PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models

Supplementary Material

A. Implementation Details

Our implementation is based on InternVL2 [83]. The visual encoder is a ViT-L/14 [20] whose weights are initialized from InternViT-300M-448px-V1.5 [15]. The progressive encoding is introduced in the last 8 layers of the 24-layer ViT (*i.e.*, $\tilde{L} = 8$ and $L = 24$), and the weights of temporal attention (T-MHA) are randomly initialized. The input resolution for the visual encoder is set to 448×448, with dynamic resolution [15] enabled for image data in both stages. The number of image tiles ranges up to 12, based on the image's aspect ratio and resolution. InternLM2-Chat-1.8B [7] is employed as the LLM for $\text{PVC}_{\text{InternVL2}}$-2B, while InternLM2.5-Chat-7B [7] is utilized for $\text{PVC}_{\text{InternVL2}}$-8B.

Training settings of the pre-training and instruction tuning stages are listed in Tab. 6.

Table 6. **Training settings of PVC.**

Training stage	Pre-training	Instruction tuning
Max sequence length	8192	
Max tile/image	12	
Token/frame (tile)	64	
Number of image repeats	4	
Number of video frames	16-96, uniform	
Optimizer	AdamW	
Learning rate	2e-4	4e-5
Weight decay	0.01 (2B), 0.05 (8B)	
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
Learning rate schedule	constant with warmup	cosine decay
Warmup steps	100	240
Training steps	25k	8k
Batch size	2048	1024

B. Dataset Details

The data used in the pre-training stage are listed in Tab 7. All image-text data is adopted from InternVL2 [83]. For video datasets, we initially utilize large-scale but mixed-quality datasets including InternVid-10M [90], WebVid-10M [3], TextVR [93] and OpenVid-1M [70] that primarily feature short video captions. To enhance the model's understanding of visual details, we further incorporate densely captioned video-text datasets of varied video lengths including MiraData [35], ShareGPT4Video [14] and the Vript series [94]. To improve capabilities in multi-turn conversation and visual reasoning, we employ VideoGPT+ [62], STAR [91], EgoTaskQA [33], CLEVRER [95] and Mementos [89]. Additionally, NTU RGB+D [78] is used to boost robustness to action recognition. Lastly, to enhance the model's holistic abilities, we utilize comprehensive datasets VideoChat2-IT [49] and FineVideo [21], which aggregate elements of multiple-choice answering, open-ended question-answering, and conversations.

Table 7. **Summary of datasets used in the pre-training stage.** *IT refers to the instruction tuning data in VideoChat2.

task	dataset
Short Caption	Laion (en&zh) [74], COYO [6], COCO [51]
OCR	Wukong-OCR [23], LaionCOCO-OCR [75]
Detection	GRIT [72], Objects365 [79]
Conversation	All-Seeing (en&zh) [87]
Image-text instruction data (see Tab. 8a)	

(a) Image-text datasets used in the pre-training stage.

task	dataset
Short Caption	InternVid-10M [90], WebVid [3], OpenVid [70], TextVR [93]
Detailed Caption	MiraData [35], ShareGPT4Video [14], Vript [94], Vript_Chinese [94], LSMDC [73]
VQA	STAR [91], VideoGPT+ [62], EgoTaskQA [33] CLEVRER [95], Mementos [89]
Classification	NTU RGB+D [78]
Comprehensive	VideoChat2-IT* [49], FineVideo [21]

(b) Video-text datasets used in the pre-training stage.

Datasets used for instruction tuning are listed in Tab 8. The image-text data is also adopted from InternVL2 [83]. For video-text data, low-quality datasets used in the pre-training stage are replaced by compositional high-quality datasets like LSMDC[73], TVQA [45], HiREST [99] and LLaVA-Video [103].

In Tab. 9, we list the training data scale of our PVC and some existing image-video general VLMs. Compared to the previous state-of-the-art model, Qwen2-VL, we use significantly less data while achieving similar or better performance on most benchmarks. Compared to other VLMs, our PVC requires more pre-training data since the temporal attention layers in ViT introduce new parameters that are randomly initialized.

C. Appended Ablation Studies

C.1. Number of Tokens per Frame

We adjust token-per-frame while keeping the total number of visual tokens fixed. Reducing tokens per frame increases image repetitions and video input frames. Compared to 256 tokens, 64 tokens perform similarly on image tasks but significantly better on long video tasks (*e.g.*,

Table 8. **Summary of datasets used in the instruction tuning stage.** *IT refers to the instruction tuning data in VideoChat2.

task	dataset
General QA	VQAv2 [22], GQA [32], OKVQA [64], VSR [54]
Science	AI2D [40], ScienceQA [60], Chemistry Data [48] TQA [41]
Medical	PMC-VQA [101], VQA-RAD [44], VQA-Med [4] Medical-Diff-VQA [29], PathVQA [26], SLAKE [52], PMC-CaseReport [92]
Chart	ChartQA [66], LRV-Instruction [55], PlotQA [69] Unichart [67], MMC-Inst [56], DVQA [36] TableMWP [61], FigureQA [37], MapQA [10] SciTSR [16], Fintabnet [104]
Mathematics	CLEVR [34], MetaMath [96], GeoQA+ [8] Geometry3k [59], GeoS [76], Unigeo [12] Super-CLEVR [50], MathQA [1]
Knowledge	Art500k [63], MovieNet [30], KonIQ-10k [27] KVQA [77], ViQuAE [46]
OCR	InfoVQA [68], TextVQA [80], ArT [17] CASIA [53], Chart-to-text [38], COCO-text [84] CTW [97], EATEN [24], ICDAR2019-LSVT [82] ICPR MTWI [25], NAF [19], ReCTS [100] TextOCR [81], LLaVAR [102], HME-100k [98] POIE [42], SROIE [31], ST-VQA [5] EST-VQA [88], IAM [65]
Document	DocVQA [18], DocReason25k [28]
Grounding	RefCOCO [39], RefCOCO+ [39], RefCOCOg [39] RD-BoxCoT [13]
Conversation	ALLaVA [11], LAION-GPT4V [43] MMDU [58], TextOCR-GPT4V [9]
Detection	Objects365 [79], V3Det [85]

(a) Image-text datasets used in the instruction tuning stage.

task	dataset
Detailed Caption	ShareGPT4Video (en&zh) [14], Vript_Chinese [94] Vript [94], LSMDC [73]
VQA	STAR [91], EgoTaskQA [33], Mementos [89] TVQA [45], HiREST [99], PerceptionTest [71]
Classification	NTU RGB+D [78]
Comprehensive	VideoChat2-IT* [49], LLaVA-Video [103]

(b) Video-text datasets used in the instruction tuning stage.

Table 9. **Training data scale of our PVC and existing VLMs.**

Model	# Pre-training samples	# SFT samples
Oryx-MLLM	4.5M	1.3M
LLaVA-OneVision	4.5M	4.8M
InternVL2	45M	7.3M
Qwen2-VL	1.4T tokens	unknown
PVC (ours)	50M (102B tokens)	8.2M (33B tokens)

VideoMME), with a 16% speed decrease. Compressing to 16 tokens performs worse and requires much more computation with more input frames. Therefore, to balance performance and efficiency, we choose 64 tokens per frame as the final setting.

C.2. Number of Temporal Attention Layers

As described in Sec.3.2, we add temporal attention to the last $\tilde{L}$ layers of the ViT. As shown in Tab. 11, $\tilde{L}=8$ performs better than $\tilde{L}=1$ or $\tilde{L}=4$. Setting $\tilde{L}=24$ (adding temporal attention to each layer) does not provide significant improvements over $\tilde{L}=8$ but increases computational overhead. Therefore, we choose $\tilde{L}=8$, *i.e.*, adding temporal attention to the last 8 layers of the ViT.

Table 10. **Ablation of the number of tokens per frame.** The ablation is conducted on 2B model with shortened pre-training on 10M samples. FPS is tested with the same setting as Tab. 5.

# token /frame	# image repeat.	# video frame.	Info VQA	MMB	MVBench	Video MME	FPS
256	1	16	**58.7**	74.8	62.1	45.0	**12.7**
64	4	64	58.5	**74.9**	**62.4**	**46.7**	10.6
16	16	256	53.3	74.9	61.8	46.4	7.2

Table 11. **Ablation of the number of temporal attention layers in ViT.** The ablation study is conducted on 2B model with shortened pre-training on 10M samples.

# temp. attn. layer ($\tilde{L}$)	MVBench	VideoMME	InfoVQA	MMB
1	61.6	46.3	55.8	74.2
4	62.1	46.6	57.3	74.7
8	62.4	46.7	58.5	74.9
24	62.5	46.7	58.8	74.8

C.3. Training Strategy

During the pre-training phase, we unfreeze the parameters of the ViT and LLM, which differs from existing methods [2, 15, 47, 57, 83, 86]. The ablation results in the Tab. 12 empirically explain why we adopt this training strategy. For InternVL2, keeping ViT and LLM fixed or trainable during pre-training has minimal impact on the final performance. However, for our PVC, unfreezing the ViT and LLM during pre-training leads to significantly better performances. We suppose this improvement is due to the following reasons: (1) Training the ViT jointly with the newly added progressive encoding module enables better capture of complementary information and minimizes redundancy. (2) The LLM's inherent capability to process multi-frame videos is limited, especially for integrating different information extracted from the repeated frames of an image. Thus, additional training is needed for effective adaptation.

D. Qualitative Results

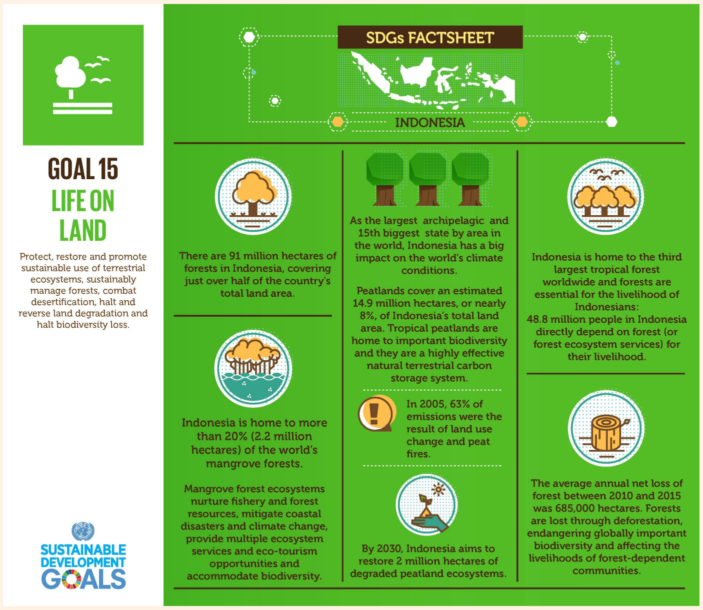
Image Progressive Encoding. As shown in Fig. 4, when an image is repeated as multiple frames, our PVC can extract more precise details and supplementary information from the subsequent frames. For instance, the model extracts the title “Goal 15 life on land” from the first frame and corresponding detailed goals from subsequent frames. Incorrect contents, *e.g.* “10% of the total land area”, are also corrected using the information extracted from subsequent frames.

Table 12. **Ablation of the training strategy**. The ablation study is conducted on 2B model with shortened pre-training on 10M samples. “Unfreeze ViT & LLM” means unfreezing the parameters of the ViT and LLM during the pre-training stage with a lower learning rate ($\times 0.1$). * InternVL2 model is trained under our ablation setting.

Model	Unfreeze ViT & LLM	MVBench	VideoMME	MLVU	DocVQA	InfoVQA	MMB
InternVL2*		60.9	44.7	50.0	86.1	58.7	74.2
InternVL2*	✓	61.9	44.9	50.5	86.4	59.1	75.1
$\text{PVC}_{\text{InternVL2}}$		60.6	45.5	53.2	84.8	57.3	74.4
$\text{PVC}_{\text{InternVL2}}$	✓	62.4	46.7	55.1	86.7	58.5	74.9

Video Dynamic Capture. As shown in Fig. 5, the video contains several consecutive collisions, testing the model's ability to capture spatiotemporal dynamics. InternVL2 [83] describes each collision moment independently but lacks a description of the whole process and includes inaccuracies. LLaVA-OneVision [47] and Qwen2-VL [86] exhibit numerous errors when capturing object interactions. In contrast, our PVC not only accurately describes the process of each collision and the associated object movements but also identifies the causal relationships between multiple collisions.

Image Progressive Encoding



User: What is the content in this picture?

PVC (image repeat × 1):

The image is an infographic related to Sustainable Development Goal 15, which focuses on Life on Land. It highlights the importance of forests in Indonesia and the challenges they face. Here is a detailed breakdown of the content:

Title and Header
- SDGs FACTSHEET
- INDONESIA
- GOAL 15
- LIFE ON LAND

Key Points
1. Forest Area
- There are 91 million hectares of forests in Indonesia, covering just over 10% of the total land area.

2. Archipelagic and State Forests
- Indonesia is home to the third largest tropical forests, which are essential for the livelihood of 48.8 million Indonesians. These forests cover 14.9 million hectares, or nearly 8% of Indonesia's total land area.

3. Biodiversity
- Tropical forests are home to important biodiversity, providing a highly effective natural terrestrial carbon storage system.

4. Mangrove Forests
- Mangrove forests provide multiple ecosystem services, including fishery resources, coastal protection, and biodiversity support.

5. Land Use and Emissions
- In 2005, 63% of emissions were the result of land use change and peatland degradation. Mangrove forests are also threatened.

6. Deforestation and Loss
- The average annual net loss of forest is 685,000 hectares. Forests are lost through deforestation, biodiversity loss, and affecting the livelihoods of forest-dependent communities.

7. Future Goals
- By 2030, Indonesia aims to restore 2 million hectares of degraded peatland ecosystems.

Visual Elements
- **Icons:** The infographic uses icons of trees, forests, and other related imagery to visually represent the information.
- **Colors:** Predominantly green, symbolizing nature and forests, with white and brown accents for text and details.

Conclusion
The infographic emphasizes the critical role of forests in Indonesia, the threats they face, and the country's efforts to address these issues through sustainable practices and restoration goals.

PVC (image repeat × 4):

The image is an infographic related to Goal 15 of the Sustainable Development Goals (SDGs), which focuses on Life on Land. The infographic provides information about Indonesia's forests and peatlands, emphasizing the importance of forest conservation and restoration. Here is a detailed breakdown of the content:

Title and Goal
- SDGs FACTSHEET
- INDONESIA
- Goal 15: Life on Land
  - Protect, restore, and promote the sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, halt and reverse land degradation, and halt biodiversity loss.

Key Points
1. Forest Area
- There are 91 million hectares of forests in Indonesia, covering just over half of the country's total land area.

2. Archipelagic and State by Area
- Indonesia is the largest archipelagic and 15th biggest state by area in the world. It has a significant impact on the world's climate conditions.

3. Dependence on Forests
- Indonesia is home to the third largest tropical forests. 48.8 million people in Indonesia directly depend on forest (or forest ecosystem services) for their livelihood.

4. Peatlands
- Peatlands cover an estimated 14.9 million hectares, or nearly 8%, of Indonesia's total land area. Tropical peatlands are home to important biodiversity and a highly effective natural terrestrial carbon storage system.

5. Mangrove Forests
- Indonesia is home to more than 20% (2.2 million hectares) of the world's mangrove forests. Mangrove forest ecosystems nurture fishery and forest resources, mitigate coastal disasters, and climate change, provide multiple ecosystem services, and eco-tourism opportunities, and accommodate biodiversity.

6. Emissions and Degradation
- In 2005, 63% of emissions were the result of land use change and peat fires.
- The average annual net loss of forest between 2010 and 2015 was 685,000 hectares. Forests are lost through deforestation, endangering globally important biodiversity and affecting the livelihoods of forest-dependent communities.

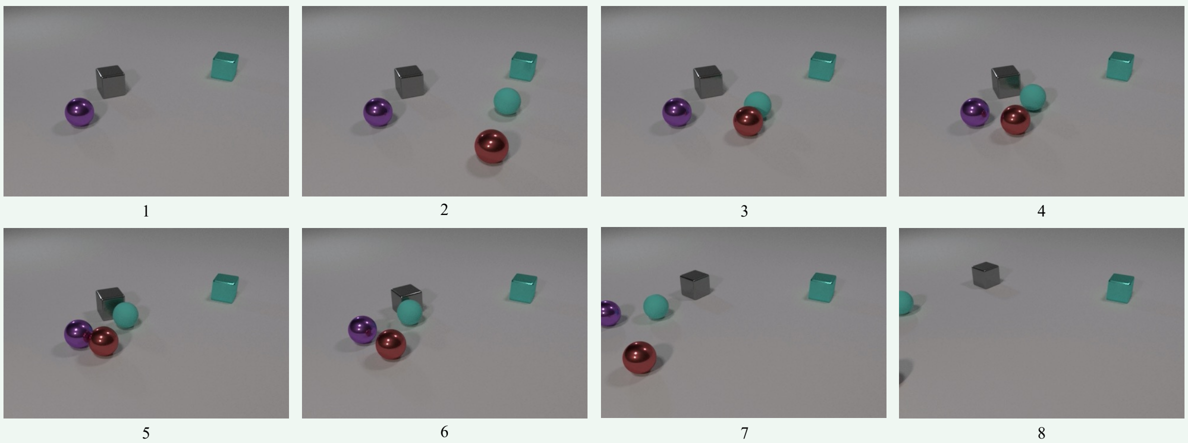
7. Future Goals
- By 2030, Indonesia aims to restore 2 million hectares of degraded peatland ecosystems.

Visual Elements
- The infographic includes icons of trees, forests, and peatlands.
- It uses a green color scheme to emphasize the environmental focus.
- The United Nations Sustainable Development Goals logo is present at the bottom left.

Conclusion
The infographic highlights the critical role of forests and peatlands in Indonesia, emphasizing the need for sustainable forest management and restoration to combat deforestation, land degradation, and biodiversity loss. It underscores the interdependence of local communities on forest ecosystems and the global impact of Indonesia's forests on climate conditions.

Figure 4. **PVC achieves image progressive encoding**. The image is repeated once (left) and four times (right). Supplementary contents are marked in blue, incorrect contents in red, and corrected contents in green.

Video Dynamic Capture



User: Describe the collisions in the video in detail.

InternVL2-8B:
In the video, the purple metal sphere and the red metal sphere collide. The red metal sphere and the cyan rubber sphere collide. The gray metal cube and the cyan rubber sphere collide. The gray metal cube and the red metal sphere collide.

LLaVA-OneVision-7B:
The video shows a sequence of geometric shapes, specifically spheres and cubes, that are arranged on a flat surface. The shapes come in different colors: purple, red, cyan, and gray. The objects move around the surface, with some colliding with each other. Starting from the beginning, there is a single purple sphere visible. As the video progresses, more objects appear, including a red sphere, a gray cube, and a cyan sphere. The objects move independently for a while before they start to interact. One notable collision occurs when the gray cube collides with the cyan sphere. This collision causes both objects to change their positions slightly. Following this, the red sphere moves towards the gray cube and appears to make contact, causing the gray cube to shift position again. The video continues with the red sphere moving away from the gray cube and the cyan sphere remaining stationary. The purple sphere is also visible but does not seem to be involved in any collisions at this point. Towards the end of the video, the red sphere moves closer to the gray cube once more, and it looks like it might collide with it again. However, the red sphere stops short of making contact, leaving the gray cube in its original position.

Qwen2-VL-7B:
The video shows a sequence of collisions between different colored spheres and cubes on a flat surface. Initially, there are three objects: a purple metallic sphere, a gray metallic cube, and a cyan metallic cube. A cyan rubber sphere enters the scene and collides with the purple sphere. The purple sphere then collides with the gray cube. The cyan rubber sphere collides with the red metallic sphere, which then collides with the purple sphere. The purple sphere collides with the gray cube. The cyan rubber sphere collides with the gray cube. The red metallic sphere collides with the gray cube.

PVC-8B (ours):
The video shows a series of interactions between different objects. Initially, there are three objects: a purple sphere, a gray cube, and a cyan cube. The purple sphere and the gray cube are stationary. Then, a red sphere enters the scene and collides with the cyan sphere, causing the cyan sphere to move away. The red sphere continues to move and eventually collides with the purple sphere, causing the purple sphere to move. The gray cube also moves slightly due to the collision with the cyan sphere. Finally, the red sphere and the cyan sphere exit the scene, leaving the gray cube in its new position.

Figure 5. **PVC effectively captures spatiotemporal dynamics in videos**. Correct descriptions of the movements and interactions of the objects are marked in blue, while incorrect descriptions are marked in red. For visualization, we select the above 8 key frames from the video, while the entire video is fed into the models.

References

[1] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2

[3] 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

[4] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019. 2

[5] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 2

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1

[7] 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. 1

[8] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022. 2

[9] Jimmy Carter. Textocr-gpt4v. https://huggingface.co/datasets/jimmycarter/textocr-gpt4v, 2024. 2

[10] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 2

[11] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 2

[12] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 2

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[14] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1, 2

[15] 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. 1, 2

[16] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 2

[17] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pages 1571–1576. IEEE, 2019. 2

[18] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*, 2017. 2

[19] Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. Deep visual template-free form parsing. In *ICDAR*, pages 134–141. IEEE, 2019. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[21] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024. 1

[22] 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 *CVPR*, pages 6904–6913, 2017. 2

[23] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *NeurIPS*, 35: 26418–26431, 2022. 1

[24] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction. In *ICDAR*, pages 254–259. IEEE, 2019. 2

[25] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *ICPR*, pages 7–12. IEEE, 2018. 2

[26] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 2

[27] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep

learning of blind image quality assessment. *IEEE TIP*, 29: 4041–4056, 2020. 2

[28] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 2

[29] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images, 2023. 2

[30] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, pages 709–727. Springer, 2020. 2

[31] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, pages 1516–1520. IEEE, 2019. 2

[32] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 2

[33] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*, 35:3343–3360, 2022. 1, 2

[34] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 2

[35] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024. 1

[36] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 2

[37] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 2

[38] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. 2

[39] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2

[40] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 2

[41] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007, 2017. 2

[42] Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *ICDAR*, pages 36–53. Springer, 2023. 2

[43] LAION. Laion-gpt4v dataset. `https://huggingface.co/datasets/laion/gpt4v-dataset`, 2023. 2

[44] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2

[45] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1, 2

[46] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR*, pages 3108–3120, 2022. 2

[47] 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

[48] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024. 2

[49] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 1, 2

[50] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, pages 14963–14973, 2023. 2

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[52] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ISBI*, pages 1650–1654. IEEE, 2021. 2

[53] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *ICDAR*, pages 37–41. IEEE, 2011. 2

[54] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2

[55] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2023. 2

[56] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 2

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 2

[58] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 2

[59] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 2

[60] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022. 2

[61] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 2

[62] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 1

[63] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *ACM MM*, pages 1183–1191, 2017. 2

[64] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019. 2

[65] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002. 2

[66] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 2

[67] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 2

[68] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 2

[69] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. 2

[70] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 1

[71] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *NeurIPS*, 36, 2024. 2

[72] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1

[73] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015. 1, 2

[74] 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. *NeurIPS*, 35: 25278–25294, 2022. 1

[75] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/, 2022. 1

[76] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP*, pages 1466–1476, 2015. 2

[77] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, pages 8876–8884, 2019. 2

[78] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 1, 2

[79] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 1, 2

[80] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 2

[81] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812, 2021. 2

[82] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *ICDAR*, pages 1557–1562. IEEE, 2019. 2

[83] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 1, 2, 3

[84] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark

for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 2

[85] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, pages 19844–19854, 2023. 2

[86] 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, 3

[87] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 1

[88] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020. 2

[89] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024. 1, 2

[90] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[91] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 1, 2

[92] Chaoyi Wu. Pmc-casereport. https://huggingface.co/datasets/chaoyi-wu/PMC-CaseReport, 2023. 2

[93] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818, 2025. 1

[94] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *arXiv preprint arXiv:2406.06040*, 2024. 1, 2

[95] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 1

[96] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023. 2

[97] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *J. Comput. Sci. Tech.*, 34(3):509–521, 2019. 2

[98] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. *arXiv preprint arXiv:2203.01601*, 2022. 2

[99] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, 2023. 1, 2

[100] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *ICDAR*, pages 1577–1581. IEEE, 2019. 2

[101] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2

[102] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[103] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2

[104] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *WACV*, pages 697–706, 2021. 2