps

Decoded pixels minus
a motion-warped reference

Alignment before video instruction tuning



Stage 1: host-space alignment

Match pooled interval features to frozen endpoint visual differences with InfoNCE and cosine loss.

Stage 2: video QA

Project unpooled tokens into the LLM input.
Train adapter, projector and internal LLM LoRA.

Gains beyond video instruction tuning

Configuration	Visual tokens	Accuracy (%)
Dense host, untouched	44.7k	61.22
Dense host, same video SFT	44.7k	62.04
HIMMEL, same video SFT	12.2k	63.52

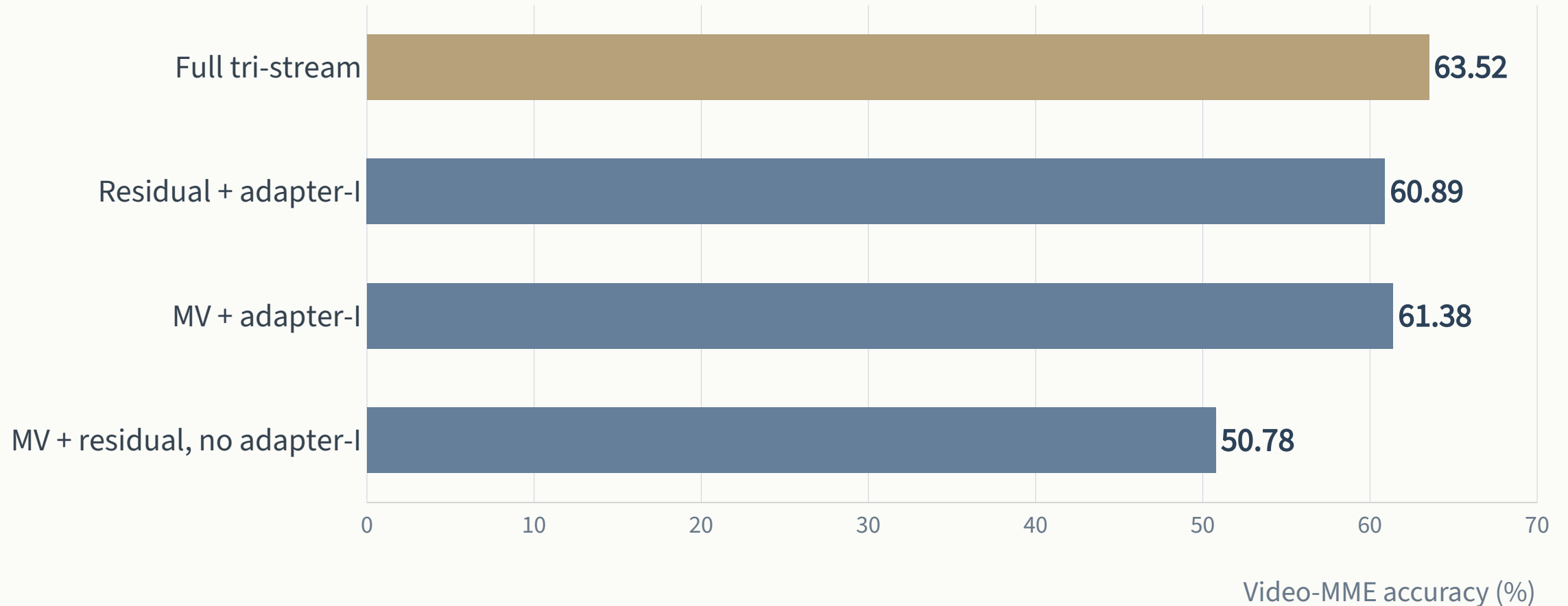
+1.48 pp

over the dense host with the same video SFT

+1.78 pp

over RGB anchors at the same 12.2k context
RGB control: 61.74%, HIMMEL: 63.52%

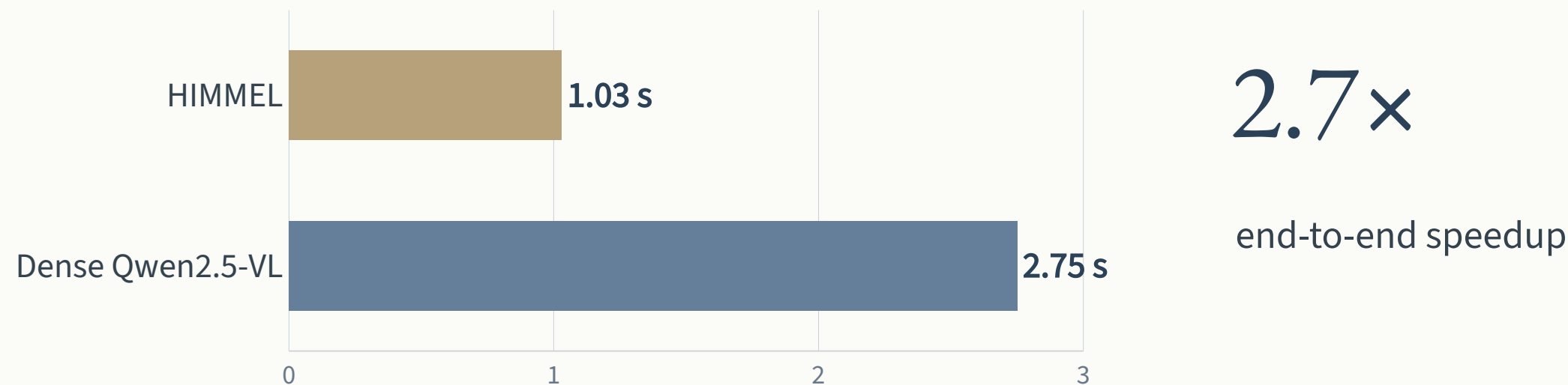
Appearance grounding supports motion fusion



The full tri-stream combines displacement, prediction error and appearance context.

All rows retain the host semantic-anchor path. Adapter-I means the separate appearance-context branch.

Lower end-to-end inference cost



Metric	Dense host	HIMMEL
Throughput (questions/s)	0.364	0.974
Peak allocated memory (GiB)	20.09	17.75

Latency includes extraction, encoding, prefill and generation. Memory uses 256 non-visual prompt tokens.

Video-MME gains across four hosts

Host	Dense baseline	HIMMEL	Gain (pp)
Qwen2.5-VL-7B	61.2	63.5	+2.3
Qwen3-VL-8B	62.7	64.9	+2.2
InternVL3-8B	62.3	64.3	+2.0
LLaVA-OV-7B	58.0	59.6	+1.6

The adapter improves each evaluated host at the compact visual-context operating point.

Codec compatibility and signal provenance

Input route	Motion field	Residual
Supported decoder side data	Exported block motion vectors	Motion-compensated pixel difference
Fallback route	Estimated block matching	Difference from the estimated warped reference

HEVC, VP9 and AV1 use the estimated route in this backend.

Dense decoding may still occur. The architectural saving comes from avoiding dense RGB processing through the expensive host ViT.

Small changes in mathematical capability

Training	MathVista testmini	GSM8K test
Untouched host	682/1000 68.20%	1137/1319 86.20%
Video-only SFT + LoRA	681/1000 68.10%	1129/1319 85.60%
With mathematical rehearsal	682/1000 68.20%	1136/1319 86.13%

Rehearsal replaces 10% of the video examples with 5% text mathematics and 5% visual mathematics.

These results support retention in the evaluated settings, not universal out-of-domain robustness.

Compact context, complementary motion evidence

HIMMEL allocates expensive host vision to sparse anchors and lightweight motion processing to dense intervals.

3.6×

fewer visual-context tokens

+1.48 pp

over the same-data SFT
baseline

2.7×

faster end-to-end
inference

Code and paper

github.com/Sunbeam23333/HIMMEL

Backup: gated fusion and token injection

Normalized inputs, difference and element-wise product

$$Z = [\text{LN}(P), \text{LN}(S), \text{LN}(P) - \text{LN}(S), \text{LN}(P) \odot \text{LN}(S)]$$

$$G = \text{sigmoid}(Z W^T + b)$$

$$F(P, S) = \text{LN}(G \odot P + (1 - G) \odot S + D(Z))$$

First fuse MV with residual. Then fuse the result with appearance context.

Stage 2 projects all 64 unpooled tokens per interval into the LLM width.
Out-of-place scatter replaces motion-pad embeddings and preserves gradients.