Supplementary Material: *A Stitch in Time Saves Nine*: Small VLM is a Precise Guidance for Accelerating Large VLMs

Wangbo Zhao[1*] Yizeng Han[2*] Jiasheng Tang[2,3] Zhikai Li[1] Yibing Song[2,3]
Kai Wang[1†] Zhangyang Wang[4] Yang You[1†]
[1]National University of Singapore [2]DAMO Academy, Alibaba Group
[3]Hupan Lab [4]The University of Texas at Austin

1. SGP with SEE Towards Improved Efficiency

In Figure 1 and Figure 2, we further validate the superiority of our SGL by incorporating both SGP and SEE mechanisms, on SEED [11] and RefCOCO [18] benchmarks. It can be observed that, with the 26B large VLM, our method SGP without SEE yields slower inference compared to FastV and ToMe. This is attributed to the computational overhead of the 2B small VLM, particularly on RefCOCO, where it requires a non-negligible amount of time to auto-regressively generate a greater number of tokens compared to other datasets *e.g.* SEED. However, scaling the large VLM to 40B and 76B results in competitive inference speeds and superior performance relative to FastV and ToMe, particularly at low token retention ratios.

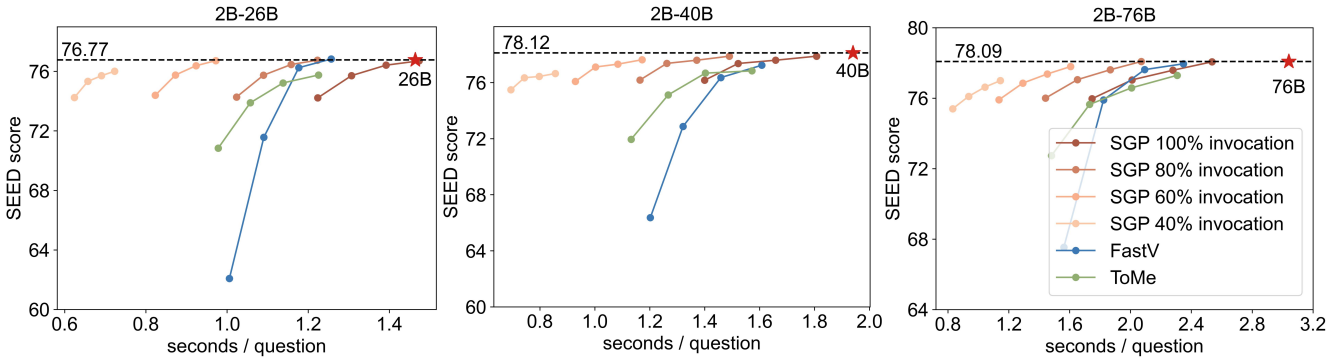

Figure 1. Performance-efficiency curves of SGL (SGP + SEE) on SEED [11]. The results with 18%, 35%, 50%, and 64% visual token retention ratios are presented as a curve. For the 26B and 40B, we use an NVIDIA H20 GPU, and the 76B is sharded on two GPUs.

2. Memory Efficiency

In this section, we analyze the memory allocation of SGL. Our approach incorporates a small VLM *e.g.* InternVL-2B in addition to the large VLM, which may introduce some additional memory overhead. Fortunately, the small VLM consumes only a minimal portion of memory compared to the large model. As a result, our method retains memory efficiency, as verified in Table 1.

3. Visualization

In Figure 3, we provide additional visualizations of examples where the small VLM (2B) fails to produce correct predictions, while the large VLM (26B), with visual tokens pruned by SGP, successfully predicts the correct answers. Notably, in these

*Equal contribution. Work done during an internship at DAMO Academy, Alibaba Group. wangbo.zhao96@gmail.com
†Corresponding author.

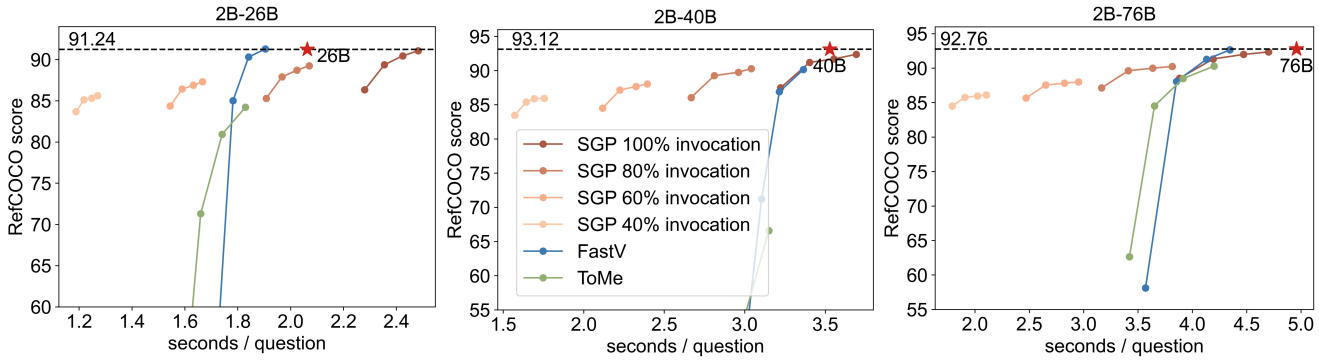

Figure 2. Performance-efficiency curves of SGL (SGP + SEE) on RefCOCO [18]. The results with 18%, 35%, 50%, and 64% visual token retention ratios are presented as a curve. For the 26B and 40B, we use an NVIDIA H20 GPU, and the 76B is sharded on two GPUs.

small VLM	small VLM memory	large VLM	large VLM peak memory	large VLM with SGL peak memory	Δ
2B	4.48 GiB	26B	51.60 GiB	54.24 GiB	+2.64GiB (5.11%)
2B	4.48 GiB	40B	77.94 GiB	80.60 GiB	+2.66GiB (3.41%)
2B	4.48 GiB	76B	147.64 GiB	147.25 GiB	-0.39 GiB (0.26%)

Table 1. **Mmeory analysis of SGL**. The meory of our method is measured with 9% average retention ratio. "small VLM memory" refers to the memory required to load the single small VLM. "Large VLM peak memory" represents the peak memory usage during inference with only the large VLM. "Large VLM with SGL peak memory" indicates the peak memory usage during inference of the large VLM when using the proposed SGL method (guided by a 2B model). Δ is defined as the difference between "Large VLM with SGL peak memory" and "Large VLM peak memory". We report the ratio of Δ relative to "Large VLM peak memory".

cases, the large VLM with FastV [5] also fails.

4. Model Descriptions

The configurations of InternVL [6], QWen2-VL [15], and LLaVa-OV [12] are comprehensively detailed in Tables 2, 3, and 4, respectively.

model name	language model	vision encoder	checkpoint
InternVL-1B	Qwen2-0.5B [17]	InternViT-300M [7]	link
InternVL-2B	InternLM2-chat-1.8B [3]	InternViT-300M [7]	link
InternVL-4B	Phi-3-mini-128k-instruct [1]	InternViT-300M [7]	link
InternVL-26B	InternLM2-chat-20B [3]	InternViT-6B [7]	link
InternVL-40B	Nous-Hermes-2-Yi-34B [14]	InternViT-6B [7]	link
InternVL-76B	Hermes-2-Theta-Llama-3-70B [13]	InternViT-6B [7]	link

Table 2. **Model descriptions of InternVL [6]**

model name	language model	vision encoder	checkpoint
Qwen2-VL-2B	Qwen2-1.5B [17]	ViT [8]	link
Qwen2-VL-76B	Qwen2-72B [17]	ViT [8]	link

Table 3. **Model descriptions of QWen2-VL [15]**

5. Generalization to Video Benchmarks

Understanding video content is a critical capability of VLMs [4, 9, 10, 16]. Unlike image tasks, video tasks require VLMs to process significantly more visual tokens, posing additional challenges. To demonstrate the effectiveness of our method in this

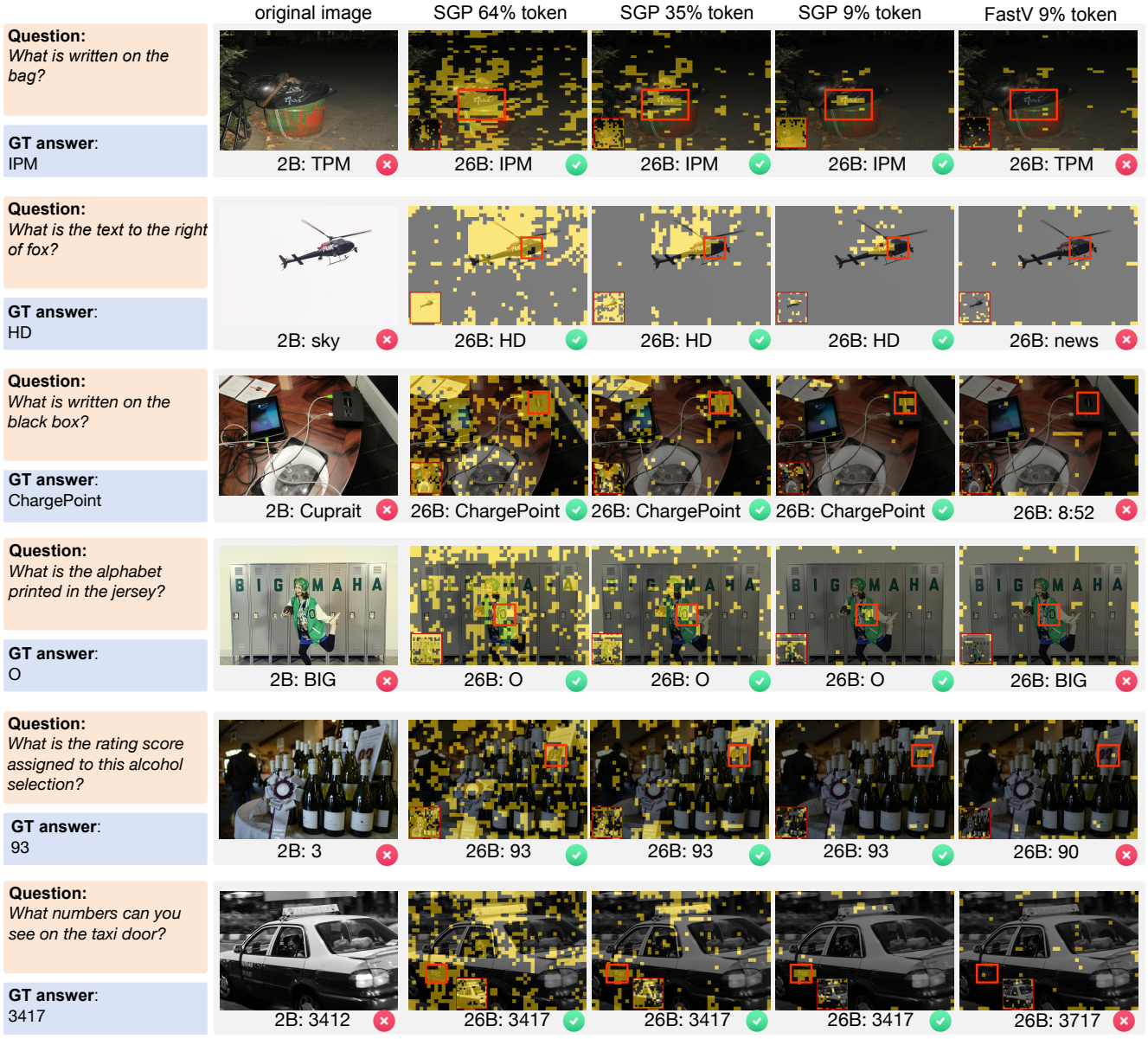

Figure 3. **Additional visualization of SGP under different visual token retention ratios and answers.** Visual tokens are pruned by 60%, 80%, and 95% at the 19th, 9th, and 2nd layers of the large VLM of 26B, which comprises 48 layers. This results in average token retention ratios of 64%, 35%, and 9%, respectively. Retained tokens are highlighted with ■. Thumbnails employed in InternVL are presented in the left corner.

model name	language model	vision encoder	checkpoint
LLaVa-OV-0.5B	Qwen2-0.5B [17]	SigLIP [19]	link
LLaVa-OV-72B	Qwen2-72B [17]	SigLIP [19]	link

Table 4. **Model descriptions of LLaVa-OV [12].**

context, we present results on three video benchmarks—VideoMME [10], MMBench-Video [9], and LongVideoBench [16]—in Table 5. The results show that our method consistently outperforms ToMe [2] and FastV [5] across all benchmarks.

method	token ratio	VideoMME				MMBench-Video			LongVideoBench
		Short	Medium	Long	Overall	Perception	Reasoning	Overall	Overall
InternVL-26B [6]	100%	63.0	50.3	44.2	52.5	1.69	1.64	1.68	53.9
InternVL-2B [6]	100%	55.0	40.8	35.4	43.7	1.47	1.39	1.44	45.0
26B w/ ToMe [2]	64%	61.8	49.1	44.4	51.8	1.66	1.64	1.66	53.6
	35%	61.1	49.6	43.6	51.4	1.61	1.62	1.61	52.4
	9%	52.0	45.7	41.8	46.5	1.39	1.48	1.43	47.6
26B w/ FastV [5]	64%	63.8	50.4	45.2	53.1	1.65	1.62	1.65	53.6
	35%	51.3	45.8	41.9	46.3	1.49	1.50	1.50	48.5
	9%	41.6	42.2	40.1	41.3	1.24	1.36	1.28	42.4
26B w/ SGP (ours)	64%	63.1	50.2	44.1	52.5	1.67	1.63	1.66	54.0
	35%	61.8	50.3	42.4	51.5	1.66	1.59	1.64	52.2
	9%	54.6	46.4	41.0	47.3	1.48	1.50	1.49	49.0

Table 5. **Comparison between SGP and previous visual token pruning methods on video benchmarks.**

6. Frequent Questions

Whether the large and small models can come from different model families?
Our method can generalize scenarios where small&large VLMs come from different families. For example, using InternVL-2B to guide pruning 35% of visual tokens in LLaVa-OV-72B achieves a speedup of **2.05x** and a 75.36 score on TextVQA (vs. 79.30). However, using small VLMs from the same family is simpler, as a VLM family often include models of different sizes.

Can the proposed method achieve acceleration for a 7B model?
SGL performs better when the size gap between the small VLM and the large VLM is greater. To evaluate the effect on a smaller model, we conduct experiments on LLaVa-OV 0.5B-7B models, finding that $\text{SGP}_{60\%\text{SEE}}$ achieves a 1.2x speedup with 9% token retention, yielding a TextVQA score of 70.55 (vs. 75.91). Therefore, we recommend applying SGL for larger VLMs.

Compared to speculative decoding, which also uses a small model for acceleration, what are the advantages of this method in inference speedup?
(i) Our method avoids invoking the large VLM for easy questions, unlike speculative decoding (SD), which frequently and inevitably activates the large model, leading to overhead, particularly on tasks (*e.g.* QA tasks) generating short sequences. For example, SD for 2B-76B takes 3.17s/question (vs. 3.02s for 76B) on TextVQA. In contrast, $\text{SGP}_{40\%\text{SEE}}$ (retaining 64% of visual tokens) reduces latency to 1.73s/question.
(ii) The proposed SGP, a visual token pruning method, is orthogonal to SD. Combining it with SD for long-sequence generation tasks would be an promising direction.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 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. 3, 4

[3] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 2

[4] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 2

[5] 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. *arXiv preprint arXiv:2403.06764*, 2024. 2, 3, 4

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 4

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 2, 3

[10] 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. 2, 3

[11] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3

[13] NousResearch. Hermes-2-theta-llama-3-70b. https://huggingface.co/NousResearch/Hermes-2-Theta-Llama-3-70B, . 2

[14] NousResearch. Nous-hermes-2-yi-34b. https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B, . 2

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[16] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 2, 3

[17] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2, 3

[18] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1, 2

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3