

## Appendix

### A. Details of ELVA Model

#### A.1. Training Details

Tab 1 summarizes the hyperparameters used across different training stages. During the spatial pretraining stage, we adopt a low number of frames, increasing to 32 frames for both the spatial-temporal pretraining and supervised fine-tuning (SFT) stages.

Table 1. **Hyper-parameter Settings for Training Details.** PE denotes Patch Embedding, TC represents the Temporal-Capture Block, TH refers to the Task Head, and LM indicates the language model.

Hyperparameter	Stage 1	Stage 2	Stage 3
Data Scale	4M	3M	843K / 3M
Batch Size	256	256	128 / 256
Video Frame	1	16	32
Hierarchical Merge	$\times$	$\times$	$\checkmark$
Learning Rate (lr)	4e-5	4e-5	2e-5
LR Schedule	cosine decay	cosine decay	cosine decay
LR Warmup Ratio	0.03	0.01	0.01
Epoch	1	2	1
Weight Decay		0	
Optimizer		AdamW	
DeepSpeed stage		2	

We utilize a total of 4M image samples, comprising 1M from Densefusion and 3M from re-annotated CC3M and COCO in stage 1. For stage 2, we employ 3M re-annotated samples, including 2M from WebVid [2] and 1M from VALOR [4]. See Table 2 for a detailed breakdown of data sources.

#### A.2. Prompt Engineering

We utilize the following prompt in table 3 to generate detailed captions for the provided images and videos using Qwen2-VL (7B) [14]. For image data, we limit the maximum pixel count to  $1280 \times 28 \times 28$  to ensure computational efficiency. Using 16 Nvidia A100 GPUs, generating 3 million high-quality image descriptions takes approximately two days. For video data, we process frames at a rate of 1 fps, with the maximum pixel count per frame set to  $360 \times 420$ . Generating 3 million video captions under these settings requires three days with 16 Nvidia A100 GPUs.

### B. Evaluations on Image-Language Benchmarks.

**Evaluation on Image-Language Benchmarks.** We evaluate ELVA on a series of general visual understanding benchmarks including GQA [9], SEED-Bench [10], MME [8], MMBench [12]. Part of the image results of Chameleon

Table 2. Data used in pre-training and multimodal supervised fine-tuning stages. \* indicates the data is used only in ELVA-7B (HD).

Stage	Dataset	Scale	Source
Stage 1	ELVA-Image	3M 1M	CC-3M, COCO DenseFusion
Stage 2	ELVA-Video	3M	Webvid-2.5M, VALOR-1M
Stage 3	LLaVA-Video	178K	NeXT-QA, ActivityNetQA, PerceptionTest, LLaVA-Hound
	LLaVA-665K /	665K /	COCO, VG, OCR-VQA, GQA, TextVQA
	LLaVA-OneVision*	3M*	High-Quality Single-Image

Image Prompt	Video Prompt
<p>You are a powerful multimodal model and you should generate detailed descriptions of this image, including information such as [World Knowledge], [Objects], and [OCR]. Although the information may contain errors or be incomplete, you should disregard any inaccuracies. If any information is not used, do not specify why.</p> <p><b>[Additional Information]:</b></p> <ul style="list-style-type: none"> <li>• <b>[World Knowledge]:</b> {SHORT CAPTION}</li> <li>• <b>[Objects]:</b> <ul style="list-style-type: none"> <li>- {OBJECT AA}</li> <li>- {OBJECT BB}</li> <li>- ...</li> </ul> </li> <li>• <b>[OCR]:</b> <ul style="list-style-type: none"> <li>- {SENTENCE A}</li> <li>- {SENTENCE B}</li> <li>- ...</li> </ul> </li> </ul> <p><b>[IMAGE]:</b></p>	<p>You are a powerful multimodal expert in understanding scene transitions based on visual features in a video. You are requested to create the descriptions for the current clip sent to you, which includes multiple sequential frames.</p> <p><b>[Guidelines For Clip Description]:</b></p> <ul style="list-style-type: none"> <li>• Analyze the narrative progression implied by the sequence of frames, interpreting the sequence as a whole.</li> <li>• Note that since these frames are extracted from a clip, adjacent frames may show minimal differences. These should not be interpreted as special effects in the clip.</li> <li>• When referring to people, use their characteristics, such as clothing, to distinguish different people.</li> <li>• <b>***IMPORTANT**</b> Please provide as many details as possible in your description, including colors, shapes, and textures of objects, actions and characteristics of humans, as well as scenes and backgrounds.</li> </ul> <p><b>[VIDEO]:</b></p>

Table 3. Prompt for Caption Engine.

Table 4. **Evaluation on image-language benchmarks.** Our evaluation involves GQA, MME, MM Bench (MMB), and SEED. The results in **bold** and underline are the best and second-best results among encoder-free models, respectively.

Model	Data	GQA	SEED <sub>1</sub>	MME	MMB
<b>Encoder-based Models</b>					
Qwen-VL [1]	7.2B	57.5	58.2	1487.5	60.6
LLaVA-v1.5 [11]	0.4B+	62.0	58.6	1510.7	64.3
LLaVA-1.6 (HD) [11]	0.4B+	64.2	64.7	1519.3	67.4
<b>Encoder-free Models</b>					
Fuyu-8B [3]	-	-	-	-	10.7
Chameleon [13]	1.4B+	-	30.6	170	31.1
EVE [6]	33M	60.8	61.3	1217.3	49.5
SOLO [5]	43.7M	-	64.4	1260	-
EVE (HD) [6]	33M	<b>62.6</b>	64.6	<b>1305.7</b>	52.3
<b>ELVA</b>	7M	60.5	<u>66.5</u>	1262.1	<u>53.8</u>
<b>ELVA (HD)</b>	7M	<u>61.1</u>	<b>67.2</b>	<u>1291.5</u>	<b>58.2</b>

and EVE are evaluated with VLMEvalKit [7] or from the OpenCompass. Following the EVE approach, we introduce a high-resolution training strategy exclusively during the SFT stage to develop the ELVA-7B (HD) model, where the longest image edge is set to 1,344 pixels.

#### B.1. Visualization on Data

We provide additional examples from the ELVA datasets in Table 5. For enhanced visualization, different colors are used to highlight distinct types of information within the descriptions.

## 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*, 2023. [1](#)
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. [1](#)
- [3] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşirlar. Introducing our multimodal models, 2023. [1](#)
- [4] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Wein-ing Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. [1](#)
- [5] Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. Solo: A single transformer for scalable vision-language modeling, 2024. [1](#)
- [6] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024. [1](#)
- [7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. [1](#)
- [8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. [1](#)
- [9] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. [1](#)
- [10] 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](#)
- [11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. [1](#)
- [12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. [1](#)
- [13] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. [1](#)
- [14] 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. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. [1](#)

Table 5. Visualizations of the descriptions in ELVA-Image and ELVA-Video. For enhanced clarity, information related to *objects/attributes*, *spatial positions/scene changes*, and *text information* is highlighted using distinct colors.

Visualization	Detailed Caption
	The image depicts a <b>red and black train</b> , stationed at a <b>platform</b> . The train is numbered <b>96104</b> and is positioned on the tracks, ready to depart or having just arrived. <b>The platform</b> is equipped with a <b>yellow safety line</b> and a <b>metal fence on the right side</b> . <b>In the background</b> , there is a <b>pedestrian bridge</b> and <b>some buildings</b> , indicating an urban setting. The weather appears to be clear, and the overall scene is well-lit.
	The image depicts a <b>cozy living room</b> with a modern and somewhat industrial aesthetic. The room features <b>two black leather sofas</b> , one of which is positioned <b>against the wall</b> and the other <b>facing the television</b> . The sofas are arranged around a <b>wooden coffee table</b> , which has a few items on it, including a <b>small plant</b> and a <b>cup</b> . ... <b>On the right side of the room</b> , there is a <b>flat-screen television</b> mounted on the wall. <b>Below the television</b> , there is a <b>small shelf</b> with a few items on it. <b>In the background</b> , there is a <b>poster</b> on the wall with the text <b>"ROXY PRO FIJI"</b> visible, suggesting a connection to surfing or a surfing event. The room also contains a <b>guitar leaning against the wall</b> , adding a personal touch to the space.
	The video <b>begins with a title screen</b> that reads <b>"A LIFESTYLE."</b> The scene then shifts to a <b>man wearing a white t-shirt, black cap, and sunglasses</b> , standing in front of a <b>car</b> at night. The man appears to be talking or speaking, and the background shows a <b>brightly lit gas station with an American flag</b> and a <b>sign that reads "2 for \$2.22."</b> The scene suggests a casual, relaxed atmosphere, possibly indicating a lifestyle that involves leisure activities and spending time outdoors.
	The video <b>opens with a close-up shot of a white fireplace with a warm fire burning inside</b> . The flames flicker and dance, casting a cozy glow on the surrounding area. <b>As the camera pans out</b> , we see that the fireplace is situated in a <b>well-decorated room</b> , with a potted plant and a <b>few decorative items</b> placed on the <b>mantel</b> . The walls are painted a warm beige color, and the <b>floor</b> is covered with a <b>soft, plush carpet</b> . The overall atmosphere is one of warmth and comfort, with the fire providing a focal point for the room.