SOVEREIGN: CAT: Content-Adaptive Image Tokenization

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Most existing image tokenizers encode images into a fixed number of tokens or patches, overlooking the inherent variability in image complexity. To address this, we introduce Content-Adaptive Tokenizer (CAT), which dynamically adjusts representation capacity based on the image content and encodes simpler images into fewer tokens. We design a caption-based evaluation system that leverages large language models (LLMs) to predict content complexity and determine the optimal compression ratio for a given image, taking into account factors critical to human perception. Trained on images with divers

1 Introduction

Analysis of: CAT: Content-Adaptive Image Tokenization. Research goal: How does the on-demand expert allocation in AnyExperts affect inference throughput and FLOPs efficiency on multimodal reasoning tasks, compared to fixed top-k routing in MoE-VLMs?.

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

13 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.7/10 $\rightarrow$ 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
CAT dynamically adjusts representation capacity based on image content and encodes simpler images into fewer tokens.	✓	0.38
CAT uses a caption-based evaluation system that leverages large language models (LLMs) to predict content complexity and	✓	0.35
CAT is trained on images with diverse compression ratios.	✓	0.24
CAT demonstrates robust performance in image reconstruction.	✓	0.24
CAT's variable-length latent representations are used to train Diffusion Transformers (DiTs) for ImageNet generation.	✓	0.27
CAT improves the FID score over fixed-ratio baselines trained with the same flops.	✓	0.32
CAT boosts inference throughput by 18.5%.	✓	0.19

References

- https://www.semanticscholar.org/paper/d514a6cddf5b2da9d4c8f1288c8f79fc5a6ba972
- https://www.semanticscholar.org/paper/054f51d1286ea3eae4893fabddf90389164347dd
- http://arxiv.org/abs/2604.23996v1