Supplementary Materials

In the appendix, we provide comprehensive details including benchmark specifications, extended ablation study results, and supplementary visualizations.

A. Benchmark Details

We conducted experiments on these widely used benchmarks:

GQA. The GQA [5] benchmark evaluates compositional visual reasoning through scene graph-driven question-answering. It assesses spatial understanding, attribute recognition, and multi-step inference capabilities using real-world images paired with structurally annotated questions.

POPE. POPE [7] measures object hallucination in large vision-language models via binary presence queries. It employs accuracy, precision, recall, and F1-score across three object sampling strategies to characterize false positive generation tendencies.

MME. MME [2] provides a comprehensive evaluation across 14 perceptual and cognitive subtasks through manually curated instruction-response pairs, specifically designed to mitigate evaluation data contamination.

ScienceQA. ScienceQA [9] benchmarks multimodal scientific reasoning across natural, social, and language sciences. Its hierarchical taxonomy and multi-choice format test models' capacity for domain-specific knowledge integration and logical inference.

TextVQA. TextVQA [10] evaluates text-rich visual reasoning by requiring models to read, localize, and comprehend textual elements within natural images to answer open-ended questions.

VizWiz. VizWiz [4] contains 31,000+ visual questions captured by blind users via mobile devices. This unique dataset challenges models with low-quality images, conversational spoken-style queries, and inherent unanswerable questions due to visual ambiguity.

MMBench. MMBench [8] employs a three-level hierarchical taxonomy spanning 20 fine-grained ability dimensions. Its circular evaluation strategy and comprehensive coverage enable systematic assessment of both perception and reasoning capabilities.

Table 1. **Ablation study on approximation error metrics across three image understanding tasks.** Experiments are performed with LLaVA-1.5-7B and Qwen2.5-VL-7B at token compression ratios of 11.1% and 10%, respectively.

Model	Method	GQA	POPE	MME
LLaVA-1.5-7B	Cosine Similarity	54.1	80.9	1576
	L1 Norm	56.6	84.0	1685
	L2 Norm	**56.9**	**84.4**	**1714**
Qwen2.5-VL-7B	Cosine Similarity	54.7	81.1	1909
	L1 Norm	55.9	82.0	2014
	L2 Norm	**56.1**	**82.4**	**2055**

VQA-v2. VQA-v2 [3] evaluates general visual understanding using 265K real-world images with open-ended questions. Each question includes 10 human-annotated answers to address annotation ambiguity and enable robust performance measurement.

TGIF. TGIF [6] extends visual QA to the video domain with 165K QA pairs derived from animated GIFs. It comprises four task categories: three spatio-temporal reasoning tasks (repetition counting, action localization, state transition) and single-frame QA, requiring integrated motion and content understanding.

MSVD. The MSVD [1] benchmark comprises 1,970 video clips with 50.5K open-ended questions across five categories (what, who, how, when, where), supporting both video question answering and captioning evaluation.

MSRVTT. MSRVTT [11] challenges models with 243K questions over 10K diverse video clips. Its five question types assess comprehensive video understanding by requiring integration of spatial, temporal, and semantic information.

B. More Ablation Studies

Our work employs approximation errors to quantify token significance. Table 1 presents a comparative evaluation of three metrics: ℓ_1-norm, ℓ_2-norm and Cosine Similarity. Results demonstrate that the ℓ_2-norm metric yields superior performance on both LLaVA-1.5-7B and Qwen2.5-VL-7B models, motivating its adoption as the approximation error measure in ApET.

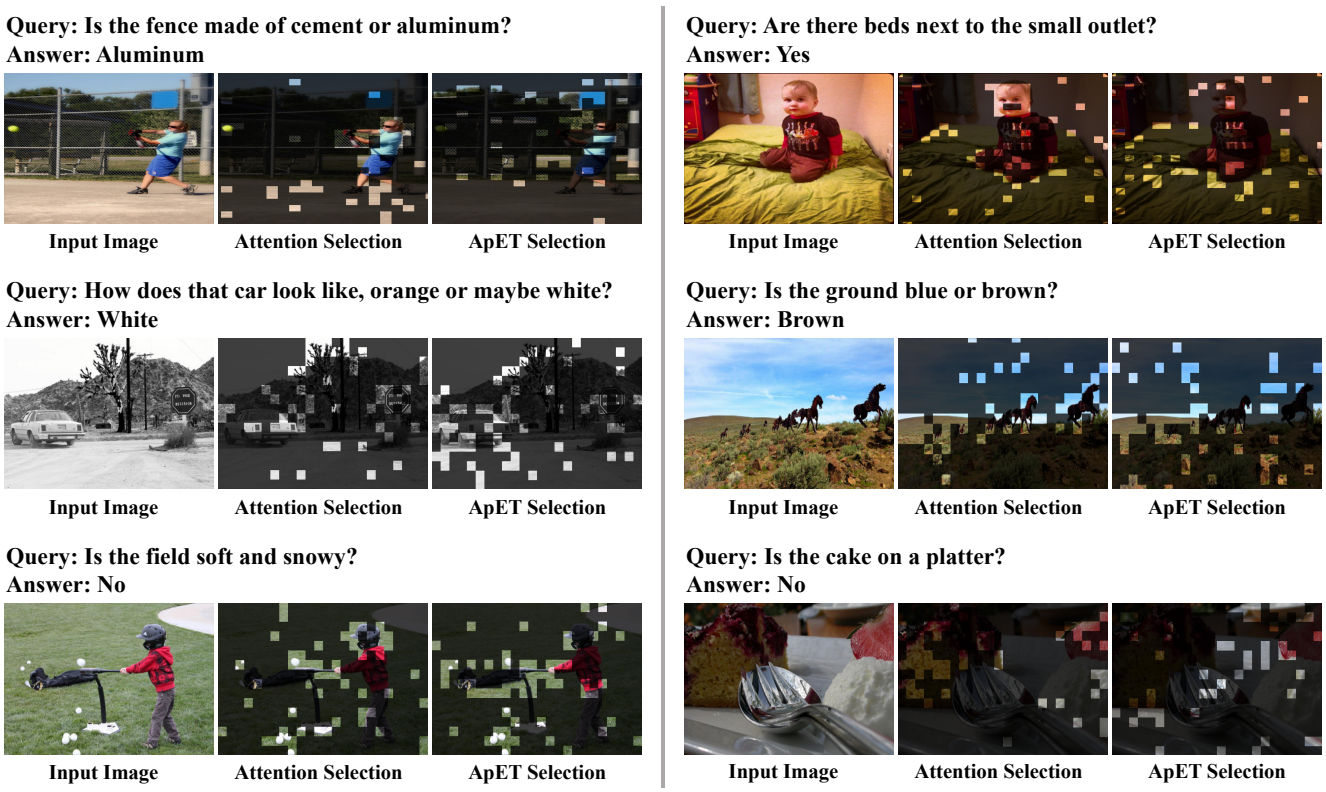

Figure 1. **Visualization of token compression.** We present more representative failure cases in which attention-driven token selection misguides the final prediction. For each case, we visualize the input question and its ground-truth answer, the original image, the subset of visual tokens preserved when ranking by attention weights, and the subset preserved by the approximation-error criterion proposed in this work. The retained tokens are highlighted for clear comparison.

C. More Visualizations

In Figure 1, we present supplementary visualizations of ApET's token compression across diverse scenarios. These results demonstrate that approximation-error-based token filtering effectively retains semantically rich visual tokens without requiring attention mechanisms.

References

[1] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 1

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 1

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[4] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[6] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1

[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[8] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He,

Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[9] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 1

[10] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[11] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1