DUET-VLM: Dual-stage Unified Efficient Token reduction for VLM Training and Inference

Aditya Kumar Singh* Hitesh Kandala* Pratik Prabhanjan Brahma Zicheng Liu Emad Barsoum
Advanced Micro Devices, Inc. (AMD) *Equal contribution

CVPR JUNE 3-7, 2026 DENVER COLORADO

WHY DUET-VLM?

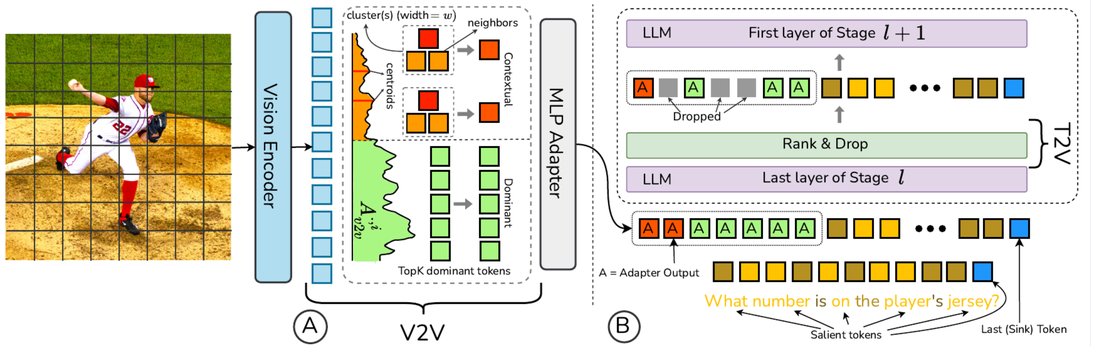
Vision tokens dominate VLM compute. DUET-VLM first merges redundant visual tokens, then lets text saliency drop the rest inside the LLM — **keeping accuracy while cutting compute.**

>99% accuracy retained
67% fewer visual tokens

>97% accuracy retained
89% fewer visual tokens

31% training time saved
LLaVA-1.5-7B backbone

METHOD: TWO-STAGE TOKEN COMPRESSION



(A) V2V merge
keep dominant tokens, locally merge context

(B) T2V prune
drop visual tokens guided by text saliency

Training uses the same compressed-token path as inference, so the model adapts to small visual budgets instead of treating pruning as a post-hoc trick.

RESULTS: ACCURACY SURVIVES COMPRESSION

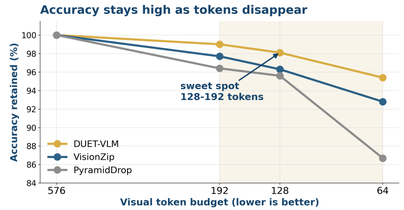


Fig. 1 Inference accuracy vs visual-token budget on LLaVA-1.5-7B (avg of 5 benchmarks).

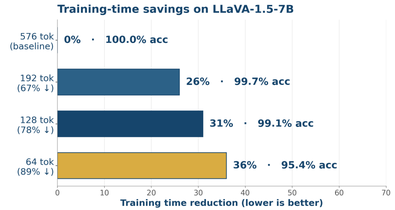


Fig. 2 End-to-end training time saved while preserving baseline accuracy.

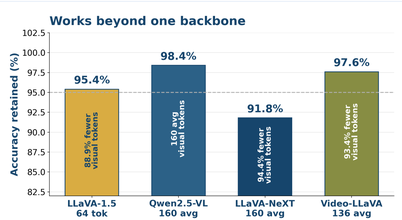


Fig. 3 DUET-VLM generalizes across LLaVA-1.5, LLaVA-NeXT, Qwen2.5-VL and Video-LLaVA backbones.

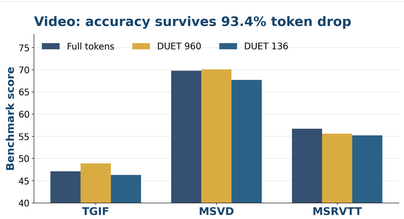
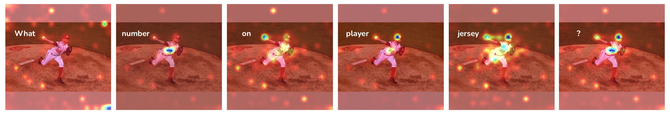


Fig. 4 Video-LLaVA-7B: 93.4% of visual tokens dropped, only ~2% accuracy lost.

QUALITATIVE EVIDENCE

Text-guided attention keeps the patches needed for the query.
The token budget shrinks, but semantically useful visual regions remain visible.



Query: What number is on the player's jersey? Salient tokens = [What, number, on, player, jersey, ?], Last token = ? Layer: 9/32

Query: What is the brand of this camera? Salient tokens = [What, brand, this, camera, ?], Last token = ? Layer: 24/32

TAKEAWAY + LINKS

What to remember
DUET-VLM is a plug-in token budget controller.
It keeps accuracy high across image and video VLMs.

Paper | Code | Project page

Interactive links: scan or click the QR cards
Paper | Code | AMD-AGI

DUET-VLM, CVPR 2026 | Token reduction for VLM training and inference | References: FastV, HiRED, PruMerge, PyramidDrop, VisionZip, FitPrune, SparseVLM, Video-LLaVA, LLaVA-1.5, LLaVA-NeXT