SOVEREIGN: How does content-adaptive tokenization affect the inference latency and accuracy of multimodal vision-language

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Modern multimodal large language models (MLLMs) typically keep the language model fixed and train a visual projector that maps the pixels into a sequence of tokens in its embedding space, so that images can be presented in essentially the same form as text. However, the language model has been optimized to operate on discrete, semantically meaningful tokens, while prevailing visual projectors transform an image into a long stream of continuous and highly correlated embeddings. This causes the visual tokens to behave differently from the word-like units that LLMs are originally trained to under

1 Introduction

Analysis of: A More Word-like Image Tokenization for MLLMs. Research goal: How does content-adaptive tokenization affect the inference latency and accuracy of multimodal vision-language models on high-resolution image datasets compared to fixed-patch baselines?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

5 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Modern multimodal large language models typically keep the language model fixed and train a visual projector that maps p	✓	0.35
Prevailing visual projectors transform an image into a long stream of continuous and highly correlated embeddings.	✓	0.29
Disentangled Visual Tokenization (DiVT) clusters patch embeddings into coherent semantic units, so each token correspond	✓	0.39
DiVT adapts its token budget to image complexity, providing an explicit accuracy-compute trade-off modifying neither the	✓	0.36
Across diverse multimodal benchmarks, DiVT matches or surpasses baselines with significantly fewer visual tokens.	✓	0.29
DiVT demonstrates robustness under limited token budgets, significantly reducing memory cost and latency while making vi	✓	0.31
The code for DiVT is available at https://github.com/snuviplab/DiVT.	✓	0.20

References

- http://arxiv.org/abs/2509.15704v2
- https://www.semanticscholar.org/paper/74fb8a716d2b6ee11572810136d7151a441cf7f2
- https://www.semanticscholar.org/paper/4fb29c98fe9d9a504095228b3fc8285331c2e6fb