

Are We Using the Right Benchmark: An Evaluation Framework for Visual Token Compression Methods

Chenfei Liao^{1,2,6} Wensong Wang^{3,2} Zichen Wen^{2,5} Xu Zheng^{1,4,6}
Yiyu Wang² Haocong He² Yuanhuiyi Lyu^{1,6} Lutao Jiang^{1,6} Xin Zou^{1,6}
Yuqian Fu⁴ Bin Ren⁷ Linfeng Zhang^{2,†} Xuming Hu^{1,6,†}

¹The Hong Kong University of Science and Technology (Guangzhou)

²Shanghai Jiao Tong University ³Northeastern University

⁴INSAIT, Sofia University “St. Kliment Ohridski” ⁵Shanghai AI Laboratory

⁶The Hong Kong University of Science and Technology ⁷MBZUAI

Abstract

Recent efforts to accelerate inference in Multimodal Large Language Models (MLLMs) have largely focused on visual token compression. The effectiveness of these methods is commonly evaluated by measuring the accuracy drop on existing MLLM benchmarks before and after compression. However, these benchmarks are originally designed to assess general perception and reasoning abilities, rather than the specific challenges posed by visual token compression, leading to a fundamental task mismatch. In this work, we uncover a counterintuitive yet consistent phenomenon: **simple image downsampling outperforms many advanced visual token compression methods across multiple widely used benchmarks**. Through a comprehensive empirical study spanning eight popular benchmarks and multiple state-of-the-art compression techniques, we show that (i) current benchmarks contain substantial noise (task-irrelevant samples) for evaluating visual token compression, and (ii) downsampling can act as an effective data filter that distinguishes between simple and difficult samples with respect to compression sensitivity. Motivated by these findings, we propose **VTC-Bench**¹², an evaluation framework that explicitly leverages downsampling as a discriminator to denoise existing benchmarks, enabling a fairer and more meaningful additional assessment of visual token compression methods.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown impressive abilities in understanding, reasoning, and generating content across vision and language (Chen et al., 2024d; Kang et al., 2025; Zou et al., 2025b; Chai et al., 2024; Wen

et al., 2026), enabling applications such as embodied AI (Yin et al., 2024; Fu et al., 2025; Zhang et al., 2025c,b, 2026). However, their efficiency is often constrained by the high computational cost of processing images, particularly at high resolutions (Liu et al., 2025). This bottleneck arises because visual tokens, derived from image patches, typically far outnumber textual tokens, leading to substantial memory consumption and inference latency (Wang et al., 2025; Chen et al., 2025a; Wen et al., 2025b). To mitigate this issue, numerous visual token compression methods have been proposed to reduce redundancy while retaining essential visual information (Yang et al., 2025a; Xing et al., 2024; Xiong et al., 2025; Zou et al., 2025a).

Yet, these methods are typically evaluated on general MLLM benchmarks (Li et al., 2024b), which are not designed for compression, therefore failing to provide appropriate evaluation criteria. Thus, in this paper, we uncover a surprising finding: as in Figure 1, **simple image downsampling consistently outperforms many advanced compression methods across multiple widely used benchmarks**. This suggests that current evaluation frameworks don’t adequately capture the challenges inherent in visual token compression.

To investigate this, we conduct a comprehensive empirical study comparing multiple state-of-the-art visual token compression methods against a simple downsampling baseline across eight widely used benchmarks. Based on the study results in Table 1 and 2, two crucial findings are concluded: ① The counterintuitive phenomenon mentioned above generally exists in popular benchmarks, proving that **current benchmarks are noisy for the visual token compression task**. Many task-irrelevant samples serve as noise when evaluating visual token compression methods. ② The correct sample group under downsampling methods has achieved significantly better accuracy than the incorrect sample group under downsampling methods across various

[†]Corresponding authors.

¹Code: <https://github.com/Chenfei-Liao/VTC-Bench>.

²Project Page: <https://chenfei-liao.github.io/VTC-Bench-Page>.

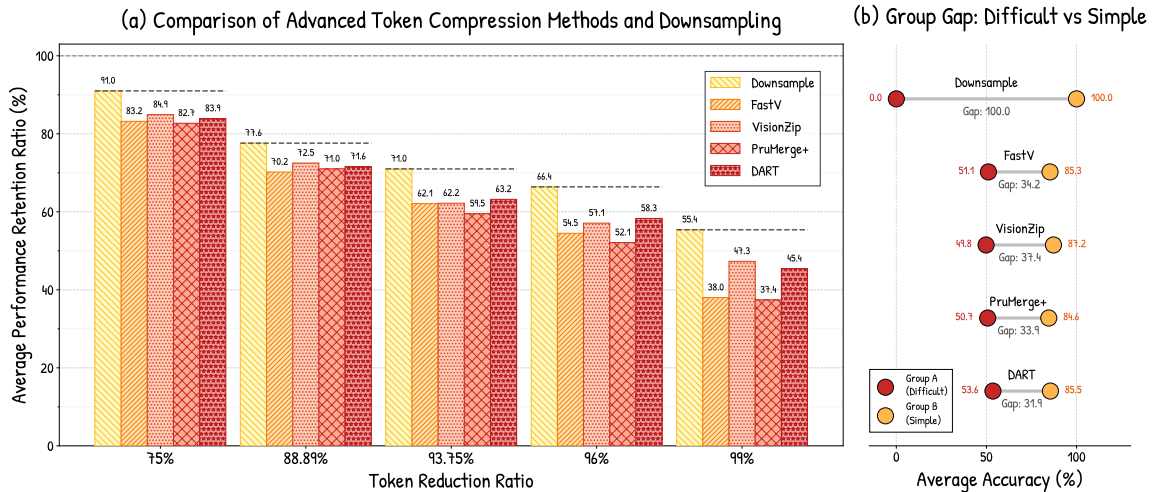


Figure 1: (a) Average Performance Retention Ratio (APRR) of five visual token compression methods on eight benchmarks (Model: Qwen2-VL-7B; Benchmark: as shown in Table 1), supporting *the observation (i): current benchmarks contain substantial noise (task-irrelevant samples) for evaluating visual token compression*. (b) Comparison of advanced token compression methods and downsampling on Qwen2-VL-7B by groups at 75% compression ratio, supporting *the observation (ii): downsampling can act as an effective data filter that distinguishes between simple and difficult samples with respect to compression sensitivity*. Group A refers to the “difficult” samples. Group B refers to the “simple” samples. The grouping strategy is explained in Section 3.4.

compression methods and benchmarks, proving that *downsampling can serve as a data filter to evaluate the difficulty of samples upon the visual token compression task*.

Based on these findings, we propose VTC-Bench, an evaluation framework specifically designed to optimize current existing benchmarks, aiming to serve as an additional reference benchmark for evaluating visual token compression methods. By explicitly distinguishing between “simple” and “difficult” samples through downsampling, VTC-Bench adaptively and fairly selects “difficult” samples that satisfy the requirements of evaluating visual token compression methods.

Overall, our contributions are threefold: ① We identify and validate the data noise (task-irrelevant samples) in existing MLLM benchmarks on the visual token compression task. ② We introduce a data filtering mechanism using downsampling as a discriminator to categorize benchmark samples by difficulty. ③ We propose VTC-Bench, the first evaluation framework tailored for fairly evaluating visual token compression methods, aiming to foster more meaningful progress in this emerging field.

2 Related Work

2.1 Visual Token Compression for MLLMs

Since visual tokens typically outnumber text tokens in MLLMs, compressing visual tokens has

emerged as a promising strategy to accelerate inference (Yao et al., 2026; Shao et al., 2025; Wu et al., 2026; Liu et al., 2025; Yang et al., 2025a; Wang et al., 2025; Shang et al., 2024). Leveraging the inherent redundancy in visual tokens, a variety of *training-free* methods have been proposed. FastV (Chen et al., 2024b), the first to explore visual token compression in MLLMs, prunes redundant tokens based on their average attention scores. Other subsequent methods mainly have two technical features: a) dividing the compression process into multiple stages (Xing et al., 2024; Han et al., 2026; Liu et al., 2024a; Chen et al., 2024a; Endo et al., 2025; Chen et al., 2025b) to enable more precise identification of redundant tokens; b) seeking better metrics as the basis for visual token compression (Han et al., 2026; Zhang et al., 2025a; Xu et al., 2024; Zou et al., 2025a; Wen et al., 2025a; Zheng et al., 2025). Specifically, VFlowOpt (Yang et al., 2025c) introduces the importance map and a recycling mechanism to achieve a progressive and effective token pruning. Similarly, G-Prune (Jiang et al., 2025) identifies critical tokens through a graph-based perspective, while LUV (Zheng et al., 2025) applies a spectrum pruning unit to LLM, gradually pruning redundant visual tokens. In contrast, DART (Wen et al., 2025a) instead prioritizes token duplication as a key criterion, achieving surprisingly strong compress-

Table 1: Comparison of Advanced Token Compression Methods and Downsampling on Qwen2-VL-7B. APRR refers to *the average performance retention ratio*, which is the average value of the retention ratio compared to the vanilla performance, across the eight benchmarks. The right side of each cell shows *the relative change percentage compared to the performance of the downsampling method*.

Method	GQA	MMB	MMB ^{CN}	MME	POPE	MMStar	OCRBench	ChartQA	APRR
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Upper Bound. All Tokens (100%)									
Vanilla	62.3	78.9	78.0	2306	88.4	57.1	80.7	81.6	100.0
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 75.00%)									
+ Downsample (Baseline)	59.2	75.0	73.8	2259	86.2	50.1	64.9	65.0	91.0
+ FastV (Chen et al., 2024b)	57.0 ↓3.7%	73.7 ↓1.7%	73.1 ↓0.9%	2083 ↓7.8%	84.5 ↓2.0%	44.6 ↓11.0%	42.0 ↓35.3%	58.1 ↓10.6%	83.2 ↓8.6%
+ VisionZip (Yang et al., 2025a)	58.6 ↓1.0%	71.1 ↓5.2%	70.5 ↓4.5%	2062 ↓8.7%	87.1 ↑1.0%	47.2 ↓5.8%	42.1 ↓35.1%	66.9 ↑2.9%	84.9 ↓6.7%
+ PruMerge+ (Shang et al., 2024)	59.4 ↑0.3%	72.1 ↓3.9%	72.0 ↓2.4%	2044 ↓9.5%	87.2 ↑1.2%	48.0 ↓4.2%	33.9 ↓47.8%	56.2 ↓13.5%	82.7 ↓9.1%
+ DART (Wen et al., 2025a)	56.9 ↓3.9%	72.5 ↓3.3%	70.2 ↓4.9%	2066 ↓8.5%	84.7 ↓1.7%	47.2 ↓5.8%	52.5 ↓19.1%	52.7 ↓18.9%	83.9 ↓7.8%
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 88.89%)									
+ Downsample (Baseline)	55.5	69.0	70.2	2127	82.9	44.0	48.8	24.8	77.6
+ FastV (Chen et al., 2024b)	52.3 ↓5.8%	65.0 ↓5.8%	65.5 ↓6.7%	1854 ↓12.8%	77.4 ↓6.6%	40.3 ↓8.4%	25.9 ↓46.9%	32.9 ↑32.6%	70.2 ↓9.5%
+ VisionZip (Yang et al., 2025a)	53.3 ↓4.0%	62.9 ↓8.8%	63.0 ↓10.3%	1820 ↓14.4%	83.6 ↑0.8%	40.2 ↓8.6%	25.1 ↓48.6%	48.4 ↑95.2%	72.5 ↓6.6%
+ PruMerge+ (Shang et al., 2024)	54.8 ↓1.3%	62.2 ↓9.9%	61.3 ↓12.7%	1806 ↓15.1%	84.3 ↑1.7%	38.4 ↓12.7%	22.2 ↓54.5%	44.2 ↑78.2%	71.0 ↓8.5%
+ DART (Wen et al., 2025a)	51.9 ↓6.5%	61.3 ↓11.2%	61.8 ↓12.0%	1915 ↓10.0%	80.5 ↓2.9%	39.8 ↓9.5%	41.0 ↓16.0%	30.8 ↓24.2%	71.6 ↓7.7%
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 93.75%)									
+ Downsample (Baseline)	52.6	66.4	66.8	1994	79.5	40.9	40.3	12.7	71.0
+ FastV (Chen et al., 2024b)	49.0 ↓6.8%	57.1 ↓14.0%	57.9 ↓13.3%	1684 ↓15.5%	74.9 ↓5.8%	37.5 ↓8.3%	18.7 ↓53.6%	20.6 ↑62.2%	62.1 ↓12.5%
+ VisionZip (Yang et al., 2025a)	49.0 ↓6.8%	54.8 ↓17.5%	54.0 ↓19.2%	1704 ↓14.5%	80.2 ↓0.9%	35.2 ↓13.9%	15.9 ↓60.5%	28.0 ↑120.5%	62.2 ↓12.4%
+ PruMerge+ (Shang et al., 2024)	48.7 ↓7.4%	48.4 ↓27.1%	48.1 ↓28.0%	1679 ↓15.8%	79.2 ↓0.4%	33.2 ↓18.8%	14.4 ↓64.3%	30.0 ↑136.2%	59.5 ↓16.2%
+ DART (Wen et al., 2025a)	49.2 ↓6.5%	53.4 ↓19.6%	54.0 ↓19.2%	1786 ↓10.4%	78.1 ↓1.8%	33.6 ↓17.8%	33.7 ↓16.4%	19.2 ↑51.2%	63.2 ↓11.0%
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 96.00%)									
+ Downsample (Baseline)	50.1	62.0	61.4	1938	78.8	37.5	32.3	11.7	66.4
+ FastV (Chen et al., 2024b)	46.1 ↓8.0%	43.9 ↓29.2%	46.6 ↓24.1%	1589 ↓18.0%	72.4 ↓8.1%	33.6 ↓10.4%	14.4 ↓55.4%	15.8 ↑35.0%	54.5 ↓17.9%
+ VisionZip (Yang et al., 2025a)	46.4 ↓7.4%	49.5 ↓20.2%	50.0 ↓18.6%	1628 ↓16.0%	77.8 ↓1.3%	33.4 ↓10.9%	12.0 ↓62.8%	19.4 ↑65.8%	57.1 ↓14.0%
+ PruMerge+ (Shang et al., 2024)	45.0 ↓10.2%	39.1 ↓36.9%	40.9 ↓33.4%	1544 ↓20.3%	74.0 ↓6.1%	30.5 ↓18.7%	10.5 ↓67.5%	20.9 ↑78.6%	52.1 ↓21.5%
+ DART (Wen et al., 2025a)	45.6 ↓9.0%	47.9 ↓22.7%	48.2 ↓21.5%	1701 ↓12.2%	74.7 ↓5.2%	31.7 ↓15.5%	29.3 ↓9.3%	16.6 ↑41.9%	58.3 ↓12.2%
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 99.00%)									
+ Downsample (Baseline)	43.5	51.6	51.9	1589	72.8	33.8	13.2	12.1	55.4
+ FastV (Chen et al., 2024b)	38.2 ↓12.2%	23.9 ↓53.7%	24.5 ↓52.8%	1189 ↓25.2%	55.0 ↓24.5%	26.1 ↓22.8%	5.8 ↓56.1%	11.9 ↓1.7%	38.0 ↓31.4%
+ VisionZip (Yang et al., 2025a)	41.9 ↓3.7%	40.5 ↓21.5%	40.5 ↓22.0%	1335 ↓16.0%	65.5 ↓10.0%	30.8 ↓8.9%	4.9 ↓62.9%	12.8 ↑5.9%	47.3 ↓14.6%
+ PruMerge+ (Shang et al., 2024)	39.0 ↓10.3%	23.7 ↓54.1%	24.4 ↓53.0%	1165 ↓26.7%	51.6 ↓29.1%	25.7 ↓24.0%	3.5 ↓73.5%	13.9 ↑14.9%	37.4 ↓32.5%
+ DART (Wen et al., 2025a)	40.5 ↓6.9%	30.8 ↓40.3%	30.7 ↓40.8%	1346 ↓15.3%	60.0 ↓17.6%	28.8 ↓14.8%	23.2 ↑75.8%	11.8 ↓2.5%	45.4 ↓18.1%

sion performance. Beyond these, GreedyPrune (Pei et al., 2025) and ToDRE (Li et al., 2025) cast token compression as an optimization problem and employ greedy algorithms to search for efficient pruning strategies.

However, as in Section 3, we have a surprising observation: across most MLLM benchmarks, these sophisticated visual token compression methods underperform compared to simply reducing the original image resolution, which motivates a deeper investigation into the underlying causes.

2.2 MLLM Benchmarks

Existing MLLM benchmarks primarily focus on areas such as perception and reasoning (Li et al., 2024b; Song et al., 2025). For example, MME (Yin et al., 2024), MMBench (Liu et al., 2024b), and MM-Vet (Yu et al., 2023, 2024) provide broad perception-focused evaluations of MLLMs’ visual understanding. In parallel, domain benchmarks target specific applications such as autonomous driving (Sima et al., 2024; Qian et al., 2024) and remote sensing (Muhtar et al., 2024). For visual token compression in MLLMs, only one bench-

mark currently exists: EffiVLM (Wang et al., 2025). It offers a unified framework for benchmarking training-free acceleration methods but relies on existing datasets (e.g., DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022)) rather than data tailored to token compression. Building on data-driven insights into compression behavior, we introduce *VTC-Bench*, the first dedicated, challenging evaluation framework as an additional tool for evaluating visual token compression in MLLMs. We aim for VTC-Bench to catalyze new research and insights, enabling fair comparisons and sharper evaluations of token-compression methods.

3 Experiments & Findings

3.1 Motivation

Some recent MLLMs, such as Qwen2-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025), natively support inputs of varying resolutions. A trivial yet efficient method to handle high-resolution images is to simply downsample them to a lower resolution, effectively using naive pixel sampling as a form of compression. However, as shown in

Section 2.1, most token compression methods for MLLMs choose to adaptively drop useless tokens or merge similar tokens instead of directly downsampling the original image, which should be more intelligent and reasonable methods. While in recent works (Yang et al., 2025b), it is surprising that *image downsampling exceeds other sophisticated methods under some settings*. VisionThink (Yang et al., 2025a) considers this finding as a motivation to design more efficient MLLMs. Different from VisionThink (Yang et al., 2025a), we would like to further investigate the causes of this anomalous phenomenon, thus deciding to comprehensively compare the results of the downsampling methods with other methods under various settings.

3.2 Experiments Setup

Before conducting experiments, it is crucial to choose a suitable MLLM for achieving the downsampling method. Most MLLMs only support fixed-resolution inputs, which makes it impossible to achieve the downsampling method. In other words, for such MLLMs, no matter which resolution the original image is downsampled to, the image will finally be resized to a fixed resolution, making the downsampling meaningless. Considering that Qwen2-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025), based on the naive dynamic resolution mechanism and M-RoPE techniques, are the open-source MLLMs closest to realizing the concept of allowing arbitrary resolution inputs, we choose Qwen2-VL³ in our comparison experiments, which supports the downsampling method the best. In order to ensure that downsampling occurs at the original resolution as much as possible without adding extra resizing operations, we set Qwen2-VL’s max pixels and min pixels to 2408448 and 3136. In this case, only a few extremely high-resolution images are resized before downsampling to ensure sufficient GPU memory.

To guarantee comprehensive experiments, we select four typical token compression methods: FastV (Chen et al., 2024b), VisionZip (Yang et al., 2025a), PruMerge+ (Shang et al., 2024), and DART (Wen et al., 2025a), with the token compression ratio set to: 75.00%, 88.89%, 93.75%, 96.00%, and 99.00%. For the token compression ra-

³When conducting this experiment, on one hand, there is no repository that systematically implements visual token compression method based on Qwen2.5-VL; on the other hand, there are few open-source visual token compression methods adapted for Qwen2.5-VL. Therefore, we use Qwen2-VL as the main experimental model.

tio C , the downsampling method applies an equivalent downsampling ratio D for fairness. The rule is shown in Eq. 1. Moreover, we choose eight popular benchmarks, including six general benchmarks (GQA (Hudson and Manning, 2019), MM-Bench_EN (Liu et al., 2024b), MMBench_CN (Liu et al., 2024b), MME (Yin et al., 2024), POPE (Li et al., 2023), and MMStar (Chen et al., 2024c)) and two resolution-sensitive OCR benchmarks (OCR-Bench (Liu et al., 2024c), and ChartQA (Masry et al., 2022)).

$$\frac{1}{D^2} \times 100\% = 1 - C \quad (1)$$

3.3 Results Analysis

Comparison between different methods: Across diverse compression ratios and general-purpose benchmarks, simple image downsampling consistently outperforms sophisticated token compression methods. For instance, at 93.75% compression, downsampling achieves 66.4% on MMBench, outperforming the best advanced method, DART, by a 24.3% relative improvement. Meanwhile, with a constraint of 70% of vanilla performance, downsampling can achieve a compression ratio of 93.75%, while other methods can only meet this condition at a compression ratio of 88.89%. The results verify a basic phenomenon in the field of visual token compression: *a substantial portion of samples in general-purpose benchmarks can be correctly answered using only low-resolution global information, without requiring the fine-grained visual details that advanced methods strive to preserve.*

Comparison between different compression ratios: As compression becomes increasingly aggressive (96.00% and 99.00%), all sophisticated token compression methods experience performance degradation, while image downsampling demonstrates remarkably graceful degradation. At 99.00% compression, downsampling maintains a score of 51.6% on MMBench, while FastV and PruMerge+ decrease to approximately 24%. The results further verify the phenomenon above: *in the existing general-purpose benchmarks, image downsampling can fully meet the acceleration requirements for most samples.*

Comparison between different tasks: On tasks requiring fine-grained visual understanding, particularly chart comprehension, we observe a reversal of the phenomenon mentioned above. At moder-

Table 2: Comparison of advanced token compression methods on Qwen2-VL-7B. Values are formatted as: **Group A** (Group B) with difference Δ below. Δ refers to *the absolute gap between groups*.

Method	GQA	MMB	MMB ^{CN}	MME	POPE	MMStar	OCRBench	ChartQA	Average
Group A (Group B)	Token Reduction (\downarrow 75.00%)								
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	57.8 (87.6) Δ 29.8	45.2 (95.9) Δ 50.7	56.5 (95.8) Δ 39.3	78.9 (96.7) Δ 17.8	65.4 (94.8) Δ 29.4	41.0 (76.0) Δ 35.0	29.1 (57.2) Δ 28.1	35.0 (78.1) Δ 43.1	51.1 (85.3) Δ 34.2
+ VisionZip (Yang et al., 2025a)	59.3 (91.2) Δ 31.9	42.4 (93.8) Δ 51.4	42.2 (93.6) Δ 51.4	54.9 (95.3) Δ 40.4	72.5 (96.8) Δ 24.3	45.9 (81.4) Δ 35.5	29.6 (58.1) Δ 28.5	51.2 (87.3) Δ 36.1	49.8 (87.2) Δ 37.4
+ PruMerge+ (Shang et al., 2024)	57.7 (91.9) Δ 34.2	51.2 (95.1) Δ 43.9	52.6 (94.6) Δ 42.0	62.0 (95.9) Δ 33.9	72.1 (97.5) Δ 25.4	48.1 (82.3) Δ 34.2	21.2 (46.2) Δ 25.0	40.5 (73.6) Δ 33.1	50.7 (84.6) Δ 33.9
+ DART (Wen et al., 2025a)	58.9 (88.1) Δ 29.2	54.8 (94.9) Δ 40.1	52.2 (94.6) Δ 42.4	67.6 (94.9) Δ 27.3	69.4 (94.5) Δ 25.1	47.0 (77.7) Δ 30.7	40.2 (70.2) Δ 30.0	39.0 (69.0) Δ 30.0	53.6 (85.5) Δ 31.9
Group A (Group B)	Token Reduction (\downarrow 88.89%)								
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	44.5 (82.5) Δ 38.0	39.2 (90.3) Δ 51.1	44.1 (90.8) Δ 46.7	59.4 (94.0) Δ 34.6	46.8 (88.7) Δ 41.9	31.0 (73.0) Δ 42.0	17.8 (41.3) Δ 23.5	28.4 (61.7) Δ 33.3	38.9 (77.8) Δ 38.9
+ VisionZip (Yang et al., 2025a)	49.4 (83.4) Δ 34.0	33.2 (89.0) Δ 55.8	44.4 (88.1) Δ 43.7	48.1 (92.2) Δ 44.1	70.0 (92.3) Δ 22.3	30.3 (73.0) Δ 42.7	22.0 (36.4) Δ 14.4	49.7 (74.4) Δ 24.7	43.4 (78.6) Δ 35.2
+ PruMerge+ (Shang et al., 2024)	50.4 (85.8) Δ 35.4	36.9 (87.2) Δ 50.3	38.4 (86.4) Δ 48.0	42.9 (91.9) Δ 49.0	71.5 (94.2) Δ 22.7	28.8 (71.6) Δ 42.8	18.1 (33.0) Δ 14.9	43.5 (73.8) Δ 30.3	41.3 (78.0) Δ 36.7
+ DART (Wen et al., 2025a)	47.5 (81.2) Δ 33.7	40.5 (87.7) Δ 47.2	40.9 (86.9) Δ 46.0	49.6 (91.7) Δ 42.1	57.7 (90.9) Δ 33.2	35.4 (70.0) Δ 34.6	31.5 (63.2) Δ 31.7	27.3 (57.6) Δ 30.3	41.3 (78.6) Δ 37.3
Group A (Group B)	Token Reduction (\downarrow 93.75%)								
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	35.7 (81.4) Δ 45.7	31.9 (85.7) Δ 53.8	35.3 (86.6) Δ 51.3	48.8 (91.5) Δ 42.7	37.4 (88.1) Δ 50.7	22.8 (74.5) Δ 51.7	13.3 (33.0) Δ 19.7	14.8 (74.8) Δ 60.0	30.0 (77.0) Δ 47.0
+ VisionZip (Yang et al., 2025a)	41.0 (79.0) Δ 38.0	34.5 (81.9) Δ 47.4	33.3 (82.2) Δ 48.9	43.5 (88.4) Δ 44.9	66.3 (89.4) Δ 23.1	24.3 (69.8) Δ 45.5	14.0 (25.2) Δ 11.2	26.1 (71.3) Δ 45.2	35.4 (73.4) Δ 38.0
+ PruMerge+ (Shang et al., 2024)	43.0 (76.7) Δ 33.7	29.6 (76.9) Δ 47.3	34.1 (76.1) Δ 42.0	43.0 (87.8) Δ 44.8	67.7 (87.6) Δ 19.9	25.5 (65.5) Δ 40.0	12.6 (21.8) Δ 9.2	29.4 (68.9) Δ 39.5	35.6 (70.2) Δ 34.6
+ DART (Wen et al., 2025a)	41.9 (78.8) Δ 36.9	33.8 (81.8) Δ 48.0	38.4 (80.4) Δ 42.0	46.9 (88.9) Δ 42.0	57.0 (88.5) Δ 31.5	26.2 (61.8) Δ 35.6	25.6 (57.4) Δ 31.8	14.5 (67.1) Δ 52.6	35.5 (75.6) Δ 40.1

ate compression ratios (93.75% and 88.89%), VisionZip and FastV outperform image downsampling on ChartQA by significant margins. This divergence is highly informative: while image downsampling uniformly preserves global information at the expense of local details, the sophisticated compression methods can selectively retain text regions and numeric values that are critical for chart understanding, which can be considered difficult to compress. Thus, a deeper observation of the above phenomenon can be concluded: *the sophisticated token compression methods demonstrate the expected effectiveness in tasks that require fine-grained visual understanding.*

The comparisons across methods, compression ratios, and tasks provide compelling evidence that current benchmarks contain a substantial simplicity bias. The performance advantage of image downsampling emerges not from its sophistication but from its ability to adequately address samples

that don't require fine-grained visual understanding, precisely the samples that dominate current benchmarks. Thus, based on the experimental results and the comparisons, we propose a well-founded hypothesis in Section 3.4.

3.4 Hypothesis

In the field of visual token compression, there is a general reliance on existing benchmarks, without ever considering whether these data are suitable for the visual token compression task. Based on the above experimental results in Table 1, we propose a bold hypothesis: *Some data in the existing benchmarks is overly simplistic and irrelevant to evaluating visual token compression methods, leading to the unreasonable phenomenon that even the downsampling method is sufficient to deal with the visual token compression task.*

To validate this hypothesis, we design a data-centric analysis using downsampling as a discrimi-

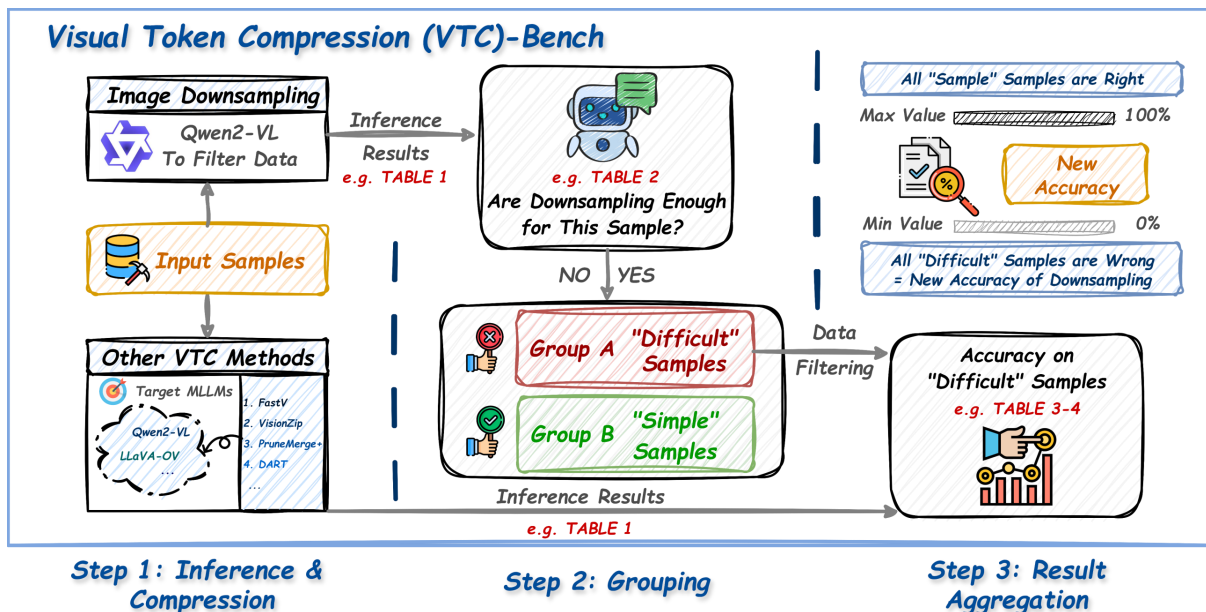


Figure 2: The VTC-Bench is a simple but effective framework that can transform any existing benchmarks to a subset that can fairly evaluate VTC (Visual Token Compression) methods. The samples that are answered correctly by the original Qwen2-VL model without downsampling form the input samples. More details in Section 4.1.

nator. We first drop out the samples answered incorrectly at the original resolution, which we consider are too hard for the original models to understand, not to mention the compressed models. Then, for a given compression ratio, we classify each sample in a benchmark into one of two groups based on the performance of the downsampling method: ① Difficult Samples (Group A): Samples that are answered incorrectly by the downsampling method. ② Simple Samples (Group B): Samples that are answered correctly by the downsampling method. We then evaluate all compression methods on these two groups separately to assess whether the sophisticated methods demonstrate their expected superiority on the “difficult” samples where image downsampling fails. Results from Table 2 strongly confirm our hypothesis, followed by two key conclusions as follows.

① **Significant performance gap between groups:** Across all benchmarks and compression methods, the accuracy on “simple” samples (Group B) is dramatically higher than on “difficult” samples (Group A). For instance, on GQA at 75% compression, the accuracy of all methods on simple samples is above 87.6%, while on difficult samples, it drops to a maximum of 59.3% (VisionZip). This stark contrast is common in Table 2, demonstrating that the two groups represent essentially different levels of visual comprehension difficulty. The existence of this gap validates our core hypothesis that

the benchmark comprises a mixture of “simple” and “difficult” samples. In other words, the current benchmarks are noisy for evaluating the visual token compression methods, where the noise refers to the task-irrelevant samples. Moreover, the significant gap also proves that *downsampling can serve as a clever filter to distinguish between “simple” and “difficult” samples, which can be the key to denoise the current benchmarks.*

② **Ideal reference points brought by downsampling:** The 0%/100% dichotomy created by image downsampling provides ideal reference points for evaluation. In Group B, where downsampling achieves 100% accuracy, advanced methods show comparable but not superior performance (e.g., 87.6-91.9% on GQA at 75% compression), confirming that their sophisticated approaches offer no advantage for simple samples. In Group A, where downsampling fails completely (0% accuracy), advanced methods demonstrate their true value by significantly exceeding this baseline. For instance, DART achieves 40.2% on OCRBench and VisionZip reaches 51.2% on ChartQA at 75% compression, proving their ability to preserve crucial details that downsampling loses.

3.5 Summary

In this section, we conduct two comprehensive experiments to further understand the anomalous phenomenon: image downsampling exceeds other

Table 3: VTC-Bench results on Qwen2-VL-7B. Take Qwen2-VL-7B as the data filter model.

Method	GQA	MMB	MMB ^{CN}	MME	POPE	MMStar	OCRBench	ChartQA	Average
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 75.00%)									
+ Downsample (Baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+ FastV (Chen et al., 2024b)	57.8	45.2	56.5	78.9	65.4	41.0	29.1	35.0	<u>51.1</u>
+ VisionZip (Yang et al., 2025a)	59.3	42.4	42.2	54.9	72.5	45.9	29.6	51.2	49.8
+ PruMerge+ (Shang et al., 2024)	57.7	51.2	52.6	62.0	<u>72.1</u>	48.1	21.2	40.5	50.7
+ DART (Wen et al., 2025a)	<u>58.9</u>	54.8	52.2	67.6	69.4	<u>47.0</u>	40.2	39.0	53.6
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 88.89%)									
+ Downsample (Baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+ FastV (Chen et al., 2024b)	44.5	39.2	44.1	59.4	46.8	31.0	17.8	28.4	38.9
+ VisionZip (Yang et al., 2025a)	49.4	33.2	44.4	48.1	70.0	30.3	22.0	49.7	43.4
+ PruMerge+ (Shang et al., 2024)	50.4	36.9	38.4	42.9	71.5	28.8	18.1	<u>43.5</u>	41.3
+ DART (Wen et al., 2025a)	47.5	40.5	40.9	<u>49.6</u>	57.7	35.4	31.5	27.3	<u>41.3</u>
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 93.75%)									
+ Downsample (Baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+ FastV (Chen et al., 2024b)	35.7	31.9	<u>35.3</u>	48.8	37.4	22.8	13.3	14.8	30.0
+ VisionZip (Yang et al., 2025a)	41.0	34.5	33.3	43.5	<u>66.3</u>	24.3	14.0	26.1	35.4
+ PruMerge+ (Shang et al., 2024)	43.0	29.6	34.1	43.0	67.7	25.5	12.6	29.4	35.6
+ DART (Wen et al., 2025a)	<u>41.9</u>	<u>33.8</u>	38.4	<u>46.9</u>	57.0	26.2	25.6	14.5	<u>35.5</u>
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 96.00%)									
+ Downsample (Baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+ FastV (Chen et al., 2024b)	29.6	24.7	<u>33.1</u>	35.6	35.6	21.5	<u>11.0</u>	9.3	25.0
+ VisionZip (Yang et al., 2025a)	38.6	<u>31.2</u>	32.6	37.9	60.0	24.5	<u>11.0</u>	<u>15.1</u>	31.4
+ PruMerge+ (Shang et al., 2024)	38.8	26.0	29.3	37.0	<u>56.4</u>	22.6	9.2	17.0	29.5
+ DART (Wen et al., 2025a)	36.4	33.7	36.4	37.9	53.2	22.3	24.7	10.5	31.9
<i>Qwen2-VL-7B (Wang et al., 2024)</i>									
Token Reduction (↓ 99.00%)									
+ Downsample (Baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+ FastV (Chen et al., 2024b)	18.3	18.3	21.5	21.5	44.3	15.0	4.2	3.8	18.4
+ VisionZip (Yang et al., 2025a)	23.4	28.8	32.2	28.5	53.6	19.4	3.7	5.5	24.4
+ PruMerge+ (Shang et al., 2024)	20.7	17.8	21.1	22.9	<u>52.6</u>	17.1	2.5	7.1	20.2
+ DART (Wen et al., 2025a)	24.5	<u>26.5</u>	<u>28.1</u>	30.6	41.5	<u>19.2</u>	25.6	4.2	25.0

Table 4: VTC-Bench results on LLaVA-OV-7B. Take Qwen2-VL-7B as the data filter model.

Method	GQA	MMB	MMB ^{CN}	POPE	MMStar
<i>LLaVA-OV-7B (Li et al., 2024a)</i>					
Token Reduction (↓ 75.00%)					
+ FastV (Chen et al., 2024b)	54.3	70.5	69.1	63.8	48.6
+ VisionZip (Yang et al., 2025a)	59.0	67.7	71.3	80.8	44.8
+ PruMerge+ (Shang et al., 2024)	60.4	74.2	73.5	<u>75.6</u>	48.6
<i>LLaVA-OV-7B (Li et al., 2024a)</i>					
Token Reduction (↓ 88.89%)					
+ FastV (Chen et al., 2024b)	45.3	64.6	66.4	39.1	42.4
+ VisionZip (Yang et al., 2025a)	56.6	71.9	71.2	69.6	43.5
+ PruMerge+ (Shang et al., 2024)	57.4	68.8	71.5	76.0	45.8
<i>LLaVA-OV-7B (Li et al., 2024a)</i>					
Token Reduction (↓ 93.75%)					
+ FastV (Chen et al., 2024b)	36.7	51.2	53.3	29.7	32.6
+ VisionZip (Yang et al., 2025a)	49.1	64.3	62.4	53.6	36.6
+ PruMerge+ (Shang et al., 2024)	50.2	66.6	65.3	59.9	<u>34.8</u>
<i>LLaVA-OV-7B (Li et al., 2024a)</i>					
Token Reduction (↓ 96.00%)					
+ FastV (Chen et al., 2024b)	31.4	37.4	43.1	24.5	28.6
+ VisionZip (Yang et al., 2025a)	42.6	55.4	56.9	45.4	30.8
+ PruMerge+ (Shang et al., 2024)	42.7	57.8	59.6	49.9	31.3
<i>LLaVA-OV-7B (Li et al., 2024a)</i>					
Token Reduction (↓ 99.00%)					
+ FastV (Chen et al., 2024b)	25.7	25.8	29.6	39.3	21.9
+ VisionZip (Yang et al., 2025a)	28.3	28.1	32.8	42.1	24.7
+ PruMerge+ (Shang et al., 2024)	25.3	25.5	28.5	40.4	25.2

sophisticated methods under some settings. The first experiment validates the universality of this anomalous phenomenon and introduces our basic hypothesis: Some data in the existing benchmarks is overly simplistic, leading to the unreasonable phenomenon that even the simplest downsampling method is sufficient to deal with the visual token compression task. Furthermore, the second experiment further validates this hypothesis and proves that the current benchmarks are noisy for evaluat-

ing the visual token compression methods. Moreover, the second experiment simultaneously demonstrates that downsampling can serve as a clever filter to distinguish between “simple” and “difficult” samples, which can be the key to denoise the current benchmarks.

4 Evaluation Framework

4.1 Framework Construction

To address the simplicity bias and denoise existing benchmarks for the visual token compression task, we propose the VTC-Bench (Visual Token Compression Benchmark) framework, a novel framework specifically designed for the fair and effective evaluation of visual token compression methods. The construction is based on the key insight, validated in Section 3.4, that “downsampling can serve as a clever filter to distinguish between ‘simple’ and ‘difficult’ samples”. We leverage this idea to construct a challenging benchmark comprising predominantly “difficult” samples that require fine-grained visual understanding. This process, summarized in Figure 2, does not create new data but rather applies a rigorous filtering mechanism to existing benchmarks to identify a challenging evaluation and noise-free set. The pipeline consists of three critical steps executed for each candidate

sample and dynamically adapts to different compression ratios:

Step 1: Inference & Compression: Given a sample and a target token compression ratio, we run two inference pipelines: ① a downsampling baseline (the filter) including one model that applies the equivalent ratio from Eq. 1 for a fair comparison and another original model without downsampling, implemented with Qwen2-VL which has a similar number of parameter with the target MLLM; and ② advanced visual token compression methods (e.g., FastV, VisionZip, DART) evaluated directly on the target MLLM. This step both establishes a fair basis for assessing compression approaches and provides signals for subsequent sample filtering.

Step 2: Grouping: We first drop out the samples that are incorrectly answered by the original Qwen2-VL. Then, we use the performance of the downsampling method as a binary discriminator to categorize the sample into two groups: ① Group A: Samples considered as “difficult”, which are incorrectly answered by the downsampling method. ② Group B: Samples considered as “simple”, which are correctly answered by the downsampling method. This step effectively tags each sample with the labels of “simple” or “difficult”, filtering the existing benchmarks and removing noisy data that is not applicable for evaluating the visual token compression methods.

Step 3: Result Aggregation: Based on the classification in Step 2 and the inference results of visual token compression methods in Step 1, we perform a statistical analysis on the accuracy of the “difficult” samples in the methods to be evaluated. Thus, an indicator that can truly reflect the visual compression methods fairly can be obtained.

In summary, we develop VTC-Bench, a simple but effective framework for evaluating visual token compression methods. The building pipeline of VTC-Bench is shown in Figure 2. Importantly, the VTC-Bench framework can be applied easily to any existing benchmark, transforming it into a more effective benchmark for evaluating visual token compression methods. Meanwhile, the VTC-Bench framework dynamically and reasonably provides a corresponding benchmark subset for each compression ratio, while offering explainable theoretical upper and lower bounds for the final metrics.

4.2 Evaluation Results & Discussions

We conduct extensive experiments across multiple mainstream MLLMs and benchmarks based on VTC-Bench. We select Qwen2-VL-7B (Wang et al., 2024) and LLaVA-OV-7B (Li et al., 2024a) as the base MLLMs and evaluate various visual token compression methods (including FastV, VisionZip, PruMerge+, DART) on a subset of “difficult” samples filtered by VTC-Bench. The experimental results are shown in Table 3 and 4, followed by several analysis:

Is downsampling all you need? Across many benchmarks, simple image downsampling often beats more advanced compression methods, suggesting that sophisticated approaches are unnecessary. VTC-Bench overturns this impression: when we restrict evaluation to the compression-relevant “difficult” samples (Group A), the trend reverses. The apparent superiority of downsampling largely stems from original benchmarks being saturated with easy cases that do not require fine-grained cues. By filtering out such samples, VTC-Bench reveals that for truly challenging instances that test visual understanding, advanced compression methods are not only effective but necessary.

What makes an effective benchmark? Simple cross-benchmark comparisons (e.g., “Benchmark A outperforms Benchmark B”) only imply that one is harder, without revealing which skills drive the difficulty or whether it is relevant to visual token compression. VTC-Bench addresses this by filtering out samples that do not inform compression performance, yielding an analysis set that is explicitly sensitive to token compression. This suggests a design principle for future work: effective benchmarks for visual token compression should deliberately increase the share of compression-relevant hard cases.

Further expand the accuracy gap: VTC-Bench amplifies and clarifies method differences. At 75% compression on ChartQA, the VisionZip–FastV gap widens from 8.8% to 16.2%; at 96% compression on GQA, it grows from 0.3% to 9.0%. These phenomenon effectively indicates that VTC-Bench indeed eliminates data noise unrelated to the visual token compression task, thereby promoting the fairness and effectiveness of the benchmark in the visual token compression task.

5 Conclusion

This paper systematically analyzes the task mismatch problem presented in current MLLM benchmarks when evaluating visual token compression methods. Based on a surprising and counterintuitive finding: simple image downsampling consistently outperforms many advanced compression methods across multiple widely used benchmarks, we conduct a comprehensive empirical study across several advanced visual token compression methods. Thus, two crucial findings are concluded based on the empirical study: ① Current benchmarks are noisy for the visual token compression task. ② Downsampling can serve as a data filter to evaluate the difficulty of samples upon the visual token compression task. Furthermore, we propose VTC-Bench, a new evaluation framework specifically designed to optimize and denoise current existing benchmarks by explicitly distinguishing between “simple” and “difficult” samples through downsampling. Through this work, we hope to not only advance the field of visual token compression but also inspire more discussions within the community on “how to properly evaluate efficient MLLMs.”

6 Limitations

Dependence on a single base model: The construction of the sample filtering mechanism is entirely dependent on Qwen2-VL. While chosen for its technical necessity, this reliance limits generalizability, as different models may define “difficult” samples slightly differently. In the future, this framework will be extended to more base models that have great dynamic resolution support.

Single criterion for filtering: We use only downsampling as the criterion for identifying “difficult” samples. Although effective, employing a more diverse set of baseline methods for ensemble filtering could yield a more robust definition. We will explore a more robust filtering mechanism for the subsequent version of VTC-Bench in future work.

Lack of a formal theoretical definition: Our identification of “difficult” and “simple” samples is primarily based on the experiment results in Table 1 and 2. A formal theoretical framework and mathematical proof for a more explainable sample filtering mechanism is yet to be established.

Limited coverage of methods: Due to the rapidly evolving field and adaptation constraints, our experimental comparison doesn’t include all visual token compression methods. However, the selected

methods are sufficient to support our core claims, and the framework remains open and extensible.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality; the Shanghai Science and Technology Program (Grant No. 25ZR1402278).

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*.
- Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. 2024a. Efficient large multi-modal models via visual context compression. *Advances in Neural Information Processing Systems*, 37:73986–74007.
- Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. 2025a. Variation-aware vision token dropping for faster large vision-language models. *arXiv preprint arXiv:2509.01552*.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024b. 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.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024c. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

- Yuan Chen, Zichen Wen, Yuzhou Wu, Xuyang Liu, Shuang Chen, Junpeng Ma, Weijia Li, Conghui He, and Linfeng Zhang. 2025b. Ipcv: Information-preserving compression for mllm visual encoders. *arXiv preprint arXiv:2512.18747*.
- Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024d. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.
- Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2025. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22826–22835.
- Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. 2025. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6530–6540.
- Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. 2026. Filter, correlate, compress: Training-free token reduction for mllm acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4601–4609.
- Drew A Hudson and Christopher D Manning. 2019. 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.
- Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4075–4083.
- Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. 2025. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Duo Li, Zuhao Yang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2025. Todre: Effective visual token pruning via token diversity and task relevance. *arXiv preprint arXiv:2505.18757*.
- Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024b. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. 2024a. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*.
- Xuyang Liu, Zichen Wen, Shaobo Wang, Junjie Chen, Zhishan Tao, Yubo Wang, Tailai Chen, Xiangqi Jin, Chang Zou, Yiyu Wang, Chenfei Liao, Xu Zheng, Honggang Chen, Weijia Li, Xuming Hu, Conghui He, and Linfeng Zhang. 2025. Shifting ai efficiency from model-centric to data-centric compression. *arXiv preprint arXiv:2505.19147*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocr-bench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer.
- Ruiguang Pei, Weiqing Sun, Zhihui Fu, and Jun Wang. 2025. Greedyprune: Retenting critical visual token set for large vision language models. *arXiv preprint arXiv:2506.13166*.
- Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of*

- the AAAI Conference on Artificial Intelligence, volume 38, pages 4542–4550.
- Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.
- Kele Shao, TAO Keda, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. 2025. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*.
- Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.
- Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. 2025. Video-mmlu: A massive multi-discipline lecture understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6099–6113.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Zekun Wang, Minghua Ma, Zexin Wang, Rongchuan Mu, Liping Shan, Ming Liu, and Bing Qin. 2025. Effivlm-bench: A comprehensive benchmark for evaluating training-free acceleration in large vision-language models. *arXiv preprint arXiv:2506.00479*.
- Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025a. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*.
- Zichen Wen, Shaobo Wang, Yufa Zhou, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, and Linfeng Zhang. 2025b. Efficient multi-modal large language models via progressive consistency distillation. *arXiv preprint arXiv:2510.00515*.
- Zichen Wen, Boxue Yang, Shuang Chen, Yaojie Zhang, Yuhang Han, Junlong Ke, Cong Wang, Yicheng Fu, Jiawang Zhao, Jiangchao Yao, Xi Fang, Zhen Wang, Henxing Cai, Lin Yao, Zhifeng Gao, Yanhui Hong, Nang Yuan, Yixuan Li, Guojiang Zhao, and 15 others. 2026. Innovator-vl: A multimodal large language model for scientific discovery. *arXiv preprint arXiv:2601.19325*.
- Hao Wu, Junlong Tong, Xudong Wang, Yang Tan, Changyu Zeng, Anastasia Antsiferova, and Xiaoyu Shen. 2026. From data to model: A survey of the compression lifecycle in mllms.
- Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. Pyramidrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.
- Minhao Xiong, Zichen Wen, Zhuangcheng Gu, Xuyang Liu, Rui Zhang, Hengrui Kang, Jiabing Yang, Junyuan Zhang, Weijia Li, Conghui He, Yafei Wang, and Linfeng Zhang. 2025. Prune2drive: A plug-and-play framework for accelerating vision-language models in autonomous driving. *arXiv preprint arXiv:2508.13305*.
- Bingxin Xu, Yuzhang Shang, Yunhao Ge, Qian Lou, and Yan Yan. 2024. freepruner: A training-free approach for large multimodal model acceleration. *arXiv preprint arXiv:2411.15446*.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025a. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802.
- Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025b. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*.
- Sihan Yang, Runsen Xu, Chenhang Cui, Tai Wang, Dahua Lin, and Jiangmiao Pang. 2025c. Vflowopt: A token pruning framework for llms with visual information flow-guided optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23924–23934.
- Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, Qidong Huang, Haotian Wang, Feng Wu, Yuanxing Zhang, Pengfei Wan, Zhouchen Lin, and Xu Sun. 2026. Towards efficient multimodal large language models: A survey on token compression.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024.

Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025a. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning*.

Zixin Zhang, Kanghao Chen, Xingwang Lin, Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Litao Guo, Yinchuan Li, and Ying-Cong Chen. 2025b. Phystool-bench: Benchmarking physical tool understanding for mllms. *arXiv preprint arXiv:2510.09507*.

Zixin Zhang, Kanghao Chen, Hanqing Wang, Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Litao Guo, and Ying-Cong Chen. 2025c. A4-agent: An agentic framework for zero-shot affordance reasoning. *arXiv preprint arXiv:2512.14442*.

Zixin Zhang, Chenfei Liao, Hongfei Zhang, Harold Haodong Chen, Kanghao Chen, Zichen Wen, Litao Guo, Bin Ren, Xu Zheng, Yinchuan Li, Xuming Hu, Nicu Sebe, and Ying-Cong Chen. 2026. Panoramic affordance prediction. *arXiv preprint arXiv:2603.15558*.

Dehua Zheng, Mouxiao Huang, Borui Jiang, Hailin Hu, and Xinghao Chen. 2025. Towards lossless ultimate vision token compression for vlms. *arXiv preprint arXiv:2512.09010*.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2025a. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *arXiv preprint arXiv:2510.02912*.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Ken-ting Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. 2025b. Look twice before you answer: Memory-space visual re-tracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning*.

A Experiment Details

In this paper, all the experiments are conducted based on one A800 GPU. For the downsampling method and DART, we apply the official code of DART (Wen et al., 2025a). As to the downsampling method, we resize the image before it enters the MLLM. As to DART, we control the compression ratio through the parameter "Reduction_Ratio". For VisionZip, PruMerge+, and FastV, we apply the EffiVLM-Bench (Wang et al., 2025), which offers a unified toolkit to evaluate efficient MLLM. As to these three methods, we control the compression ratio through the parameter "Budget". Considering this paper focuses on the evaluation, it is not related to hyperparameter search. All results come from a single run. The code environment includes Python=3.10, torch=2.6.0, torchvision=0.21.0, and torchaudio=2.6.0. Bicubic interpolation is used to achieve downsampling. We will release all the results, including the output results of each sample and the accuracy results of each benchmark.

B Benchmark Details

GQA: GQA (Hudson and Manning, 2019) is a large-scale benchmark for visual reasoning and compositional question answering. Based on a strict distribution control, GQA offers 22M valuable reasoning questions.

MMBench: MMBench (Liu et al., 2024b) is a comprehensive benchmark designed to evaluate the capabilities of MLLMs. It includes 3217 multiple-choice questions spanning 20 fine-grained dimensions, supporting several languages such as Chinese and English.

MME: MME (Yin et al., 2024) provides a systematic framework for evaluating the perceptual and cognitive abilities of MLLMs. It encompasses 14 subtasks across the domains of visual perception, text understanding, reasoning, and cross-modal alignment.

POPE: POPE (Li et al., 2023) is a benchmark designed to evaluate object hallucination in MLLMs. The pipeline of POPE measures hallucination under random, popular, and adversarial sampling strategies.

MMStar: MMStar (Chen et al., 2024c) is a vision-dependent benchmark for evaluating the reasoning and perception abilities. It has 1500 samples, covering six core abilities with 18 sub-dimensions.

OCRBench: OCRBench (Liu et al., 2024c)

is a comprehensive benchmark for evaluating the OCR capabilities of multimodal large models. The benchmark includes 1000 manually verified samples from 29 datasets.

ChartQA: ChartQA (Masry et al., 2022) evaluates visual and logical reasoning over real-world charts. It includes 9.6k human-written and 23.1k automatically generated questions across different kinds of charts.

C Complete VTC-Bench Results

Due to the page limitation, we are unable to offer complete results in the experiment sections. Thus, we provide the evaluation results by group of Qwen2-VL-7B and LLaVA-OV-7B on eight benchmarks here, as shown in Table 5 and 6.

D Concurrent Similar Research

D.1 VisionThink

VisionThink (Yang et al., 2025b) identifies a key observation: downsampling shows surprising effectiveness on most general VQA tasks except for OCR and detail-sensitive benchmarks. Leveraging this finding, VisionThink is proposed as a novel paradigm that builds an effective and smart LVLM. Instead of applying a fixed compression rate, VisionThink starts inference with a low-resolution image and employs a reinforcement learning framework to enable the model itself to dynamically decide whether the visual information is sufficient or if a high-resolution original image is needed. This approach effectively utilizes the observation by optimizing for efficiency on simple samples while preserving necessary detail for complex, OCR-heavy tasks.

Different from VisionThink’s perspective, we further investigate this observation in comprehensive settings and analyze it through the benchmark perspective. Through extensive experimental results in Table 1 and 2, firstly, the task mismatch between current benchmarks and the visual token compression task is clarified. Secondly, the potential of downsampling as a data filter is proven, motivating the subsequent design of VTC-Bench.

D.2 EffiVLM-Bench

EffiVLM-Bench (Wang et al., 2025) is a comprehensive evaluation framework for systematically assessing training-free acceleration techniques in LVLMs, including KV cache compression and token prune methods. Thanks to the open-source of

Table 5: Comparison of advanced token compression methods on Qwen2-VL-7B across five reduction ratios. Values are formatted as: **Group A** (Group B) with difference Δ below. Δ refers to *the absolute gap between groups*.

Method	GQA	MMB	MMB ^{CN}	MME	POPE	MMStar	OCRBench	ChartQA	Average
<i>Token Reduction (\downarrow 75.00%)</i>									
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	57.8 (87.6) Δ 29.8	45.2 (95.9) Δ 50.7	56.5 (95.8) Δ 39.3	78.9 (96.7) Δ 17.8	65.4 (94.8) Δ 29.4	41.0 (76.0) Δ 35.0	29.1 (57.2) Δ 28.1	35.0 (78.1) Δ 43.1	51.1 (85.3) Δ 34.2
+ VisionZip (Yang et al., 2025a)	59.3 (91.2) Δ 31.9	42.4 (93.8) Δ 51.4	42.2 (93.6) Δ 51.4	54.9 (95.3) Δ 40.4	72.5 (96.8) Δ 24.3	45.9 (81.4) Δ 35.5	29.6 (58.1) Δ 28.5	51.2 (87.3) Δ 36.1	49.8 (87.2) Δ 37.4
+ PruMerge+ (Shang et al., 2024)	57.7 (91.9) Δ 34.2	51.2 (95.1) Δ 43.9	52.6 (94.6) Δ 42.0	62.0 (95.9) Δ 33.9	72.1 (97.5) Δ 25.4	48.1 (82.3) Δ 34.2	21.2 (46.2) Δ 25.0	40.5 (73.6) Δ 33.1	50.7 (84.6) Δ 33.9
+ DART (Wen et al., 2025a)	58.9 (88.1) Δ 29.2	54.8 (94.9) Δ 40.1	52.2 (94.6) Δ 42.4	67.6 (94.9) Δ 27.3	69.4 (94.5) Δ 25.1	47.0 (77.7) Δ 30.7	40.2 (70.2) Δ 30.0	39.0 (69.0) Δ 30.0	53.6 (85.5) Δ 31.9
<i>Token Reduction (\downarrow 88.89%)</i>									
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	44.5 (82.5) Δ 38.0	39.2 (90.3) Δ 51.1	44.1 (90.8) Δ 46.7	59.4 (94.0) Δ 34.6	46.8 (88.7) Δ 41.9	31.0 (73.0) Δ 42.0	17.8 (41.3) Δ 23.5	28.4 (61.7) Δ 33.3	38.9 (77.8) Δ 38.9
+ VisionZip (Yang et al., 2025a)	49.4 (83.4) Δ 34.0	33.2 (89.0) Δ 55.8	44.4 (88.1) Δ 43.7	48.1 (92.2) Δ 44.1	70.0 (92.3) Δ 22.3	30.3 (73.0) Δ 42.7	22.0 (36.4) Δ 14.4	49.7 (74.4) Δ 24.7	43.4 (78.6) Δ 35.2
+ PruMerge+ (Shang et al., 2024)	50.4 (85.8) Δ 35.4	36.9 (87.2) Δ 50.3	38.4 (86.4) Δ 48.0	42.9 (91.9) Δ 49.0	71.5 (94.2) Δ 22.7	28.8 (71.6) Δ 42.8	18.1 (33.0) Δ 14.9	43.5 (73.8) Δ 30.3	41.3 (78.0) Δ 36.7
+ DART (Wen et al., 2025a)	47.5 (81.2) Δ 33.7	40.5 (87.7) Δ 47.2	40.9 (86.9) Δ 46.0	49.6 (91.7) Δ 42.1	57.7 (90.9) Δ 33.2	35.4 (70.0) Δ 34.6	31.5 (63.2) Δ 31.7	27.3 (57.6) Δ 30.3	41.3 (78.6) Δ 37.3
<i>Token Reduction (\downarrow 93.75%)</i>									
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	35.7 (81.4) Δ 45.7	31.9 (85.7) Δ 53.8	35.3 (86.6) Δ 51.3	48.8 (91.5) Δ 42.7	37.4 (88.1) Δ 50.7	22.8 (74.5) Δ 51.7	13.3 (33.0) Δ 19.7	14.8 (74.8) Δ 60.0	30.0 (77.0) Δ 47.0
+ VisionZip (Yang et al., 2025a)	41.0 (79.0) Δ 38.0	34.5 (81.9) Δ 47.4	33.3 (82.2) Δ 48.9	43.5 (88.4) Δ 44.9	66.3 (89.4) Δ 23.1	24.3 (69.8) Δ 45.5	14.0 (25.2) Δ 11.2	26.1 (71.3) Δ 45.2	35.4 (73.4) Δ 38.0
+ PruMerge+ (Shang et al., 2024)	43.0 (76.7) Δ 33.7	29.6 (76.9) Δ 47.3	34.1 (76.1) Δ 42.0	43.0 (87.8) Δ 44.8	67.7 (87.6) Δ 19.9	25.5 (65.5) Δ 40.0	12.6 (21.8) Δ 9.2	29.4 (68.9) Δ 39.5	35.6 (70.2) Δ 34.6
+ DART (Wen et al., 2025a)	41.9 (78.8) Δ 36.9	33.8 (81.8) Δ 48.0	38.4 (80.4) Δ 42.0	46.9 (88.9) Δ 42.0	57.0 (88.5) Δ 31.5	26.2 (61.8) Δ 35.6	25.6 (57.4) Δ 31.8	14.5 (67.1) Δ 52.6	35.5 (75.6) Δ 40.1
<i>Token Reduction (\downarrow 96.00%)</i>									
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	29.6 (79.2) Δ 49.6	24.7 (74.5) Δ 49.8	33.1 (74.4) Δ 41.3	35.6 (90.1) Δ 54.5	35.6 (85.8) Δ 50.2	21.5 (68.6) Δ 47.1	11.0 (27.5) Δ 16.5	9.3 (75.3) Δ 66.0	25.0 (71.9) Δ 46.9
+ VisionZip (Yang et al., 2025a)	38.6 (75.6) Δ 37.0	31.2 (80.3) Δ 49.1	32.6 (80.2) Δ 47.6	37.9 (87.4) Δ 49.5	60.0 (86.5) Δ 26.5	24.5 (69.6) Δ 45.1	11.0 (20.4) Δ 9.4	15.1 (69.2) Δ 54.1	31.4 (71.2) Δ 39.8
+ PruMerge+ (Shang et al., 2024)	38.8 (72.5) Δ 33.7	26.0 (69.3) Δ 43.3	29.3 (69.1) Δ 39.8	37.0 (83.9) Δ 46.9	56.4 (82.2) Δ 25.8	22.6 (61.4) Δ 38.8	9.2 (18.4) Δ 9.2	17.0 (70.3) Δ 53.3	29.5 (65.9) Δ 36.4
+ DART (Wen et al., 2025a)	36.4 (74.5) Δ 38.1	33.7 (77.3) Δ 43.6	36.4 (75.1) Δ 38.7	37.9 (84.9) Δ 47.0	53.2 (84.3) Δ 31.1	22.3 (63.5) Δ 41.2	24.7 (53.7) Δ 29.0	10.5 (71.1) Δ 60.6	31.9 (73.1) Δ 41.2
<i>Token Reduction (\downarrow 99.00%)</i>									
+ Downsample	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)	0.0 (100.0)
+ FastV (Chen et al., 2024b)	18.3 (75.5) Δ 57.2	18.3 (55.6) Δ 37.3	21.5 (53.2) Δ 31.7	21.5 (75.3) Δ 53.8	44.3 (59.3) Δ 15.0	15.0 (61.0) Δ 46.0	4.2 (19.5) Δ 15.3	3.8 (73.3) Δ 69.5	18.4 (59.1) Δ 40.7
+ VisionZip (Yang et al., 2025a)	23.4 (79.4) Δ 56.0	28.8 (78.3) Δ 49.5	32.2 (76.9) Δ 44.7	28.5 (80.5) Δ 52.0	53.6 (70.2) Δ 16.6	19.4 (69.4) Δ 50.0	3.7 (17.1) Δ 13.4	5.5 (70.3) Δ 64.8	24.4 (67.8) Δ 43.4
+ PruMerge+ (Shang et al., 2024)	20.7 (73.9) Δ 53.2	17.8 (55.9) Δ 38.1	21.1 (52.6) Δ 31.5	22.9 (73.9) Δ 51.0	52.6 (49.7) Δ 2.9	17.1 (57.0) Δ 39.9	2.5 (13.0) Δ 10.5	7.1 (69.5) Δ 62.4	20.2 (55.7) Δ 35.5
+ DART (Wen et al., 2025a)	24.5 (76.1) Δ 51.6	26.5 (63.2) Δ 36.7	28.1 (59.0) Δ 30.9	30.6 (73.2) Δ 42.6	41.5 (67.4) Δ 25.9	19.2 (63.7) Δ 44.5	25.6 (43.1) Δ 17.5	4.2 (70.7) Δ 66.5	25.0 (64.6) Δ 39.6

Table 6: Comparison of advanced token compression methods on LLaVA-OV-7B across five reduction ratios. Values are formatted as: **Group A** (Group B) with difference Δ below. Δ refers to *the absolute gap between groups*.

<i>Method</i>	<i>GQA</i>	<i>MMB</i>	<i>MMB^{CN}</i>	<i>POPE</i>	<i>MMStar</i>	<i>Average</i>
<i>Token Reduction (↓ 75.00%)</i>						
+ FastV (Chen et al., 2024b)	54.3 (84.0) Δ 29.7	70.5 (93.5) Δ 23.0	69.1 (94.7) Δ 25.6	63.8 (92.2) Δ 28.4	48.6 (73.1) Δ 24.5	61.3 (87.5) Δ 26.2
+ VisionZip (Yang et al., 2025a)	59.0 (86.2) Δ 27.2	67.7 (93.4) Δ 25.7	71.3 (94.2) Δ 22.9	80.8 (95.6) Δ 14.8	44.8 (62.9) Δ 18.1	64.7 (86.5) Δ 21.8
+ PruMerge+ (Shang et al., 2024)	60.4 (87.1) Δ 26.7	74.2 (93.8) Δ 19.6	73.5 (94.0) Δ 20.5	75.6 (96.3) Δ 20.7	48.6 (62.1) Δ 13.5	66.5 (86.7) Δ 20.2
<i>Token Reduction (↓ 88.89%)</i>						
+ FastV (Chen et al., 2024b)	45.3 (76.1) Δ 30.8	64.6 (91.7) Δ 27.1	66.4 (92.3) Δ 25.9	39.1 (85.2) Δ 46.1	42.4 (66.5) Δ 24.1	51.6 (82.4) Δ 30.8
+ VisionZip (Yang et al., 2025a)	56.6 (82.8) Δ 26.2	71.9 (92.5) Δ 20.6	71.2 (92.1) Δ 20.9	69.6 (93.6) Δ 24.0	43.5 (59.1) Δ 15.6	62.6 (84.0) Δ 21.4
+ PruMerge+ (Shang et al., 2024)	57.4 (82.9) Δ 25.5	68.8 (92.8) Δ 24.0	71.5 (93.2) Δ 21.7	76.0 (93.8) Δ 17.8	45.8 (54.9) Δ 9.1	63.9 (83.5) Δ 19.6
<i>Token Reduction (↓ 93.75%)</i>						
+ FastV (Chen et al., 2024b)	36.7 (73.0) Δ 36.3	51.2 (85.8) Δ 34.6	53.3 (86.5) Δ 33.2	29.7 (81.5) Δ 51.8	32.6 (64.3) Δ 31.7	40.7 (78.2) Δ 37.5
+ VisionZip (Yang et al., 2025a)	49.1 (79.4) Δ 30.3	64.3 (91.2) Δ 26.9	62.4 (90.9) Δ 28.5	53.6 (90.9) Δ 37.3	36.6 (54.6) Δ 18.0	53.2 (81.4) Δ 28.2
+ PruMerge+ (Shang et al., 2024)	50.2 (78.7) Δ 28.5	66.6 (91.1) Δ 24.5	65.3 (91.0) Δ 25.7	59.9 (91.0) Δ 31.1	34.8 (53.6) Δ 18.8	55.4 (81.1) Δ 25.7
<i>Token Reduction (↓ 96.00%)</i>						
+ FastV (Chen et al., 2024b)	31.4 (71.9) Δ 40.5	37.4 (77.3) Δ 39.9	43.1 (77.9) Δ 34.8	24.5 (79.0) Δ 54.5	28.6 (57.1) Δ 28.5	33.0 (72.6) Δ 39.6
+ VisionZip (Yang et al., 2025a)	42.6 (76.3) Δ 33.7	55.4 (86.9) Δ 31.5	56.9 (86.8) Δ 29.9	45.4 (86.6) Δ 41.2	30.8 (49.9) Δ 19.1	46.2 (77.3) Δ 31.1
+ PruMerge+ (Shang et al., 2024)	42.7 (74.5) Δ 31.8	57.8 (84.1) Δ 26.3	59.6 (84.2) Δ 24.6	49.9 (85.9) Δ 36.0	31.3 (49.7) Δ 18.4	48.3 (75.7) Δ 27.4
<i>Token Reduction (↓ 99.00%)</i>						
+ FastV (Chen et al., 2024b)	25.7 (63.0) Δ 37.3	25.8 (50.6) Δ 24.8	29.6 (46.5) Δ 16.9	39.3 (59.0) Δ 19.7	21.9 (47.4) Δ 25.5	28.5 (53.3) Δ 24.8
+ VisionZip (Yang et al., 2025a)	28.3 (64.4) Δ 36.1	28.1 (54.5) Δ 26.4	32.8 (53.8) Δ 21.0	42.1 (61.8) Δ 19.7	24.7 (36.0) Δ 11.3	31.2 (54.1) Δ 22.9
+ PruMerge+ (Shang et al., 2024)	25.3 (60.4) Δ 35.1	25.5 (46.9) Δ 21.4	28.5 (45.5) Δ 17.0	40.4 (56.7) Δ 16.3	25.2 (32.6) Δ 7.4	29.0 (48.4) Δ 19.4

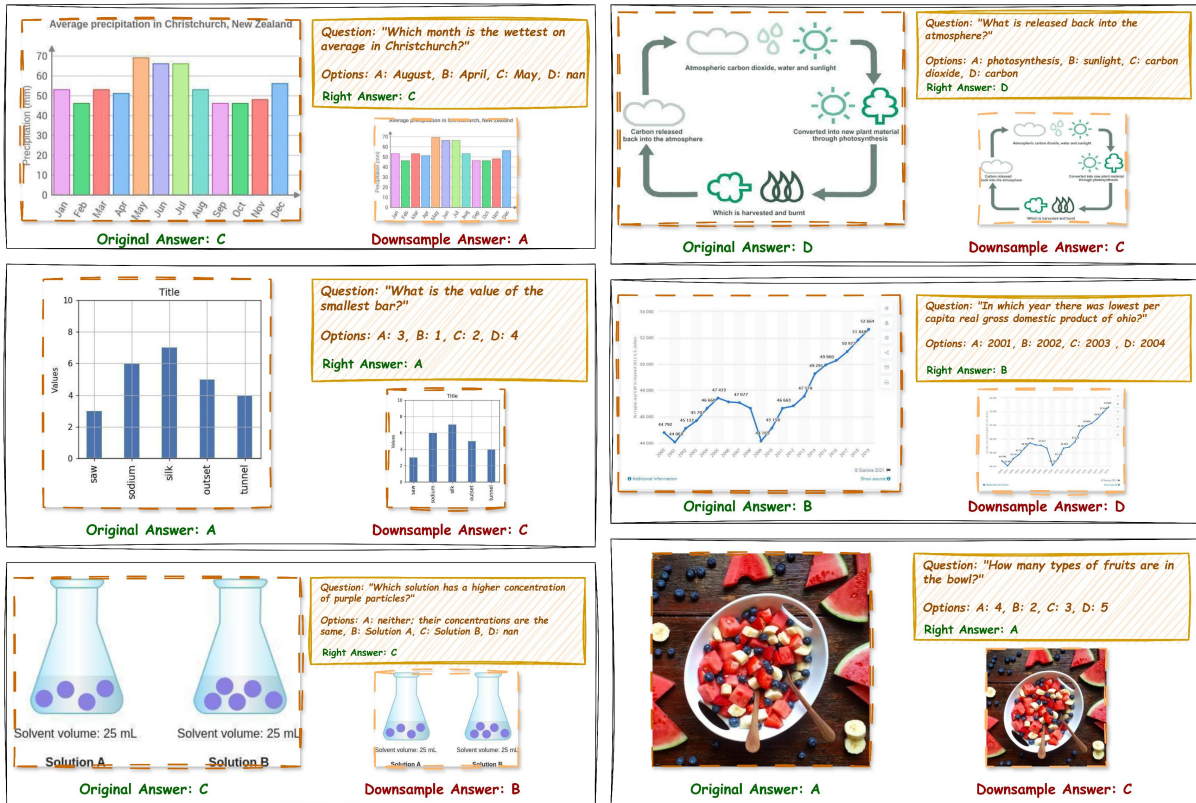


Figure 3: Visualization of “difficult” samples with downsampling ratio set to 2.

EffiVLM-Bench, researchers are capable of comparing and testing different compression methods based on the same model without the effort of adapting these methods to different models.

Different from EffiVLM-Bench, we do not aim to build an open-source toolkit for the community, but to reflect whether the benchmarking system for the visual token compression methods. Under such expectations, VTC-Bench is proposed to address the task mismatch between current benchmarks and the visual token compression task.

E Visualization between Groups

Additionally, we visualize the “difficult” and “simple” samples to observe their characteristics. As shown in Figure 3 and Figure 4, the “difficult” samples tend to require multiple detail perceptions and comparisons, such as figuring out the extreme values in a chart or performing complex counting. For example, Question (Which month is the wettest on average in Christchurch?) requires models to capture several details, including the precipitation of all 12 months, and conduct complex comparisons. Question (How many types of fruits are in the bowl?) requires models to understand large amounts of tiny fruits, which is a typical com-

plex counting question. Meanwhile, the “simple” samples usually perceive based on medium/large-sized patterns or perform simple dual-value comparisons. For example, Question (What is at the top of the food web?) only requires understanding four medium-sized patterns. Question ($f(-1)$ is $_ f(0)$.) requires only the comparison of two values. In summary, “difficult” samples often require answers based on several details, while “simple” samples usually require answers based on non-detailed or limited details.

F Statistical Analysis of Image Properties between Groups

To investigate whether the inherent “difficulty” of a sample for visual token compression could be attributed to basic, low-level visual properties, we conduct a quantitative analysis of key image statistics between samples labeled as “difficult” (Group A) and “simple” (Group B). The statistics measured included image entropy, brightness, contrast, colorfulness, and original image dimensions. This analysis is performed for both the MMStar and POPE across multiple downsampling ratios used for grouping (from 2 to 5).

The results in Table 7, and Figure 5 reveal a

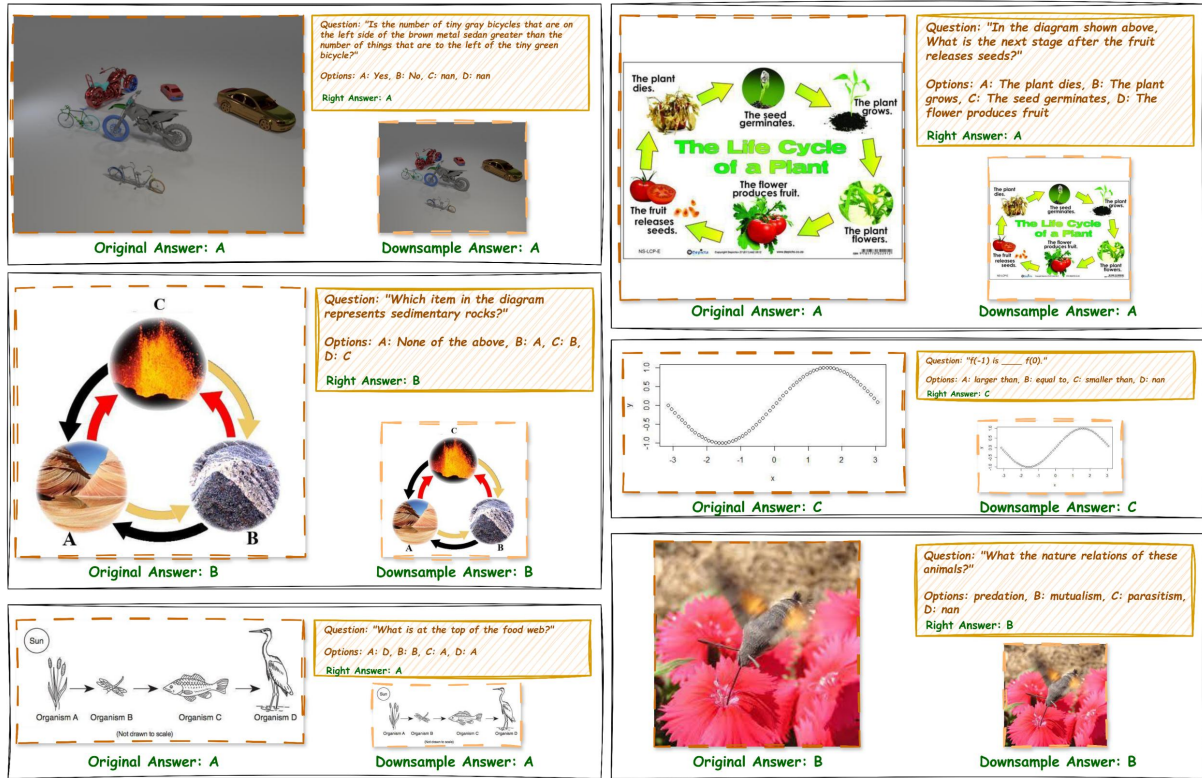


Figure 4: Visualization of "simple" samples with downsampling ratio set to 2.

Table 7: Low-level visual properties of two groups based on MMStar and POPE across different downsampling ratios.

Benchmark	Downsample	Group	Entropy	Brightness	Contrast	Colorfulness	Size ($W \pm \sigma \times H \pm \sigma$)	
MMStar (Chen et al., 2024c)	Ratio 2	Group A	4.97 ± 2.22	165.63 ± 71.60	50.99 ± 17.21	38.51 ± 28.43	$(457.87 \pm 205.37) \times (353.56 \pm 165.86)$	
		Group B	5.06 ± 2.27	167.15 ± 64.92	52.11 ± 17.62	39.79 ± 32.87	$(508.06 \pm 272.49) \times (389.30 \pm 221.20)$	
	Ratio 3	Group A	4.97 ± 2.29	163.13 ± 69.82	52.00 ± 18.38	37.91 ± 27.77	$(485.88 \pm 236.38) \times (383.46 \pm 190.36)$	
		Group B	5.08 ± 2.25	168.58 ± 64.67	51.80 ± 17.12	40.29 ± 33.75	$(502.46 \pm 270.65) \times (380.56 \pm 220.04)$	
	Ratio 4	Group A	4.99 ± 2.29	165.73 ± 68.70	52.57 ± 17.54	36.42 ± 28.05	$(486.04 \pm 230.52) \times (366.15 \pm 165.08)$	
		Group B	5.07 ± 2.24	167.50 ± 64.95	51.42 ± 17.52	41.48 ± 34.06	$(504.11 \pm 277.10) \times (391.21 \pm 234.83)$	
	Ratio 5	Group A	4.90 ± 2.31	167.97 ± 68.25	51.97 ± 17.65	36.88 ± 27.44	$(502.70 \pm 253.08) \times (382.14 \pm 198.24)$	
		Group B	5.15 ± 2.22	165.92 ± 64.96	51.79 ± 17.45	41.57 ± 34.94	$(492.74 \pm 265.51) \times (380.99 \pm 220.28)$	
	POPE (Li et al., 2023)	Ratio 2	Group A	7.39 ± 0.41	109.22 ± 28.33	62.86 ± 14.36	44.25 ± 20.71	$(583.56 \pm 89.37) \times (478.05 \pm 87.72)$
			Group B	7.40 ± 0.44	110.61 ± 28.24	61.10 ± 13.23	43.16 ± 20.46	$(583.88 \pm 86.74) \times (479.65 \pm 90.52)$
Ratio 3		Group A	7.39 ± 0.43	110.04 ± 28.62	61.35 ± 14.10	44.39 ± 20.77	$(583.67 \pm 86.12) \times (469.42 \pm 84.01)$	
		Group B	7.40 ± 0.44	110.58 ± 28.20	61.19 ± 13.21	43.09 ± 20.44	$(583.89 \pm 86.99) \times (480.72 \pm 90.99)$	
Ratio 4		Group A	7.41 ± 0.41	112.04 ± 27.46	61.89 ± 13.85	43.76 ± 19.59	$(579.60 \pm 89.42) \times (477.16 \pm 86.95)$	
		Group B	7.39 ± 0.45	110.27 ± 28.37	61.09 ± 13.21	43.14 ± 20.62	$(584.58 \pm 86.45) \times (479.95 \pm 90.91)$	
Ratio 5		Group A	7.40 ± 0.43	111.47 ± 28.01	61.66 ± 13.43	44.58 ± 20.15	$(584.90 \pm 84.97) \times (475.62 \pm 87.85)$	
		Group B	7.39 ± 0.44	110.35 ± 28.29	61.12 ± 13.28	42.98 ± 20.53	$(583.67 \pm 87.25) \times (480.28 \pm 90.80)$	

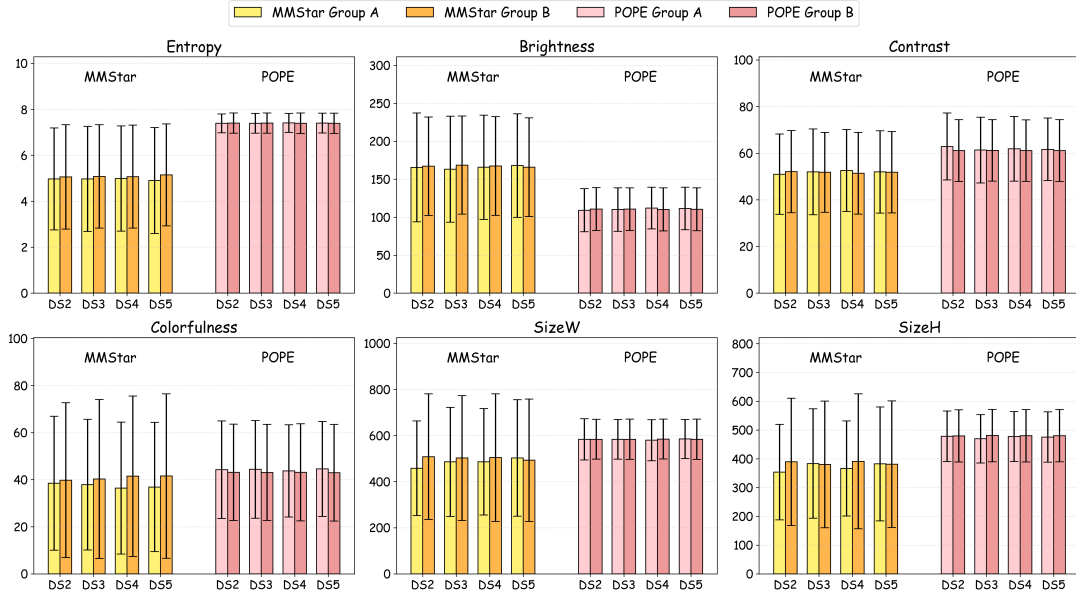


Figure 5: Visualization of low-level visual properties of two groups based on MMStar and POPE across different downsampling ratios. DSx refers to that downsampling ratio is set to x.

consistent and crucial pattern: there is no statistically significant difference in these low-level image statistics between Group A and Group B for either dataset. For instance, in MMStar, the average entropy ranges between 4.90 and 5.15 for both groups, while brightness values are centered around 163-168. Similarly, in POPE, the average entropy remains at approximately 7.40, and the average brightness varies minimally around 110 for both groups. The substantial overlap in the ranges indicated by the standard deviations confirms this lack of separation.

This finding is pivotal. It demonstrates that simple, perceptual image properties cannot explain or easily predict the “difficulty” of samples. Samples that are “difficult” are not systematically brighter, more colorful, higher in contrast, or larger in size than simpler samples. Consequently, the defining characteristic of a “difficult” sample must lie beyond these elementary features.

It is precisely the absence of a straightforward, low-level metric to distinguish difficulty that validates our methodological choice to use downsampling as an operational and effective proxy for identifying samples that challenge the model’s visual compression. However, although downsampling effectively distinguishes the “difficulty” of samples, it is equally important to further find reasonable and high-level metrics to measure difficulty, which requires us to explore further in the future.

G Inference Time Comparison

In this section, we provide a detailed comparison of the inference time between the simple downsampling baseline and one other advanced visual token compression method⁴. Considering that different methods optimize code complexity differently and use different acceleration tools, we compare DART and downsampling based on the same code repository from DART within 1 A800. As shown in Table 8, the reduction in tokens has far less impact on speed than we had anticipated in DART, while downsampling shows more normal results. We believe this is most likely due to the time overhead caused by additional metric calculations in the sophisticated visual token compression methods, which further proves the efficiency of downsampling in the previous evaluation system of visual token compression methods.

Table 8: Inference time comparison of downsampling and DART (Wen et al., 2025a). *Budget* refers to the proportion of remaining visual tokens.

<i>Budget</i>	<i>Downsample</i>	<i>DART</i>
1.0000	30566s	-
0.2500	17452s	25200s
0.1111	14846s	24094s
0.0625	13584s	24079s
0.0400	12975s	23690s
0.0100	12802s	23386s

⁴Inference time is collected through our training logs for rough comparison, which is calculated based on the inference start time and end time of all 8 benchmarks.