Switch-KD: Visual-Switch Knowledge Distillation for Vision-Language Models

Supplementary Material

Supplementary Overview

This supplementary material provides an overview of the content in each appendix section:

- **Section A**: Extended Background and Related Work.
- **Section B**: Additional Implement Details.
- **Section C**: Comparison with SOTA Distilled VLMs
- **Section D**: More Ablation and Explanatory Analysis.
- **Section E**: Detailed Results.

A. Extended Background and Related Work

We present a more comprehensive review of related work, expanding upon the brief discussion and focusing on advances in large and lightweight vision–language models.

Large Vision-Language Models: Recent advances in large vision-language models (VLMs) have led to significant progress in visual-language understanding. Early methods, such as CLIP [19], typically relied on contrastive learning over roughly 400 million image–text pairs to establish alignment between visual and textual modalities. Subsequently, models such as BLIP-2 [10] adopted a paradigm of integrating a frozen pretrained vision encoder with a large language model (LLM), and introduced a lightweight Q-Former bridging module to enable multimodal understanding and generation. More recently, research [1, 13, 22] has largely adopted the architecture of ViT–Projector–LLM, and applied instruction tuning to enable multimodal dialogue and reasoning capabilities. Regarding the performance improvements of large VLMs, numerous studies [9, 17, 18] show that their gains roughly follow the scaling law observed in language models. As model size, training data, and compute increase, performance typically improves. Nevertheless, in resource-constrained or real-time deployment scenarios, these large VLMs face serious bottlenecks in terms of compute resource requirements, storage footprint, and inference latency, making lightweight alternatives particularly necessary.

Lightweight Vision-Language Models: Recently, researchers [3, 11, 12, 23] have proposed a series of lightweight vision-language model frameworks from three perspectives: architecture design, data quality, and training strategy. TinyLLaVA [23] introduces compact LLM backbones and efficient fine-tuning strategies, achieving performance comparable to 7B-scale models within only 1B–3B parameters. Mini-Gemini [11] adopts a dual-visual-encoder design and a patch information mining mechanism to effectively integrate fine-grained details from high-resolution images while maintaining low computational cost. SPHINX-X [12] further streamlines the visual encoder design, incorporates skip tokens for architectural efficiency, and adopts a unified one-stage training scheme that facilitates more efficient deployment. MobileVLM V2 [3] employs the lightweight LDPv2 projector, a high-quality 3.6M dataset, and a full-LLM pre-training strategy, achieving higher benchmark accuracy and the fastest inference speed among similarly scaled models on mobile devices.

B. Additional Implement Details

Training Phase	Task	Datasets (# Size)
PT	Caption	LCS (558K)
SFT / DFT	VQA	VQAv2 (83K) GQA (72K) OKVQA (9K) A-OKVQA (50K)
	OCR	OCRVQA (80K) TextCaps (22K)
	Region	RefCOCO (30K) VG (86K)
	Conversation	LLaVA (158K) ShareGPT (40K)
Total		**1.2M**

Table 1. Dataset composition for each training phase.

Training datasets. Table 1 summarizes the dataset composition across different training phases. The pretraining dataset LCS-558K consists of 558k image–text pairs from LAION-CC-SBU, annotated with BLIP captions. The 665k instruction-following dataset includes a diverse collection of tasks: VQA [5, 6, 15, 20], OCR [16, 21], region-level VQA [7, 8, 14], visual conversation [13], and general language conversation data.

Training Hyperparameters. We adopt a similar set of hyperparameters to those used in LLaVA-KD. The configurations for the first-stage vision–language alignment pretraining and the second-stage distillation-based instruction tuning are summarized in Table 2.

Pseudocode for the Proposed Method. For clarity and reproducibility, we present pseudo-code illustrating the implementation of our *Visual-Switch Distillation* framework and its core *DBiLD Loss* in Algorithm 1 and Algorithm 2, respectively. The first algorithm outlines the overall visual-switch distillation process, while the second specifies the

Hyperparameter	PT	DFT
Visual Encoder	×	✓
Projector	✓	✓
LLM	×	✓
Image Resolution	384×384	
Learning Rate	1e-3	2e-5
Optimizer	AdamW	
Scheduler	Cosine decay	
Warm up ratio	0.03	
Global Batch Size	256	128
Epoch	1	
DeepSpeed stage	Zero 2	Zero 2

Table 2. Hyperparameters of Switch-KD.

dynamic bidirectional logits alignment procedure.

Algorithm 1 Visual-Switch Distillation

Input: image $\mathbf{x}_v$, prompt $\mathbf{x}_t$, ground-truth $\mathbf{y}_t$; teacher modules (V^T, P^T, L^T); student modules (V^S, P^S, L^S); temperature τ; loss weights λ_1, λ_2.
Output: total training loss $\mathcal{L}$.

// Standard Alignment Pathway
compute teacher logits $\mathbf{z}^T = L^T(P^T(V^T(\mathbf{x}_v)), \mathbf{x}_t)$
compute student logits $\mathbf{z}^S = L^S(P^S(V^S(\mathbf{x}_v)), \mathbf{x}_t)$
$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{DBiLD}}(\mathbf{z}^T, \mathbf{z}^S)$
// Visual-Switch Pathway
compute visual-switch logits
$\mathbf{z}^{Switch} = L^T(P^T(V^S(\mathbf{x}_v)), \mathbf{x}_t)$
$\mathcal{L}_{\text{VSD}} = \mathcal{L}_{\text{DBiLD}}(\mathbf{z}^T, \mathbf{z}^{Switch})$
// Language modeling objective
compute student next-token distribution
$\mathbf{p}^S = \text{softmax}(\mathbf{z}^S/\tau)$
$\mathcal{L}_{\text{CE}} = -\sum_t \log \mathbf{p}^S(\mathbf{y}_t \,|\, \mathbf{x}_v, \mathbf{y}_{<t})$
// Final objective
$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Align}} + \lambda_2 \mathcal{L}_{\text{VSD}}$
return $\mathcal{L}$

C. Comparison with SOTA Distilled VLMs

LLaVA-KD [2] represents a state-of-the-art approach for VLM distillation, achieving knowledge transfer via explicit alignment of visual logits, language logits, and self-correlation matrices of LLM-generated visual tokens. We evaluate our method on the same benchmarks and training data as LLaVA-KD to ensure a fair comparison. Switch-KD achieves notable performance improvements at equivalent model scales, with average gains of 1.1 and 0.4 points for 0.5B and 1.5B student models, respectively. To further clarify the similarities and distinctions between the two methods, we summarize them below:

Algorithm 2 Calculation of DBiLD Loss

Input: teacher logits $\mathbf{z}^t$, student logits $\mathbf{z}^s$.
Output: the DBiLD loss $\mathcal{L}_{\text{DBiLD}}$.

// Teacher-guided Loss
compute the transition point k^t on sorted $\mathbf{z}^t$
select top-k^t teacher logits $\mathbf{z}^t_{led}$
select corresponding student logits $\mathbf{z}^s_{cor}$
build pairwise differences $\mathbf{d}^t_{led}$ and $\mathbf{d}^s_{cor}$
normalize differences to probabilities $\mathbf{p}^t_{led}$ and $\mathbf{p}^s_{cor}$
$\mathcal{L}_t = D_{\text{RKL}}[\mathbf{p}^t_{led} \,\|\, \mathbf{p}^s_{cor}]$
// Student-guided Loss
compute the transition point k^s on sorted $\mathbf{z}^s$
select top-k^s student logits $\mathbf{z}^s_{led}$
select corresponding teacher logits $\mathbf{z}^t_{cor}$
build pairwise differences $\mathbf{d}^s_{led}$ and $\mathbf{d}^t_{cor}$
normalize differences to probabilities $\mathbf{p}^s_{led}$ and $\mathbf{p}^t_{cor}$
$\mathcal{L}_s = D_{\text{RKL}}[\mathbf{p}^t_{cor} \,\|\, \mathbf{p}^s_{led}]$
// Final objective
$\mathcal{L}_{\text{DBiLD}} = \mathcal{L}_t + \mathcal{L}_s$
return $\mathcal{L}_{\text{DBiLD}}$

- **Architecture Design.** Similar to LLaVA-KD, Switch-KD maintains a simple yet effective architecture for the student VLM without introducing additional complexity or specialized modules.
- **Training Scheme.** LLaVA-KD adopts a three-stage training framework comprising (1) Distilled Pre-Training (DPT) for visual–textual alignment,(2) Supervised Fine-Tuning (SFT) for task-specific knowledge acquisition, and (3) Distilled Fine-Tuning (DFT) for teacher–student knowledge transfer. In contrast, our framework uses only standard Pre-Training (PT) followed by DFT with a visual-switch distillation design, effectively achieving both efficient knowledge acquisition and knowledge transfer, without introducing any additional training stages.
- **Distillation Strategy.** LLaVA-KD employs dedicated knowledge distillation strategies (MDist/RDist) across both DPT and DFT stages. In contrast, Switch-KD introduces a dynamically attentive distillation mechanism through the proposed DBiLD loss during the DFT stage, enabling adaptive alignment of informative logits between teacher and student.

Align-KD [4] is another recent advance in multimodal distillation, achieving knowledge transfer via explicit alignment of cross-modal attention, visual tokens, and language logits. We evaluate our approach on the same benchmarks as Align-KD, and despite using only about one-third of the training data (3M vs. 1.2M samples) and a lighter language backbone (1.7B vs. 1.5B parameters), Switch-KD achieves an average performance gain of 4.4 points. To further high-

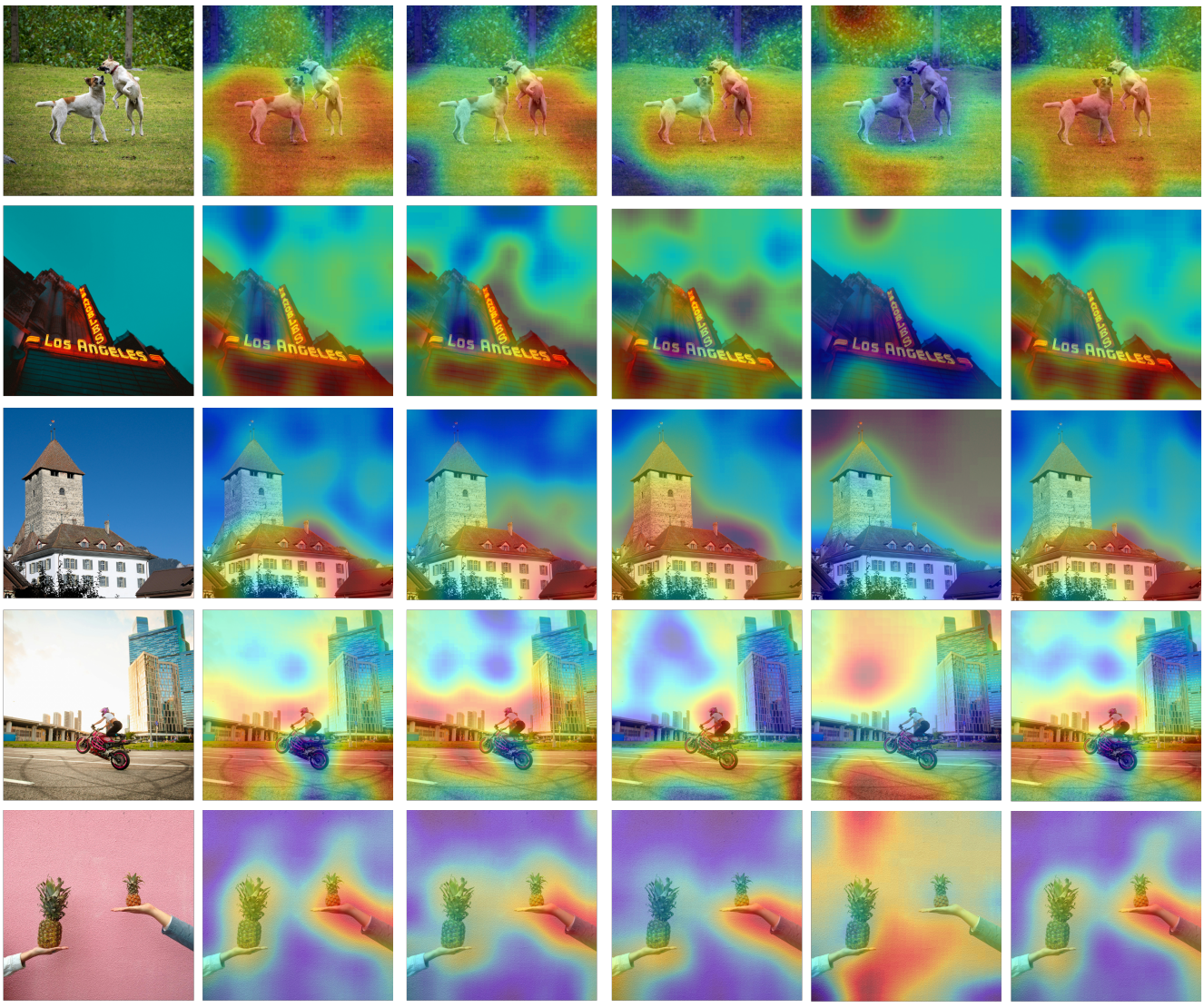

Figure 1. Visualization of attention maps for different distillation strategies.

light the similarities and differences between Align-KD and our approach, we summarize three key aspects below:

- **Architecture Design.** Both methods adopt similar vision–language model architectures, while our approach employs a lighter LLM backbone, reducing computational cost without sacrificing performance.
- **Training Scheme.** Both frameworks utilize a two-stage training paradigm in which distillation is introduced during the multi-task fine-tuning phase.
- **Distillation Strategy.** Align-KD distills knowledge from the text-query–vision part of the first attention layer and selectively, although unevenly, enhances vision tokens based on the attention focus of text tokens. In contrast, our method performs unified and adaptive multimodal alignment across the entire output distribution via the proposed DBiLD loss.

D. More Ablation and Explanatory Analysis

D.1. More Visualization of Attention Maps

Fig. 1 presents additional qualitative comparisons of visual attention across five representative image pairs. The teacher model consistently focuses on semantically critical regions that closely correspond to the textual queries. The SFT baseline produces diffuse and unstable attention, often failing to localize salient regions or spreading attention across large background areas, indicating insufficient visual grounding. LLaVA-KD mainly focuses on key semantic re-

Training Strategy	Percep. & Underst.			Cognition & Reasoning				OCR	Specific	Halluc.	Avg_{10}
	MME	MMB	MMB^{CN}	VQAv2	GQA	SciQA	MMMU	TextVQA	VizWiz	POPE	
PT+SFT (V^S)	61.5	58.9	54.2	74.8	58.3	**59.1**	**33.6**	49.2	28.9	86.1	56.5
PT+SFT (V^T)	62.0	55.2	51.3	75.7	58.0	57.9	29.7	44.7	41.3	84.6	56.0
Switch-KD	**66.8**	**63.5**	**57.8**	**79.6**	**61.6**	57.9	29.8	**52.3**	**44.9**	**87.3**	**60.1**

Table 3. Ablation study on direct visual encoder substitution. **PT+SFT** (V^S) denotes standard pre-training (PT) and supervised fine-tuning (SFT) with the original visual encoder V^S. **PT+SFT** (V^T) denotes the same PT+SFT pipeline but with the teacher visual encoder V^T replacing V^S. **Switch-KD** denotes pre-training followed by distilled fine-tuning (DFT) with the proposed visual-switch distillation. Simply substituting V^T does not outperform the baseline, whereas Switch-KD consistently achieves superior performance.

Training Scheme	Percep. & Underst.			Cognition & Reasoning				OCR	Specific	Halluc.	Avg_{10}
	MME	MMB	MMB^{CN}	VQAv2	GQA	SciQA	MMMU	TextVQA	VizWiz	POPE	
PT-SFT	61.5	58.9	54.2	74.8	58.3	**59.1**	**33.6**	49.2	28.9	86.1	56.5
DPT-SFT	**65.6**	57.4	53.2	76.5	58.6	58.4	32.2	50.2	35.4	86.8	57.4
PT-DFT	63.1	**60.8**	57.3	**77.8**	**59.7**	58.1	31.7	51.0	**41.5**	**87.0**	**58.8**
DPT-DFT	63.6	60.5	**57.4**	77.2	59.6	58.2	32.0	**51.1**	41.4	86.4	58.7

Table 4. Ablation study on different training schemes.

gions, but still exhibits noisy patches and inconsistent localization across images. Align-KD tends to focus on background regions rather than key objects, resulting in overly diffuse attention. Across all examples, Our Switch-KD maintains stable object-centric focus and effectively reconstructs the teacher's fine-grained attention pattern, demonstrating that its visual-switch architecture and dynamic logit selection effectively guides the student to learn more faithful and interpretable visual representations.

D.2. Further Validation of the Visual-Switch Hypothesis

We formulated the core hypothesis that if the student visual encoder V^S learns meaningful representations, its visual features should be correctly interpreted and decoded by the teacher's language pathway, yielding a probability distribution consistent with the teacher's original output. A natural follow-up question is whether the performance could be improved simply by replacing the student visual encoder with the teacher's visual encoder, thereby starting PT and SFT with higher-quality visual representations. To investigate this, we conduct an experiment where the student model is initialized with the teacher's visual encoder and trained under the standard PT+SFT paradigm. Surprisingly, as shown in Table 3, this direct substitution not only fails to surpass the baseline performance, but in some cases even performs worse—despite the teacher encoder providing substantially stronger visual representations. This indicates that simply transplanting a high-capacity encoder does not guarantee effective knowledge transfer, likely due to representational mismatches between the teacher's visual features and the student's language pathway, as well as the absence of explicit alignment signals during PT and SFT. In contrast, our proposed Switch-KD consistently yields notable improvements. These results suggest that Switch-KD achieves a well-balanced integration: effectively learning the teacher's visual knowledge while aligning it with the student's language backbone through unified text-probability supervision.

E. Detailed Results

E.1. Ablation Study on Training Schemes

Table 4 shows more detailed results of the ablation study on different training schemes, including PT-SFT, DPT-SFT, PT-DFT and DPT-DFT. DFT–SFT notably improves the student model's performance on the MME and VizWiz datasets, but the overall gain remains limited. PT–DFT outperforms PT–SFT by 2.3 point, indicating that DFT contributes more than SFT to downstream performance. DPT–DFT achieves results largely comparable to PT–SFT across all datasets, suggesting that the performance improvements mainly stem from DFT. Therefore, we adopt the PT–DFT training scheme to ensure an efficient and effective training paradigm.

Teacher	Student	Percep. & Underst.			Cognition & Reasoning				OCR	Specific	Halluc.	Avg_{10}
		MME	MMB	MMB^{CN}	VQAv2	GQA	SciQA	MMMU	TextVQA	VizWiz	POPE	
7B	/	77.4	74.9	74.4	81.3	64.0	73.6	41.6	60.3	53.9	86.8	68.8
3B	/	73.9	71.8	69.5	80.4	63.2	76.0	40.3	61.5	38.7	86.4	66.2
/	1.5B	72.5	68.6	63.0	78.8	62.0	**72.0**	**37.0**	57.4	43.2	85.5	64.0
7B	1.5B	72.2	71.4	**68.5**	**81.4**	**63.9**	69.3	34.9	**60.3**	**44.4**	86.8	**65.3**
3B	1.5B	**72.6**	**72.2**	68.3	80.6	62.2	70.0	33.9	59.1	42.5	**86.9**	64.8
/	0.5B	61.5	58.9	54.2	74.8	58.3	**59.1**	**33.6**	49.2	28.9	86.1	56.5
7B	0.5B	66.1	61.6	56.9	**79.7**	**62.1**	58.1	31.2	**52.6**	42.6	87.0	59.8
3B	0.5B	**66.8**	**63.5**	**57.8**	79.6	61.6	57.9	29.8	52.3	**44.9**	**87.3**	**60.1**

Table 5. Ablation study on teacher models with different sizes.

E.2. Ablation study on Teacher's Size

Table 5 shows more detailed results of the ablation study that investigates the impact of different teacher sizes. Across all teacher sizes, Switch-KD consistently improves performance, indicating that it effectively transfers useful knowledge while mitigating teacher–student mismatch. For the 1.5B student, larger teachers lead to higher accuracy: the 7B teacher achieves the best overall score of 65.3, outperforming both the 3B teacher the no-teacher baseline. This shows that stronger teachers provide richer supervisory signals, especially for reasoning-heavy tasks such as VQAv2 and SciQA. For the 0.5B student, distilling from a 3B or 7B teacher yields clear improvements over the baseline, with the 3B teacher slightly outperforming the 7B teacher. This suggests a capacity–compatibility effect: extremely large teachers may produce representations too difficult for very small students to fully understand.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 1

[2] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024. 2

[3] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 1

[4] Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4178–4188, 2025. 2

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

[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] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[9] Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024. 1

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[11] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1

[12] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang,

Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[14] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[15] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

[16] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 1

[17] Marianna Nezhurina, Tomer Porian, Giovanni Pucceti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling laws for robust comparison of open foundation language-vision models and datasets. *arXiv preprint arXiv:2506.04598*, 2025. 1

[18] Cheng Peng, Kai Zhang, Mengxian Lyu, Hongfang Liu, Lichao Sun, and Yonghui Wu. Scaling up biomedical vision-language models: Fine-tuning, instruction tuning, and multi-modal learning. *arXiv preprint arXiv:2505.17436*, 2025. 1

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[20] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1

[21] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020. 1

[22] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[23] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 1