VladVA: Discriminative Fine-tuning of LVLMs

Supplementary Material

1. Results for additional model sizes and architectures

To further showcase the generalizability of our approach, herein we report results on two additional models: LLaVA-1.5-13B [8] and Qwen2-VL-2B [14]. The 1st is a scaled-up version of the LLaVA-1.5-7B [8] used in the main manuscript and tests the scalability of our approach with size. The second follows a different architecture and training procedure and has "only" 2B parameters, testing both generalizations to different architectures and finetuning in a lower-parameters regime. As the results from Tables 1, 2 3 and 4 show, on all 6 datasets (*i.e*. Flickr, coco, nocaps, SugarCrepe, SugarCrepe++ and Winoground) for both retrieval and compositionality, in all cases we significantly improve upon the original zero-shot model performance, showing good scalability with size in both directions, *i.e*. for smaller and bigger models.

2. Compositionality evaluation on Winoground

In addition to the results from the main paper, herein, we report results on Winoground [13], a curated dataset consisting of 400 images with difficult/unusual scenarios that go beyond compositionality and largely act as a natural adversarial set [3, 15]. As the results from Table 4 show, our

Table 1. Zero-shot text-image retrieval accuracy on Flickr30K, COCO and nocaps.

Method	**image** retrieval						**text** retrieval					
	Flickr30K		COCO		nocaps		Flickr30K		COCO		nocaps	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
Qwen2-VL-2B [14]	54.1	86.0	32.4	68.2	41.2	80.1	59.6	89.2	35.3	71.8	54.0	90.3
VladVA (Ours) (`Qwen2-VL-2B`)	**80.4**	**97.3**	**52.5**	**84.4**	**68.3**	**94.9**	**93.7**	**99.9**	**71.9**	**93.9**	**86.0**	**99.4**
LLaVA-1.5-7B [8]	59.6	89.3	34.4	69.6	46.9	83.3	65.6	92.3	35.6	70.5	52.1	88.1
VladVA (Ours) (`LLaVA-1.5-7B`)	**85.0**	**98.5**	**59.0**	**88.6**	**72.3**	**96.5**	**94.3**	**99.9**	**72.9**	**94.4**	**85.7**	**99.5**
LLaVA-1.5-13B [8]	61.7	90.4	37.9	74.1	48.4	85.0	66.9	93.6	35.3	71.0	48.0	87.9
VladVA (Ours) (`LLaVA-1.5-13B`)	**85.6**	**98.6**	**58.2**	**88.4**	**74.0**	**96.6**	**94.5**	**99.8**	**75.0**	**95.6**	**85.4**	**99.6**

Table 2. Zero-shot results on SugarCrepe compositionality benchmark.

Method	Params	Replace			Swap		Add	
	(B)	Object	Attribute	Relation	Object	Attribute	Object	Attribute
Qwen2-VL-2B [14]	2.21	89.9	72.0	75.0	56.1	56.1	73.2	70.1
VladVA (Ours) (`Qwen2-VL-2B`)	2.21	**97.9**	**89.7**	**81.5**	**76.5**	**82.6**	**93.6**	**95.4**
LLaVA-1.5-7B [8]	7.06	88.0	81.6	76.1	60.9	58.8	67.0	62.4
VladVA (Ours) (`LLaVA-1.5-7B`)	7.06	**98.1**	**92.1**	**86.8**	**79.0**	**82.9**	**95.2**	**95.8**
LLaVA-1.5-13B [8]	13.35	90.0	80.6	76.3	71.8	61.9	69.3	59.1
VladVA (Ours) (`LLaVA-1.5-13B`)	13.35	**98.1**	**93.9**	**89.8**	**81.1**	**86.0**	**95.2**	**97.0**

Table 3. Zero-shot results on the SugarCrepe++ compositionality benchmark.

Method	Params	Swap Object		Swap Attribute		Replace Object		Replace Attribute		Replace Relation	
	(B)	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT	ITT	TOT
Qwen2-VL-2B [14]	2.21	32.7	27.8	30.5	25.3	73.6	65.9	46.8	43.0	57.6	**58.3**
VladVA (Ours) (`Qwen2-VL-2B`)	2.21	**50.8**	**33.5**	**60.4**	**48.2**	**93.7**	**93.8**	**74.8**	**77.5**	**63.6**	57.4
LLaVA-1.5-7B [8]	7.06	23.8	30.7	28.0	29.5	58.1	63.0	46.8	58.1	52.3	63.4
VladVA (Ours) (`LLaVA-1.5-7B`)	7.06	**56.1**	36.7	**63.0**	**62.5**	**95.0**	93.0	**78.2**	82.3	71.1	66.3
LLaVA-1.5-13B [8]	13.35	35.5	32.3	30.2	32.4	68.7	66.8	44.8	43.1	52.3	55.6
VladVA (Ours) (`LLaVA-1.5-13B`)	13.35	**55.2**	**38.3**	**65.6**	**60.6**	**94.5**	**92.5**	**80.7**	**81.1**	**73.2**	**66.4**

approach matches and outperforms prior models, including the large 18B EVA-CLIP model (17.5 vs. 15.0, 40.5 vs. 35.8 and 12.8 vs. 10.5, for image, text and respectively group set).

Table 4. Comparison with state-of-the-art on the Winoground compositionality benchmark.

Model	Image	Text	Group
CLIP (ViT-B) [9]	10.5	25.0	7.3
CLIP (ViT-L) [9]	12.3	27.5	8.3
BLIP (ViT-L) [6]	10.0	30.5	7.8
BLIP2 (ViT-L) [7]	10.5	29.5	8.5
OpenCLIP (ViT-G/14) [10]	12.8	32.0	9.3
OpenCLIP (ViT-BigG/14) [10]	15.5	35.5	12.0
EVA-02-CLIP (ViT-E/14+) [11]	14.0	33.8	10.8
EVA-CLIP (8B) [12]	14.8	36.5	10.3
EVA-CLIP (18B) [12]	15.0	35.8	10.5
NegCLIP [15]	10.5	29.5	8.0
LLaVA-1.5-7B [8]	11.3	18.5	6.5
E5-V (LLaVA-Next-8B) [4]	14.8	32.3	11.3
E5-V (LLaVA-1.5-7B) [4]	17.4	31.3	10.5
VladVA (Ours) (LLaVA-1.5-7B)	**17.5**	**40.5**	**12.8**

3. Zero-shot image recognition on ImageNet

Table 5. Zero-shot image recognition results on ImageNet dataset in terms of Top-1 and Top-5 (%) accuracy.

Model	Data. size	Top-1	Top-5
CLIP (ViT-B) [9]	400M	68.4	91.9
CLIP (ViT-L) [9]	400M	74.0	94.0
EVA-CLIP (18B) [12]	2.7B	83.5	97.2
CLIP (ViT-B) [9]	15M	32.8	-
HiDeCLIP (ViT-B) [9]	15M	45.9	-
FFF (ViT-B) [1]	15M	51.1	-
BLIP (ViT-L) [6]	129M	54.2	81.5
BLIP2 (ViT-L) [7]	129M	46.7	74.2
LLaVA-Next-8B [5]	0M	45.8	74.6
E5-V [4] (LLaVA-Next-8B)	0M	48.2	76.6
LLaVA-1.5-7B [8]	0M	42.0	74.6
VladVA (Ours) (LLaVA-1.5-7B)	8.1M	63.7	88.3
Qwen2-VL-2B [14]	0M	54.7	79.4
VladVA (Ours) (Qwen2-VL-2B)	8.1M	70.6	91.1

From an evaluation point of view, the main focus of this work is on improved zero-shot retrieval and, more generally, improved vision-language compositional ability. We focus on these tasks, as they require stronger (vision-)language understanding abilities, which we show an LVLM can offer under appropriate training regimes. As a study case, herein, for completeness, we also measure the zero-shot ability of the model for image recognition on ImageNet [2]. As the results from Table 5 show, our approach significantly improves upon the zero-shot LVLM we start from (54.7 vs 70.6%). In comparison, E5-V approach only offers modest performance gains (45.8 vs 48.2%) and has notably lower performance than our approach (48.2 vs 70.6%) despite using a bigger model. While significantly improving upon the model we start from, the low data regime we train our model in (only 8.1M samples) limits its overall performance, with contrastive models trained on billion samples performing better. This is expected as the image recognition ability of a model, especially on the highly specific categories of ImageNet, will depend on how often (if at all) they are seen in the training set. This is especially significant given that many of the datasets used for contrastive learning are filtered based on the ImageNet classes [9]. In lower data regimes, comparable with ours, we can observe that our approach produces notably better results (*e.g.* 51.1% for FFF [1], trained on 15M samples vs 70.6% for ours). Finally, when comparing it with other models focusing on retrieval (*i.e.* BLIP and BLIP2) our approach outperforms either of them by more than 15% in absolute terms despite the fact that these models were trained on 129M samples. All in all, we outperform all models trained in comparable settings, showing promising initial results in this direction too.

4. Which layer to choose the token from?

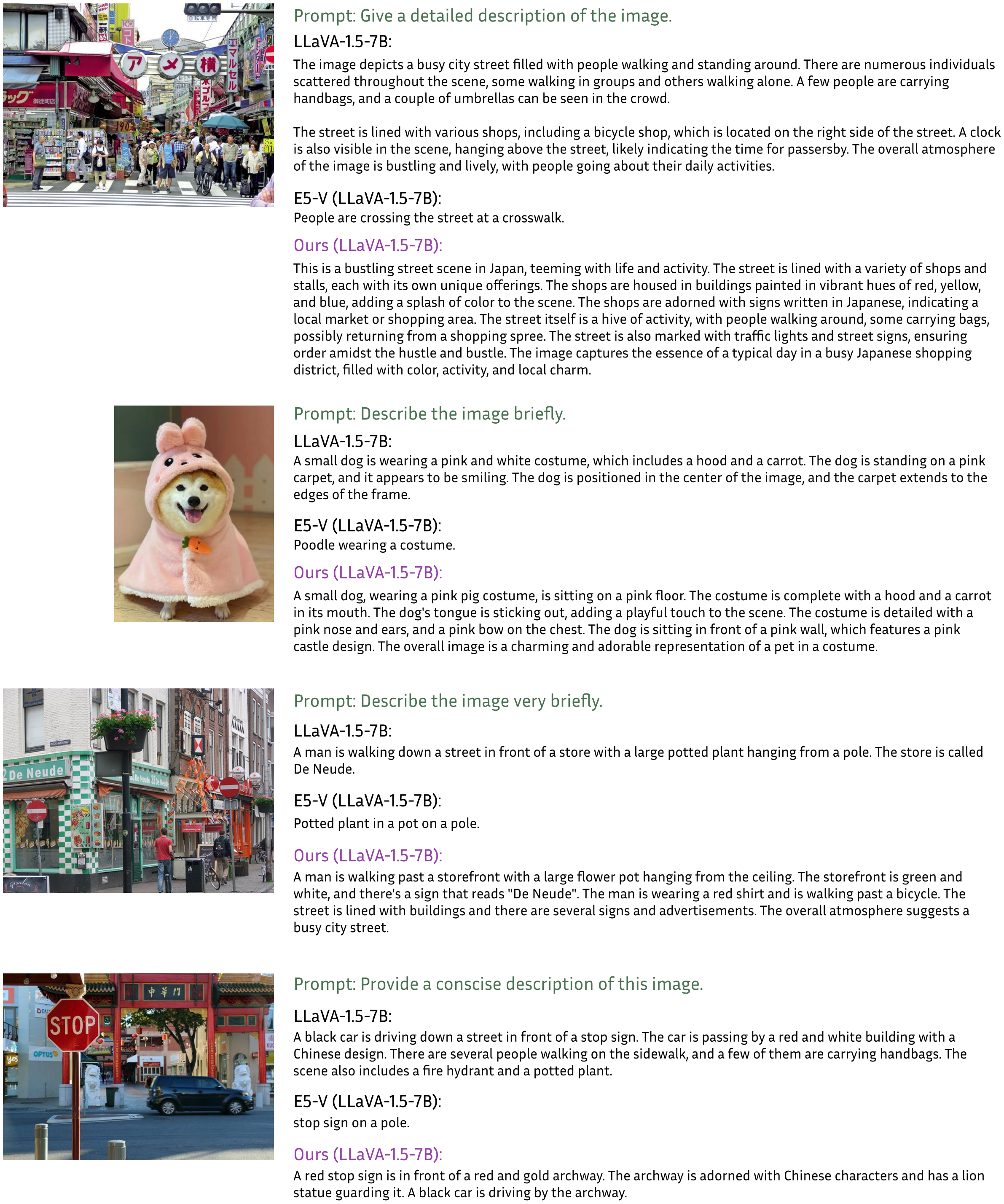
In the main paper, we've used the last token of the last layer as the summary, discriminative token. Intuitively, by selecting the last layer, we maximize the amount of parameters we can adapt, and hence adaptation plasticity. However, herein, for completeness, we report results for different layer IDs in Table 6. The results show that the last 3-4 layers have comparable performance, performance that degrades as we select earlier layers.

Table 6. Performance change when using different layer IDs, reported on SugarCrepe (averaged) and Flickr30k (I2T).

Dataset/Layer	32 (last)	31	28	24	20	16
Flickr30K	0	+0.3	-0.2	-1.1	-13.5	-59.0
SugarCrepe:	0	+0.9	+0.4	-0.1	-8.1	-20.0

5. Qualitative text generation examples post discriminative adaptation

Our main objective is to convert generative LVLMs into discriminative ones, hence the proposed approach is designed from the perspective of maximizing the discriminative abilities of the model. Still, it may be interesting to qualitatively see how our model, and the closest relevant approach E5-V behave. We note, that in principle both our approach and E5-V use LoRAs adapters, hence it is easy to switch between the discriminative and the generative mode without compromising either, by enabling or disabling the adapters. That being said, herein we present some qualitative examples post-training, so we can see the direct effect the training has on the model. As the results from Fig.1 show, generally, our approach better retains the generative capabilities of the model post-training, producing fine-grained captions, similar with the original ones. In contrast, E5-V appears to predominantly produce only very-shot, not-descriptive outputs.



Prompt: Give a detailed description of the image.

LLaVA-1.5-7B:

The image depicts a busy city street filled with people walking and standing around. There are numerous individuals scattered throughout the scene, some walking in groups and others walking alone. A few people are carrying handbags, and a couple of umbrellas can be seen in the crowd.

The street is lined with various shops, including a bicycle shop, which is located on the right side of the street. A clock is also visible in the scene, hanging above the street, likely indicating the time for passersby. The overall atmosphere of the image is bustling and lively, with people going about their daily activities.

E5-V (LLaVA-1.5-7B):

People are crossing the street at a crosswalk.

Ours (LLaVA-1.5-7B):

This is a bustling street scene in Japan, teeming with life and activity. The street is lined with a variety of shops and stalls, each with its own unique offerings. The shops are housed in buildings painted in vibrant hues of red, yellow, and blue, adding a splash of color to the scene. The shops are adorned with signs written in Japanese, indicating a local market or shopping area. The street itself is a hive of activity, with people walking around, some carrying bags, possibly returning from a shopping spree. The street is also marked with traffic lights and street signs, ensuring order amidst the hustle and bustle. The image captures the essence of a typical day in a busy Japanese shopping district, filled with color, activity, and local charm.

Prompt: Describe the image briefly.

LLaVA-1.5-7B:

A small dog is wearing a pink and white costume, which includes a hood and a carrot. The dog is standing on a pink carpet, and it appears to be smiling. The dog is positioned in the center of the image, and the carpet extends to the edges of the frame.

E5-V (LLaVA-1.5-7B):

Poodle wearing a costume.

Ours (LLaVA-1.5-7B):

A small dog, wearing a pink pig costume, is sitting on a pink floor. The costume is complete with a hood and a carrot in its mouth. The dog's tongue is sticking out, adding a playful touch to the scene. The costume is detailed with a pink nose and ears, and a pink bow on the chest. The dog is sitting in front of a pink wall, which features a pink castle design. The overall image is a charming and adorable representation of a pet in a costume.

Prompt: Describe the image very briefly.

LLaVA-1.5-7B:

A man is walking down a street in front of a store with a large potted plant hanging from a pole. The store is called De Neude.

E5-V (LLaVA-1.5-7B):

Potted plant in a pot on a pole.

Ours (LLaVA-1.5-7B):

A man is walking past a storefront with a large flower pot hanging from the ceiling. The storefront is green and white, and there's a sign that reads "De Neude". The man is wearing a red shirt and is walking past a bicycle. The street is lined with buildings and there are several signs and advertisements. The overall atmosphere suggests a busy city street.

Prompt: Provide a conscise description of this image.

LLaVA-1.5-7B:

A black car is driving down a street in front of a stop sign. The car is passing by a red and white building with a Chinese design. There are several people walking on the sidewalk, and a few of them are carrying handbags. The scene also includes a fire hydrant and a potted plant.

E5-V (LLaVA-1.5-7B):

stop sign on a pole.

Ours (LLaVA-1.5-7B):

A red stop sign is in front of a red and gold archway. The archway is adorned with Chinese characters and has a lion statue guarding it. A black car is driving by the archway.

Figure 1. **Qualitative comparison on image captioning** of the base LLaVA-1.5-7B model and its fine-tuned versions using both E5-V [4] and our proposed method. We show that with our method, the LLaVA-1.5-7B better retains its captioning capabilities, while E5-V fine-tuning appears to result in less informative captions.

References

[1] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182, 2024. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[3] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022. 1

[4] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 2, 4

[5] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 2

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[7] 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. 2

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2

[9] 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. 2

[10] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[11] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[12] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 2

[13] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[14] 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. 1, 2

[15] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 1, 2