

---

# Eagle2.5: Boosting Long-Context Post-Training for Frontier Vision-Language Models

## Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Training and Inference

### 1.1 Framework

We integrate and develop multiple technologies to optimize long-context training framework, encompassing GPU memory optimization, distributed context parallelism, and video processing acceleration.

- *GPU Memory Optimization.* We integrate Triton-based fused operators replacing PyTorch’s MLP, RMSNorm, and RoPE [114] implementations. We employ operators that fuse linear layers with cross-entropy loss to eliminate intermediate logit storage, and utilize CPU-offloading of hidden states to further reduce GPU memory usage.
- *Distributed Context Parallelism.* Building on USP [25], we adopt a two-layer communication group based on Ulysses and Ring [68]. Rather than using zigzag ring-attention, we implement zigzag Llama3-style [23] context parallelism with all-gather KV to reduce communication latency.
- *Video Decoding Acceleration.* Training often requires sampling specific sparse video frames, which can cause frame seek latency or memory management issues. We optimize this process through rapid video metadata parsing, improving long video decoding while minimizing memory consumption.
- *Inference Acceleration.* We deploy VLLM [52] for model serving and evaluation, significantly reducing memory requirements and accelerating inference speed.

### 1.2 Training Settings

To maximize resource utilization, we use the model weight from the Stage-1.5 training of Eagle-2 and extend long-context training in stage 2. Finally, our training pipeline is as follows Tab. 1.

		Eagle2.5-Stage-1	Eagle2.5-Stage-1.5	Eagle2.5-Stage-2	Eagle2.5-Stage-3	Eagle2.5-Stage-4
Vision	<b>Resolution</b>	$448 \times \{(i, j) \mid i, j \in \mathbb{Z}^+, i \times j \leq 12\}$				
	<b>Tokens</b>					
Data	<b>Dataset</b>	ALLaVA	Rich Diverse Data	Short+Long Data	Short+Long Data	Short+Long Data
	<b>#Samples</b>	1.2M	21.6M	4.6M+4.6M	4.6M+4.6M	4.6M+4.6M
Model	<b>Trainable</b>	MLP Connector	Full Model			
	<b>Qwen2.5-7B</b>	40.0M				
Training	<b>Batch Size</b>	1024	1024	256	128	128
	<b>Learning Rate</b>	$2 \times 10^{-4}$	$2 \times 10^{-5}$			
	<b>Max Length</b>	4096	8192	32768	65536	128K

Table 1: The proposed progressive training settings.

Subset	ANLS Score	Exact Match Score	F1 Score	Metric	Score
Dev	73.8	67.7	74.7	Overall F1	29.4
Test	72.7	63.2	72.3	Overall Acc	27.7

Table 2: Performance on the SlideVQA benchmark. The ANLS Score refers to Approximate Normalized Levenshtein Similarity.

Table 3: Performance on the MMLongBench-Doc.

Category	Dataset
Captioning & Knowledge	ShareGPT4o [97], KVQA [106], Movie-Posters [113], Google-Landmark [132], WikiArt [37], Weather-QA [80], Coco-Colors [31], music-sheet [24], SPARK [143], Image-Textualization [99], SAM-Caption [100], Tmdb-Celeb-10k [3], PixMo [20]
Mathematics	GeoQA+ [10], MathQA [140], CLEVR-Math/Super [65, 63], Geometry3K [74], MAVIS-math-rule-geo [149], MAVIS-math-metagen [149], InterGPS [75], Raven [148], GEOS [105], UniGeo [14]
Science	AI2D [46], ScienceQA [77], TQA [47], PathVQA [33], SciQA [4], <b>Textbooks-QA</b> , VQA-RAD [55], VisualWebInstruct [123]
Chart & Table	ChartQA [84], MMC-Inst [67], DVQA [42], PlotQA [88], LRV-Instruction [66], TabMWP [78], UniChart [85], Vistext [120], TAT-DQA [159], VQAonBD [128], FigureQA [43], Chart2Text [45], RobuT-{Wikisql, SQA, WTQ} [155], MultiHiertt [154]
Naive OCR	SynthDoG [50], MTWI [32], LVST [117], SROIE [36], FUNSD [40], Latex-Formula [96], IAM [83], Handwriting-Latex [2], ArT [17], CTW [145], ReCTs [150], COCO-Text [127], SVRD [142], Hiertext [72], RoadText [124], MapText [62], CAPTCHA [98], Est-VQA [130], HME-100K [118], TAL-OCR-ENG [118], TAL-HW-MATH [118], IMGUR5K [51], ORAND-CAR [21], Invoices-and-Receipts-OCR [94], Chrome-Writing [92], IIIT5k [89], K12-Printing [118], Memotion [102], <b>Arxiv2Markdown</b> , Handwritten-Mathematical-Expression [6], WordArt [134], RenderedText [131], Handwriting-Forms [39]
OCR QA	DocVQA [18], InfoVQA [87], TextVQA [112], ArxivQA [60], ScreencQA [35], DocReason [93], Ureader [138], FinanceQA [116], DocMatrix [56], A-OKVQA [104], Diagram-Image-To-Text [44], MapQA [12], OCRVQA [90], ST-VQA [9], SlideVQA [119], PDF-VQA [22], SQuAD-VQA, VQA-CD [81], Block-Diagram [109], MTVQA [121], ColPali [27], BenthamQA [86]
Grounding & Counting	TallyQA [1], OODVQA [126], RefCOCO+/g (en) [139, 82], GroundUI [156]
General VQA	LLaVA-150K [69], LVIS-Instruct4V [129], ALLaVA [13], Laion-GPT4V [54], LLaVAR [152], SketchyVQA [126], OminiAlign-V [153], VizWiz [30], IDK [11], AlfworldGPT, LNQA [101], Face-Emotion [26], SpatialSense [137], IndoorQA [48], Places365 [158], MMInstruct [70], DriveLM [111], YesBut [95], WildVision [79], LLaVA-Critic-113k [135], RLAI-F-V [141], VQAv2 [29], MMRA [133], KONIQ [34], MMDU [71], Spot-The-Diff [41], Hateful-Memes [49], COCO-QA [103], NLVR [115], Mimic-CGD [57], Datikz [8], Chinese-Meme [19], IconQA [76], Websight [58]
Text-only	Orca [64], Orca-Math [91], OpenCodeInterpreter [157] MathInstruct [146], WizardLM [136], TheoremQA [16], OpenHermes2.5 [122], NuminaMath-CoT [59], Python-Code-25k [28], Infinity-Instruct [7], Python-Code-Instructions-18k-Alpaca [38], Ruozhiba [73], InfinityMATH [147], StepDPO [53], TableLLM [151], UltraInteract-sft [144]

(a) Summary of the collected Eagle 2.5 SFT datasets

Category	Dataset
Captioning & Knowledge	CC3M [108], TextCaps [110], ShareGPT-4V [15], DenseFusion-1M [61]
Grounding & Counting	Object 365 [107]
Text-only	OpenMathInstruct [125]

(b) Summary of the additional Stage 1.5 datasets

Table 4: Dataset used in Eagle 2.5 for Stage 1 and Stage1.5. **Dataset in Magenta** is internal data.

## 2 Additional Benchmarks

The performance of our model is also evaluated on the SlideVQA and MMLongBench-Doc benchmarks. The test results are shown in the Table 2 and 3:

## 3 Training Data

The training of Eagle2.5 is divided into multiple stages, including stage 1 for MLP alignment, and the pretraining stage 1.5, similar to Eagle-2 [5]. It also includes the progressive long-context training proposed by Eagle 2.5. The training data used in these stages are as follows:

- **Eagle2.5-Stage1:** ALLaVA.
- **Eagle2.5-Stage1.5:** Table 6, including Eagle2.5-Image-SFT (Table 6a) and additional pretraining data (Table 6b) in Eagle2.5.
- **Eagle2.5-Stage2:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.

- **Eagle2.5-Stage3:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.
- **Eagle2.5-Stage4:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.

## 4 Eagle-Video-110K

### 4.1 Data Curation Prompts

In this section, we will introduce the prompts utilized for curating dataset Eagle-Video-110K. They consist of prompts for generating captions and textual context anchors, prompts for generating clip-level QA, and prompts for generating video-level QA.

#### 4.1.1 Prompts for Generating Captions and Anchors

```
You are an expert in understanding visual content in video clips. You are requested to create both brief and detailed captions for the current video clip titled "{title}".

#### Guidelines For Brief Caption:
- Create a concise summary (15-30 words) that captures the essential action, setting, and participants
- Focus on the most visually or narratively significant elements of the scene
- Use clear, direct language
- **IMPORTANT** Treat the video as a complete clip rather than a sequence of frames

#### Guidelines For Detailed Caption:
- Begin with a thorough analysis of the visual content shown in the clip
- **IMPORTANT** Pay special attention to the progression of actions and movements:
  * Break down complex actions into their component steps
  * Use transitional words (then, next, afterward, etc.) to show the flow of actions
  * Describe how one action leads to or connects with the next
  * Capture the natural sequence of movements and gestures
- Note that while the clip may be shown as multiple frames, it should be described as a continuous piece of footage
- For text that appears clearly in the clip: describe it in its original language and provide an English translation in parentheses). For example: [book in Chinese] [book]. Additionally, explain the meaning of the text within its context
- **IMPORTANT** If any text is unclear, partially visible, or too blurry to read with confidence, simply mention the presence of text without attempting to specify its content
- When referring to people, use their characteristics, such as clothing, to distinguish different people
- **IMPORTANT** Please provide as many details as possible in your caption, including colors, shapes, and textures of objects, actions and characteristics of humans, as well as scenes and backgrounds
- Consider how the visual content provides context and meaning

Only output your response in the following format without any additional text, explanations or notes:

```json
{"Brief Caption":"concise summary of the video",
 "Detailed Caption":"The clip begins with..., progresses by..., and concludes with..."}
```
```

Here, the brief caption will be used for textual contextual anchors.

#### 4.1.2 Prompts for generating clip-level QA

```
### Task:
```

You are tasked with generating question-answer pairs based on a detailed video caption and given question categories. Try to create challenging questions when possible, but simpler questions are also acceptable when the details are limited.

#### Input Information:

The detailed caption of the video clip is: {caption}

A brief version of the caption is also provided: {brief\_caption}

#### Question Generation Guidelines:

- Read the caption carefully
- Generate a question-answer pair for each category if possible
- Make sure the question-answer pair cannot be fully answered using only the brief caption
- Create two components for each pair:
  - \* A question that is unambiguous within the current clip
  - \* An answer (can be a word, phrase, or sentence)

#### Question Categories:

{question\_type\_pool}

#### Important Criteria:

- Questions should be unambiguous within the current clip
- Create challenging questions when the details allow
- Ensure answers require information beyond what's in the brief caption
- Generate questions for as many categories as possible
- Mark a category as null only if:
  - \* The question-answer would be fully answerable using just the brief caption
  - \* The category cannot be addressed using any information from either caption

Only output your response in the following format without any additional text, explanations or notes:

```
“‘json
{"question_type_1":{"Q":"question specific to this video","A":"answer"} or null,
"question_type_2":{"Q":"question specific to this video","A":"answer"} or null,
...,
"question_type_n":{"Q":"question specific to this video","A":"answer"} or null}“‘
```

46

47 The “question type pool” is defined in 6.2. And the “question\_type\_x”, x in (1, 2, ...n) are the types  
48 selected from the “question type pool”.

### 49 4.1.3 Prompts for generating generating video-level QA

50 #### Task:

```

You are tasked with generating question-answer pairs based on a video caption and given
question categories. Try to create challenging questions when possible, but simpler questions
are also acceptable when the details are limited.

#### Input Information:
The caption of the video clip is: {caption}

#### Question Generation Guidelines:
- Read the caption carefully
- Generate a question-answer pair for each category if possible
- Create two components for each pair:
  * A question that is unambiguous within the current clip
  * An answer (can be a word, phrase, or sentence)

#### Question Categories:
{question_type_pool}

#### Important Criteria:
- Questions should be unambiguous within the current clip
- Create challenging questions when the details allow
- Generate questions for as many categories as possible
- Mark a category as null only if:
  * The category cannot be addressed using any information from the caption

Only output your response in the following format without any additional text, explanations
or notes:

```json
{"question_type_1":{"Q":"question specific to this video","A":"answer"} or null,
"question_type_2":{"Q":"question specific to this video","A":"answer"} or null,
...,
"question_type_n":{"Q":"question specific to this video","A":"answer"} or null}```

```

Here, “caption” is composed of clip-level caption, and its format is:

```

start_1 ~ end_1: caption_1
start_2 ~ end_2: caption_2
...

```

The “start\_x” and “end\_x”, (x in 1, 2, ...) are in seconds.

## 4.2 Question type pool

As shown in table 7, we list the category names and category descriptions in the question type pool. We ask the model to generate qa pairs according to these categories.

## 5 Training Data

The training of Eagle2.5 is divided into multiple stages, including stage 1 for MLP alignment, and the pretraining stage 1.5, similar to Eagle-2 [5]. It also includes the progressive long-context training proposed by Eagle 2.5. The training data used in these stages are as follows:

- **Eagle2.5-Stage1:** ALLaVA.
- **Eagle2.5-Stage1.5:** Table 6, including Eagle2.5-Image-SFT (Table 6a) and additional pretraining data (Table 6b) in Eagle2.5.
- **Eagle2.5-Stage2:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.
- **Eagle2.5-Stage3:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.
- **Eagle2.5-Stage4:** Mixture of short- and long-context data, including Eagle2.5-Image-SFT (Table 6a), Open-Data, and Eagle-Video-110K.

Index	Category	Description
1	object_recognition	Questions about what an object is
2	object_properties	Questions about object properties, such as color, shape, material, texture
3	object_count	Questions about the number of objects
4	object_state	Questions about object states, such as stretched, compressed, cut, stationary
5	object_location	Questions about where an object is located
6	object_presence	Questions about object existence
7	human_attributes	Questions about human attributes, such as height, body type, build
8	human_pose	Questions about human posture
9	human_appearance	Questions about human external appearance, such as clothing and makeup
10	human_identity	Questions about human identity
11	human_cognitive_process	Questions about human mental processes, including intentions, motivations, decision-making rationale, problem-solving approaches, and reasoning methods
12	human_location	Questions about human location
13	human_emotion	Questions about human emotional state
14	scene_description	Questions about overall scene description
15	text_recognition	Questions about text content appearing in the video
16	text_count	Questions about frequency of text appearances in the video
17	text_location	Questions about location of text in the video
18	single_object_event_recognition	Questions about events involving a single object
19	single_object_event_count	Questions about frequency of single-object events
20	single_object_state_change	Questions about changes in single object state
21	single_object_quantity_change	Questions about changes in single object quantity
22	single_object_location_change	Questions about changes in single object location
23	single_object_trajectory	Questions about single object motion trajectory
24	single_object_speed	Questions about single object movement speed
25	single_object_presence_change	Questions about changes in single object presence
26	human_object_interaction_recognition	Questions about types of human-object interaction
27	human_object_interaction_count	Questions about frequency of human-object interactions
28	human_human_interaction_recognition	Questions about types of human-human interaction
29	object_interaction	Questions about objects' states, actions, interactions, changes, identifications (including brands), and how objects affect or interact with other objects
30	abnormal_event_detection	Questions about presence of abnormal events
31	domain_medical	Questions about medical-related professional knowledge
32	domain_education	Questions about education-related professional knowledge
33	domain_sports	Questions about sports-related professional knowledge
34	domain_movies	Questions about movie-related professional knowledge
35	domain_gaming	Questions about gaming-related professional knowledge
36	domain_technology	Questions about technology-related professional knowledge
37	domain_arts	Questions about arts-related professional knowledge
38	video_editing_effects	Questions about video editing effects, including shot transitions, editing effects, transition effects, etc.
39	camera_movement	Questions about camera motion
40	spatial_relationship	Questions about spatial relationships between objects
41	property_comparison	Questions about comparison of multiple object properties
42	quantity_comparison	Questions about comparison of multiple object quantities
43	state_comparison	Questions about comparison of multiple object states
44	human_object_relationship	Questions about human-object relationships
45	human_human_relationship	Questions about human-human relationships
46	scene_sequence	Questions about temporal relationships between scenes
47	event_sequence	Questions about temporal relationships between events
48	event_causality	Questions about causal relationships between events, including both human-initiated actions and their consequences, as well as cause-effect relationships in natural or systematic processes
49	counterfactual_reasoning	Questions about counterfactual reasoning
50	trajectory_tracking	Questions about tracking object or human positions
51	speed_comparison	Questions about speed comparison between multiple objects or humans
52	event_prediction	Questions about future event prediction
53	anomaly_reasoning	Questions about causes of anomalous phenomena
54	planning	Questions about planning for specific tasks
55	navigation	Questions about navigation to destinations
56	human_action	Questions about actions, behaviors, movements or activities performed by humans, including analysis of techniques, efficiency, and patterns of behavior
57	dialogue_content	Questions about spoken dialogue, verbal content, or conversations between characters/people
58	event_summary	Questions about overall event summary
59	object_ordering	Questions about the sequence or order in which objects are placed, arranged, or handled by individuals
60	event_location	Questions about where events or activities take place
61	process_description	Questions about identifying key components, steps, or progression in a process involving objects and/or humans
62	video_topic	Questions about the main subject, focus, or theme covered in the video
63	anomaly_recognition	Questions about identifying and interpreting anomalies

Table 5: Question type categories and their descriptions

Category	Dataset
Captioning & Knowledge	ShareGPT4o [97], KVQA [106], Movie-Posters [113], Google-Landmark [132], WikiArt [37], Weather-QA [80], Coco-Colors [31], music-sheet [24], SPARK [143], Image-Textualization [99], SAM-Caption [100], Tmdb-Celeb-10k [3], PixMo [20]
Mathematics	GeoQA+ [10], MathQA [140], CLEVR-Math/Super [65, 63], Geometry3K [74], MAVIS-math-rule-geo [149], MAVIS-math-metagen [149], InterGPS [75], Raven [148], GEOS [105], UniGeo [14]
Science	AI2D [46], ScienceQA [77], TQA [47], PathVQA [33], SciQA [4], <b>Textbooks-QA</b> , VQA-RAD [55], VisualWebInstruct [123]
Chart & Table	ChartQA [84], MMC-Inst [67], DVQA [42], PlotQA [88], LRV-Instruction [66], TabMWP [78], UniChart [85], Vistext [120], TAT-DQA [159], VQAonBD [128], FigureQA [43], Chart2Text [45], RobuT-{Wikisql, SQA, WTQ} [155], MultiHiertt [154]
Naive OCR	SynthDoG [50], MTWI [32], LVST [117], SROIE [36], FUNSD [40], Latex-Formula [96], IAM [83], Handwriting-Latex [2], ArT [17], CTW [145], ReCTs [150], COCO-Text [127], SVRD [142], Hiertext [72], RoadText [124], MapText [62], CAPTCHA [98], Est-VQA [130], HME-100K [118], TAL-OCR-ENG [118], TAL-HW-MATH [118], IMGUR5K [51], ORAND-CAR [21], Invoices-and-Receipts-OCR [94], Chrome-Writing [92], IIIT5k [89], K12-Printing [118], Memotion [102], <b>Arxiv2Markdown</b> , Handwritten-Mathematical-Expression [6], WordArt [134], RenderedText [131], Handwriting-Forms [39]
OCR QA	DocVQA [18], InfoVQA [87], TextVQA [112], ArxivQA [60], ScreencQA [35], DocReason [93], Ureader [138], FinanceQA [116], DocMatrix [56], A-OKVQA [104], Diagram-Image-To-Text [44], MapQA [12], OCRVQA [90], ST-VQA [9], SlideVQA [119], PDF-VQA [22], SQuAD-VQA, VQA-CD [81], Block-Diagram [109], MTVQA [121], ColPali [27], BenthamQA [86]
Grounding & Counting	TallyQA [1], OODVQA [126], RefCOCO+/g (en) [139, 82], GroundUI [156]
General VQA	LLaVA-150K [69], LVIS-Instruct4V [129], ALLaVA [13], Laion-GPT4V [54], LLAVAR [152], SketchyVQA [126], OminiAlign-V [153], VizWiz [30], IDK [11], AlfworldGPT, LNQA [101], Face-Emotion [26], SpatialSense [137], Indoor-QA [48], Places365 [158], MMInstruct [70], DriveLM [111], YesBut [95], WildVision [79], LLaVA-Critic-113k [135], RLAIIF-V [141], VQAv2 [29], MMRA [133], KONIQ [34], MMDU [71], Spot-The-Diff [41], Hateful-Memes [49], COCO-QA [103], NLVR [115], Mimic-CGD [57], Datikz [8], Chinese-Meme [19], IconQA [76], Websight [58]
Text-only	Orca [64], Orca-Math [91], OpenCodeInterpreter [157] MathInstruct [146], WizardLM [136], TheoremQA [16], OpenHermes2.5 [122], NuminaMath-CoT [59], Python-Code-25k [28], Infinity-Instruct [7], Python-Code-Instructions-18k-Alpaca [38], Ruozhiba [73], InfinityMATH [147], StepDPO [53], TableLLM [151], UltraInteract-sft [144]

(a) Summary of the collected Eagle 2.5 SFT datasets

Category	Dataset
Captioning & Knowledge	CC3M [108], TextCaps [110], ShareGPT-4V [15], DenseFusion-1M [61]
Grounding & Counting	Object 365 [107]
Text-only	OpenMathInstruct [125]

(b) Summary of the additional Stage 1.5 datasets

Table 6: Dataset used in Eagle 2.5 for Stage 1 and Stage1.5. **Dataset in Magenta** is internal data.

## 6 Eagle-Video-110K

### 6.1 Data Curation Prompts

In this section, we will introduce the prompts utilized for curating dataset Eagle-Video-110K. They consist of prompts for generating captions and textual context anchors, prompts for generating clip-level QA, and prompts for generating video-level QA.

#### 6.1.1 Prompts for Generating Captions and Anchors

You are an expert in understanding visual content in video clips. You are requested to create

```

both brief and detailed captions for the current video clip titled "{title}".

#### Guidelines For Brief Caption:
- Create a concise summary (15-30 words) that captures the essential action, setting, and
  participants
- Focus on the most visually or narratively significant elements of the scene
- Use clear, direct language
- **IMPORTANT** Treat the video as a complete clip rather than a sequence of frames

#### Guidelines For Detailed Caption:
- Begin with a thorough analysis of the visual content shown in the clip
- **IMPORTANT** Pay special attention to the progression of actions and movements:
  * Break down complex actions into their component steps
  * Use transitional words (then, next, afterward, etc.) to show the flow of actions
  * Describe how one action leads to or connects with the next
  * Capture the natural sequence of movements and gestures
- Note that while the clip may be shown as multiple frames, it should be described as a
  continuous piece of footage
- For text that appears clearly in the clip: describe it in its original language and provide
  an English translation in parentheses). For example: [book in Chinese] [book]. Additionally,
  explain the meaning of the text within its context
- **IMPORTANT** If any text is unclear, partially visible, or too blurry to read with
  confidence, simply mention the presence of text without attempting to specify its content
- When referring to people, use their characteristics, such as clothing, to distinguish
  different people
- **IMPORTANT** Please provide as many details as possible in your caption, including colors,
  shapes, and textures of objects, actions and characteristics of humans, as well as scenes
  and backgrounds
- Consider how the visual content provides context and meaning

Only output your response in the following format without any additional text, explanations
or notes:

```json
{"Brief Caption":"concise summary of the video",
 "Detailed Caption":"The clip begins with..., progresses by...,
  and concludes with..."}

```

77

78 Here, the brief caption will be used for textual contextual anchors.

## 79 6.1.2 Prompts for generating clip-level QA

80 **### Task:**



You are tasked with generating question-answer pairs based on a detailed video caption and given question categories. Try to create challenging questions when possible, but simpler questions are also acceptable when the details are limited.

#### Input Information:

The detailed caption of the video clip is: {caption}

A brief version of the caption is also provided: {brief\_caption}

#### Question Generation Guidelines:

- Read the caption carefully
- Generate a question-answer pair for each category if possible
- Make sure the question-answer pair cannot be fully answered using only the brief caption
- Create two components for each pair:
  - \* A question that is unambiguous within the current clip
  - \* An answer (can be a word, phrase, or sentence)

#### Question Categories:

{question\_type\_pool}

#### Important Criteria:

- Questions should be unambiguous within the current clip
- Create challenging questions when the details allow
- Ensure answers require information beyond what's in the brief caption
- Generate questions for as many categories as possible
- Mark a category as null only if:
  - \* The question-answer would be fully answerable using just the brief caption
  - \* The category cannot be addressed using any information from either caption

Only output your response in the following format without any additional text, explanations or notes:

```
“‘json
{"question_type_1":{"Q":"question specific to this video","A":"answer"} or null,
"question_type_2":{"Q":"question specific to this video","A":"answer"} or null,
...,
"question_type_n":{"Q":"question specific to this video","A":"answer"} or null}“‘
```

81

82 The “question type pool” is defined in 6.2. And the “question\_type\_x”, x in (1, 2, ...n) are the types  
83 selected from the “question type pool”.

### 84 6.1.3 Prompts for generating generating video-level QA

85 #### Task:

You are tasked with generating question-answer pairs based on a video caption and given question categories. Try to create challenging questions when possible, but simpler questions are also acceptable when the details are limited.

#### Input Information:

The caption of the video clip is: {caption}

#### Question Generation Guidelines:

- Read the caption carefully
- Generate a question-answer pair for each category if possible
- Create two components for each pair:
  - \* A question that is unambiguous within the current clip
  - \* An answer (can be a word, phrase, or sentence)

#### Question Categories:

{question\_type\_pool}

#### Important Criteria:

- Questions should be unambiguous within the current clip
- Create challenging questions when the details allow
- Generate questions for as many categories as possible
- Mark a category as null only if:
  - \* The category cannot be addressed using any information from the caption

Only output your response in the following format without any additional text, explanations or notes:

```

“‘json
{"question_type_1":{"Q":"question specific to this video","A":"answer"} or null,
"question_type_2":{"Q":"question specific to this video","A":"answer"} or null,
...,
"question_type_n":{"Q":"question specific to this video","A":"answer"} or null}“‘

```

Here, “caption” is composed of clip-level caption, and its format is:

```

start_1 ~ end_1: caption_1
start_2 ~ end_2: caption_2
...

```

The “start\_x” and “end\_x”, (x in 1, 2, ...) are in seconds.

## 6.2 Question type pool

As shown in table 7, we list the category names and category descriptions in the question type pool. We ask the model to generate qa pairs according to these categories.

## References

- [1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *AAAI*, 2019.
- [2] aidapearson. Aida calculus math handwriting recognition dataset. <https://www.kaggle.com/datasets/aidapearson/ocr-data>, 2023.
- [3] Ashraq. TMDb-Celeb-10K Dataset. <https://huggingface.co/datasets/ashraq/tmdb-celeb-10k>, 2024.
- [4] Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240, 2023.
- [5] Anonymous Authors. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. In *Submission*, 2025.
- [6] Azu. Handwritten-mathematical-expression-convert-latex. <https://huggingface.co/datasets/Azu/Handwritten-Mathematical-Expression-Convert-LaTeX>, 2023.
- [7] BAAI. Infinity-instruct dataset. <https://huggingface.co/datasets/BAAI/Infinity-Instruct>, 2024.

Index	Category	Description
1	object_recognition	Questions about what an object is
2	object_properties	Questions about object properties, such as color, shape, material, texture
3	object_count	Questions about the number of objects
4	object_state	Questions about object states, such as stretched, compressed, cut, stationary
5	object_location	Questions about where an object is located
6	object_presence	Questions about object existence
7	human_attributes	Questions about human attributes, such as height, body type, build
8	human_pose	Questions about human posture
9	human_appearance	Questions about human external appearance, such as clothing and makeup
10	human_identity	Questions about human identity
11	human_cognitive_process	Questions about human mental processes, including intentions, motivations, decision-making rationale, problem-solving approaches, and reasoning methods
12	human_location	Questions about human location
13	human_emotion	Questions about human emotional state
14	scene_description	Questions about overall scene description
15	text_recognition	Questions about text content appearing in the video
16	text_count	Questions about frequency of text appearances in the video
17	text_location	Questions about location of text in the video
18	single_object_event_recognition	Questions about events involving a single object
19	single_object_event_count	Questions about frequency of single-object events
20	single_object_state_change	Questions about changes in single object state
21	single_object_quantity_change	Questions about changes in single object quantity
22	single_object_location_change	Questions about changes in single object location
23	single_object_trajectory	Questions about single object motion trajectory
24	single_object_speed	Questions about single object movement speed
25	single_object_presence_change	Questions about changes in single object presence
26	human_object_interaction_recognition	Questions about types of human-object interaction
27	human_object_interaction_count	Questions about frequency of human-object interactions
28	human_human_interaction_recognition	Questions about types of human-human interaction
29	object_interaction	Questions about objects' states, actions, interactions, changes, identifications (including brands), and how objects affect or interact with other objects
30	abnormal_event_detection	Questions about presence of abnormal events
31	domain_medical	Questions about medical-related professional knowledge
32	domain_education	Questions about education-related professional knowledge
33	domain_sports	Questions about sports-related professional knowledge
34	domain_movies	Questions about movie-related professional knowledge
35	domain_gaming	Questions about gaming-related professional knowledge
36	domain_technology	Questions about technology-related professional knowledge
37	domain_arts	Questions about arts-related professional knowledge
38	video_editing_effects	Questions about video editing effects, including shot transitions, editing effects, transition effects, etc.
39	camera_movement	Questions about camera motion
40	spatial_relationship	Questions about spatial relationships between objects
41	property_comparison	Questions about comparison of multiple object properties
42	quantity_comparison	Questions about comparison of multiple object quantities
43	state_comparison	Questions about comparison of multiple object states
44	human_object_relationship	Questions about human-object relationships
45	human_human_relationship	Questions about human-human relationships
46	scene_sequence	Questions about temporal relationships between scenes
47	event_sequence	Questions about temporal relationships between events
48	event_causality	Questions about causal relationships between events, including both human-initiated actions and their consequences, as well as cause-effect relationships in natural or systematic processes
49	counterfactual_reasoning	Questions about counterfactual reasoning
50	trajectory_tracking	Questions about tracking object or human positions
51	speed_comparison	Questions about speed comparison between multiple objects or humans
52	event_prediction	Questions about future event prediction
53	anomaly_reasoning	Questions about causes of anomalous phenomena
54	planning	Questions about planning for specific tasks
55	navigation	Questions about navigation to destinations
56	human_action	Questions about actions, behaviors, movements or activities performed by humans, including analysis of techniques, efficiency, and patterns of behavior
57	dialogue_content	Questions about spoken dialogue, verbal content, or conversations between characters/people
58	event_summary	Questions about overall event summary
59	object_ordering	Questions about the sequence or order in which objects are placed, arranged, or handled by individuals
60	event_location	Questions about where events or activities take place
61	process_description	Questions about identifying key components, steps, or progression in a process involving objects and/or humans
62	video_topic	Questions about the main subject, focus, or theme covered in the video
63	anomaly_recognition	Questions about identifying and interpreting anomalies

Table 7: Question type categories and their descriptions

- [8] Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv:2310.00367*, 2023.
- [9] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019.
- [10] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, 2022.
- [11] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation. In *ICASSP*, 2024.
- [12] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv:2211.08545*, 2022.
- [13] 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:2402.11684*, 2024.
- [14] 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:2212.02746*, 2022.
- [15] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023.
- [16] Wenhui Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *EMNLP*, 2023.
- [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*, 2019.
- [18] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, 2018.
- [19] LLM-Red-Team Contributors. emo-visual-data: Emotion and visual data analysis project. <https://github.com/LLM-Red-Team/emo-visual-data>, 2024.
- [20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv:2409.17146*, 2024.
- [21] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsr 2014). In *International Conference on Frontiers in Handwriting Recognition*, 2014.
- [22] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. VQA: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023.
- [23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.
- [24] EmileEsmaili. sheet music clean ataset. [https://huggingface.co/datasets/EmileEsmaili/sheet\\_music\\_clean](https://huggingface.co/datasets/EmileEsmaili/sheet_music_clean), 2024.
- [25] Jiarui Fang and Shangchun Zhao. Usp: A unified sequence parallelism approach for long context generative ai. *arXiv:2405.07719*, 2024.
- [26] FastJobs. Visual emotional analysis dataset. [https://huggingface.co/datasets/FastJobs/Visual\\_Emotional\\_Analysis](https://huggingface.co/datasets/FastJobs/Visual_Emotional_Analysis), 2024.

- [27] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv:2407.01449*, 2024.
- [28] flytech. Python codes 25k dataset. <https://huggingface.co/datasets/flytech/python-codes-25k>, 2024.
- [29] 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*, 2017.
- [30] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz Grand Challenge: Answering visual questions from blind people. In *CVPR*, 2018.
- [31] hazal karakus. mscoco-controlnet-canny-less-colors dataset. <https://huggingface.co/datasets/hazal-karakus/mscoco-controlnet-canny-less-colors>, 2024.
- [32] 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*, 2018.
- [33] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv:2003.10286*, 2020.
- [34] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.
- [35] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv:2209.08199*, 2022.
- [36] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. ICDAR 2019 robust reading challenge on scanned receipts ocr and information extraction. In *ICDAR*, 2019.
- [37] HugGAN. WikiArt Dataset. <https://huggingface.co/datasets/huggan/wikiart>, 2024.
- [38] iamtarun. Python code instructions 18k alpaca dataset. [https://huggingface.co/datasets/iamtarun/python\\_code\\_instructions\\_18k\\_alpaca](https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca), 2024.
- [39] ift. Handwriting forms dataset. [https://huggingface.co/datasets/ift/handwriting\\_forms](https://huggingface.co/datasets/ift/handwriting_forms), 2024.
- [40] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR Workshops*, 2019.
- [41] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv:1808.10584*, 2018.
- [42] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *CVPR*, 2018.
- [43] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv:1710.07300*, 2017.
- [44] Kamizuru00. Diagram image to text dataset. [https://huggingface.co/datasets/Kamizuru00/diagram\\_image\\_to\\_text](https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text), 2024.
- [45] 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:2203.06486*, 2022.

- [46] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.
- [47] 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*, 2017.
- [48] keremberke. Indoor scene classification dataset. <https://huggingface.co/datasets/keremberke/indoor-scene-classification>, 2024.
- [49] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.
- [50] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *ECCV*, 2022.
- [51] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9122–9134, 2023.
- [52] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [53] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv:2406.18629*, 2024.
- [54] LAION. gpt4v-dataset. <https://huggingface.co/datasets/laion/gpt4v-dataset>, 2023.
- [55] 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.
- [56] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv:2408.12637*, 2024.
- [57] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv:2405.02246*, 2024.
- [58] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv:2403.09029*, 2024.
- [59] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-CoT>] ([https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina\\_dataset.pdf](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)), 2024.
- [60] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv:2403.00231*, 2024.
- [61] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv:2407.08303*, 2024.
- [62] Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu, and Nathalie Abadie. ICDAR 2024 competition on historical map text detection, recognition, and linking. In *ICDAR*, 2024.

- [63] 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*, 2023.
- [64] W Lian, B Goodson, E Pentland, et al. OpenOrca: An open dataset of gpt augmented flan reasoning traces, 2023.
- [65] Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning. *arXiv:2208.05358*, 2022.
- [66] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv:2306.14565*, 2023.
- [67] 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:2311.10774*, 2023.
- [68] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv:2310.01889*, 2023.
- [69] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [70] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv:2407.15838*, 2024.
- [71] 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:2406.11833*, 2024.
- [72] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. ICDAR 2023 competition on hierarchical text detection and recognition. In *ICDAR*, 2023.
- [73] LooksJuicy. Ruozhiba dataset. <https://huggingface.co/datasets/LooksJuicy/ruozhiba>, 2024.
- [74] 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:2105.04165*, 2021.
- [75] 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:2105.04165*, 2021.
- [76] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv:2110.13214*, 2021.
- [77] 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. In *NeurIPS*, 2022.
- [78] 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:2209.14610*, 2022.
- [79] Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv:2406.11069*, 2024.
- [80] Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. WeatherQA: Can multimodal language models reason about severe weather? *arXiv:2406.11217*, 2024.



- [81] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d’Andecy, and Jean-Marc Ogier. CHIC: Corporate document for visual question answering. In *ICDAR*, 2024.
- [82] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.
- [83] 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.
- [84] 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*, 2022.
- [85] 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:2305.14761*, 2023.
- [86] Minesh Mathew, Lluís Gomez, Dimosthenis Karatzas, and CV Jawahar. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition*, 24(3):235–249, 2021.
- [87] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *WACV*, 2022.
- [88] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *WACV*, 2020.
- [89] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.
- [90] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019.
- [91] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-Math: Unlocking the potential of slms in grade school math. *arXiv:2402.14830*, 2024.
- [92] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In *International Conference on Frontiers in Handwriting Recognition*, 2016.
- [93] mPLUG. DocReason25k dataset. <https://huggingface.co/datasets/mPLUG/DocReason25K>, 2024.
- [94] mychen76. Invoices and receipts ocr v1 dataset. [https://huggingface.co/datasets/mychen76/invoices-and-receipts\\_ocr\\_v1](https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1), 2024.
- [95] Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. *arXiv:2409.13592*, 2024.
- [96] OleehyO. Latex formulas dataset. <https://huggingface.co/datasets/OleehyO/latex-formulas>, 2024.
- [97] OpenGVLab. ShareGPT-4o dataset. <https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o>, 2024.
- [98] parasam. Captcha Dataset. <https://www.kaggle.com/datasets/parasam/captcha-dataset>, 2024.
- [99] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image Textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv:2406.07502*, 2024.



- [100] PixArt-alpha. SAM-LLaVA-Captions10M Dataset. <https://huggingface.co/datasets/PixArt-alpha/SAM-LLaVA-Captions10M>, 2024.
- [101] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.
- [102] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.
- [103] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NeurIPS*, 2015.
- [104] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, 2022.
- [105] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP*, 2015.
- [106] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware visual question answering. In *AAAI*, 2019.
- [107] 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*, 2019.
- [108] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- [109] shreyanshu09. Block diagram dataset. [https://huggingface.co/datasets/shreyanshu09/Block\\_Diagram](https://huggingface.co/datasets/shreyanshu09/Block_Diagram), 2024.
- [110] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.
- [111] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv:2312.14150*, 2023.
- [112] 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*, 2019.
- [113] skvarre. Movie posters-100k dataset. [https://huggingface.co/datasets/skvarre/movie\\_posters-100k](https://huggingface.co/datasets/skvarre/movie_posters-100k), 2024.
- [114] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [115] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017.
- [116] Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa, 2024. URL <https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k>.
- [117] 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*, 2019.
- [118] TAL. TAL open dataset. <https://ai.100tal.com/dataset>, 2023.
- [119] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

- [120] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv:2307.05356*, 2023.
- [121] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv:2405.11985*, 2024.
- [122] Teknium. OpenHermes 2.5: An open dataset of synthetic data for generalist llm assistants. <https://huggingface.co/datasets/teknium/OpenHermes-2.5>, 2023.
- [123] TIGER-Lab. VisualWebInstruct Dataset. <https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct>, 2024.
- [124] George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. ICDAR 2023 competition on roadtext video text detection, tracking and recognition. In *ICDAR*, 2023.
- [125] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. OpenMathInstruct-1: A 1.8 million math instruction tuning dataset. *arXiv:2402.10176*, 2024.
- [126] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv:2311.16101*, 2023.
- [127] 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:1601.07140*, 2016.
- [128] VQAonDB. Vqaondb dataset. <https://ilocr.iiit.ac.in/vqabd/>.
- [129] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574*, 2023.
- [130] 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*, 2020.
- [131] wendlerc. Renderedtext dataset. <https://huggingface.co/datasets/wendlerc/RenderedText>, 2024.
- [132] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*, 2020.
- [133] Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv:2407.17379*, 2024.
- [134] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *ECCV*. Springer, 2022.
- [135] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv:2410.02712*, 2024.
- [136] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv:2304.12244*, 2023.
- [137] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *ICCV*, 2019.
- [138] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv:2310.05126*, 2023.

- [139] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- [140] 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:2309.12284*, 2023.
- [141] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv:2405.17220*, 2024.
- [142] Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. ICDAR 2023 competition on structured text extraction from visually-rich document images. In *ICDAR*, 2023.
- [143] Youngjoon Yu, Sangyun Chung, Byung-Kwan Lee, and Yong Man Ro. SPARK: Multi-vision sensor perception and reasoning benchmark for large-scale vision-language models. *arXiv:2408.12114*, 2024.
- [144] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv:2404.02078*, 2024.
- [145] 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. *Journal of Computer Science and Technology*, 34, 2019.
- [146] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv:2309.05653*, 2023.
- [147] Bo-Wen Zhang, Yan Yan, Lin Li, and Guang Liu. Infinitymath: A scalable instruction tuning dataset in programmatic mathematical reasoning. In *ACM International Conference on Information and Knowledge Management*, 2024.
- [148] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for Relational and Analogical Visual Reasoning. In *CVPR*, 2019.
- [149] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. MAVIS: Mathematical visual instruction tuning. *arXiv:2407.08739*, 2024.
- [150] 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*, 2019.
- [151] Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, et al. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv:2403.19318*, 2024.
- [152] 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:2306.17107*, 2023.
- [153] Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhui Wang, Guangtao Zhai, et al. Omniaalign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*, 2025.
- [154] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertr: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv:2206.01347*, 2022.
- [155] Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv:2306.14321*, 2023.

- 485 [156] Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng Yan.  
486 Agentstudio: A toolkit for building general virtual agents. *arXiv:2403.17918*, 2024.
- 487 [157] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhui  
488 Chen, and Xiang Yue. OpenCodeInterpreter: Integrating code generation with execution and  
489 refinement. *arXiv:2402.14658*, 2024.
- 490 [158] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A  
491 10 million image database for scene recognition. *IEEE transactions on pattern analysis and*  
492 *machine intelligence*, 40(6):1452–1464, 2017.
- 493 [159] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua.  
494 Towards complex document understanding by discrete reasoning. In *ACMMM*, 2022.