A Comprehensive Study on Visual Token Redundancy for Discrete Diffusion-based Multimodal Large Language Models

Supplementary Material

1. Experimental Details

1.1. Benchmarks

We conduct our experiments across a diverse suite of multimodal benchmarks. For image understanding, we evaluate on ten datasets: MME [4], SQA [14], GQA [6], POPE [9], MMB [12], TVQA [19], CQA [15], MMMUP [21], IVQA [17], and DVQA [16]. For video understanding, we further assess performance on two benchmarks: VMME [5] and Video Detail Caption [13].

MME. MME is a comprehensive benchmark evaluating multimodal models on 14 perception and cognition subtasks, including OCR, counting, spatial localization, and visual recognition of scenes, landmarks, and artworks. All tasks are formulated as binary judgment questions with curated instruction–answer pairs to ensure fairness. We report the standard perception score on 2,374 image–question pairs.

SQA. ScienceQA evaluates multimodal reasoning and zero-shot generalization in scientific domains. It covers natural, language, and social sciences, with questions organized hierarchically across multiple topics and skills. Each question is a multiple choice question, often paired with an illustrative image. We evaluate on the image dataset of 2,017 question–answer pairs.

GQA. GQA evaluates structured visual reasoning using images, scene graphs, and automatically generated questions. Each image is paired with a scene graph from the Visual Genome dataset [7], providing detailed objects, attributes, and relations. We follow standard protocol and report accuracy on the test-dev set with 12,578 image–question pairs.

POPE. POPE evaluates object hallucination in vision–language models using binary questions about object presence in images from the MSCOCO dataset [10]. Performance is measured by the average F1 score over three sampling strategies, covering 8,910 image–question pairs.

MMB. MMBench provides a hierarchical evaluation of multimodal understanding across three levels—perception and reasoning (L1), six sub-skills (L2), and 20 tasks (L3)—each formulated as multiple-choice questions. It is available in English and Chinese versions, containing 4,377 and 4,329 image–question pairs, respectively. We evaluate on MMBench-EN subset.

TVQA. TextVQA benchmarks VQA models that must read and reason over text in natural images. It comprises 45,336 questions on 28,408 images (from text-rich Open Images categories), with 10 human answers per question. We follow the standard setting and evaluate accuracy on this dataset.

CQA. ChartQA benchmarks question answering over chart images that require both visual and logical reasoning. It includes 9,608 human-written questions and 23,111 questions generated from chart summaries, spanning 20,882 real-world charts collected from Statista, Pew Research, Our World in Data, and the OECD. Answers are often open-vocabulary and may involve arithmetic or comparisons. We follow the dataset's official evaluation protocol.

MMMUP. MMMU-Pro is a strengthened version of MMMU that aims to test genuine multimodal understanding and reasoning. It (i) filters out items solvable by text-only models, (ii) augments candidate options, and (iii) introduces the vision-only input setting MMMU-Pro Vision, where questions and options are embedded directly into images so models must truly "see" and read. In our experiments, we report results on the Vision subset following the paper's protocol.

IVQA. InfoVQA evaluates VQA on infographics that require joint reasoning over layout, embedded text, graphical elements, and data visualizations. The dataset contains 5,485 images with 30,035 questions; answers are mainly extractive, with some numerical ones derived via counting, sorting, or simple arithmetic. We follow the official protocol and report accuracy.

DVQA. DocVQA focuses on question answering over real document images that require both reading and layout understanding. The dataset contains 12,767 document images of varied types and content, paired with about 50,000 human-annotated question–answer pairs. Each question involves information extraction, reasoning across text blocks, or interpreting document structure. We follow the dataset's official evaluation protocol.

VMME. VideoMME is a large-scale benchmark for evaluating video understanding in LVLMs. It includes 900 videos (≈254 hours) from six domains and 30 subcategories, covering short (≤2 min), medium (4–15 min), and long (30–60 min) durations. Each video has three expert-authored multiple-choice questions, yielding 2,700 video–question pairs. We evaluate on the full dataset.

VDC. Video Detail Caption is a video captioning benchmark released by LMMs-Lab, where each video clip is paired with a detailed textual description. The test set contains 499 samples, each including a video name, a question prompt, and an answer paragraph. We follow the official evaluation protocol and assess performance using GPT-4o-mini as the evaluator.

1.2. Backbone Models

LLaDA-V. LLaDA-V [20] represents the pure diffusion paradigm in multimodal large language modeling. It extends the LLaDA diffusion language backbone with a SigLIP-2 vision encoder and a lightweight MLP projector, enabling multimodal understanding entirely through masked diffusion rather than next-token prediction. As a purely diffusion-trained model, LLaDA-V exemplifies non-autoregressive probabilistic reasoning and demonstrates strong scalability across image, document, and video understanding benchmarks.

LaViDa-Dream. LaViDa-Dream [8] represents the autoregressive-to-diffusion adaptation paradigm. It builds on Dream-7B, a discrete diffusion language model (DLM) adapted from autoregressive pretraining, and extends it to the multimodal setting through visual instruction tuning. By incorporating techniques such as complementary masking and Prefix-DLM caching, LaViDa-Dream achieves efficient multimodal reasoning while exemplifying the AR-to-diffusion adaptation route in dMLLMs.

1.3. Token Compression Methods

ToMe. ToMe [2] is a training-free efficiency method that accelerates inference by merging similar tokens instead of pruning them. It computes pairwise similarity between attention keys and merges the most redundant token pairs during encoding through a fast bipartite matching algorithm.

DivPrune. DivPrune [1] formulates visual token pruning as a diversity-driven token selection problem. It defines a min–max diversity objective, encouraging the retained tokens to be maximally dissimilar to each other, and applies a greedy selection strategy to iteratively preserve the most informative and diverse subset of visual tokens.

Benchmark(s)	gen_length	block_length	gen_steps	think_mode
LLaDA-V [20]				
MME, SQA, GQA, POPE, MMB, TVQA, MMMUP, VMME	2	2	2	no_think
CQA	16	16	8	no_think
DVQA, IVQA	32	32	16	no_think
VDC	128	128	64	think
LaViDa-Dream [8]				
MME, SQA, GQA, POPE, MMB, TVQA, MMMUP	4	4	2	no_think
CQA	16	16	8	no_think
DVQA, IVQA	32	32	16	no_think

Table 1. **Generation hyperparameters used for different benchmarks.** The settings largely follow the default configurations of the respective backbone models, with minor adjustments to ensure stable decoding across short- and long-answer tasks.

FastV. FastV [3] is a training-free method that accelerates vision–language models by pruning redundant visual tokens in the early decoding stage. It removes the least informative tokens after the second LLM layer based on averaged attention scores.

VTW. VTW [11] is a training-free acceleration method that withdraws all visual tokens after a specific transformer layer to reduce inference cost in vision–language models. The withdrawal layer is chosen via a KL divergence criterion, enabling VTW to cut FLOPs and memory usage by over 40% without significant performance degradation.

SparseVLM. SparseVLM [22] introduces adaptive cross-modal sparsity to reduce redundancy in both visual and textual tokens. It ranks token importance via cross-modal attention, dynamically applies different sparsity ratios to vision and language streams, and employs a token recycling mechanism that reuses informative pruned tokens to preserve contextual completeness.

TRIM. TRIM [18] is a training-free token reduction method that measures text–image similarity in the CLIP representation space to rank visual tokens. It selects important tokens via an IQR-based threshold and appends an aggregated representation of unselected tokens.

2. Generation Hyperparameters

We summarize in Table 1 the generation hyperparameters used in our experiments for both the LLaDA-V and LaViDa-Dream backbone models. The generation settings generally follow the default configurations provided in the original model implementations, with minor adjustments to ensure stable decoding across short- and long-answer tasks. Specifically, for LaViDa-Dream, the parameters

`gen_length`, `block_length`, and `gen_steps` are set to 4, 4, and 2, respectively, for the group of benchmarks including MME, ScienceQA, and GQA to ensure decoding stability.

References

[1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025. 2

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 2

[3] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 1

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[6] 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

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. 1

[8] Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*, 2025. 2

[9] 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

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1

[11] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5334–5342, 2025. 2

[12] 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

[13] LMMs-Lab. Videodetailcaption. Hugging Face Dataset, 2024. 1

[14] 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. In *Advances in Neural Information Processing Systems*, pages 2507–2521, 2022. 1

[15] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[16] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1

[17] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 1

[18] Alfredo Garrachón Ruiz, Tomás de la Rosa, and Daniel Borrajo. Trim: Token reduction and inference modeling for cost-effective language generation. *arXiv preprint arXiv:2412.07682*, 2024. 2

[19] 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

[20] Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025. 2

[21] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 1

[22] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 2