Focus, Don't Prune: Identifying Instruction-Relevant Regions for Information-Rich Image Understanding

Supplementary Material

A. Experimental Details

A.1. Implementation Details

LLaVA-NeXT. LLaVA-NeXT [7] employs the Dynamic High Resolution (AnyRes) module, which processes high-resolution images by dividing them into multiple patches and encoding them independently. In all experiments, we follow the same configuration and allow AnyRes to generate at most 6 patches per image. During the Region Selection stage, approximately 60% of the original image area is selected and re-encoded. Since Region Refinement operates on this smaller cropped area, we restrict AnyRes to at most 4 patches in this stage, which is sufficient to capture finer-grained visual details while further reducing computational cost.

Qwen2-VL. The Qwen2-VL [16] model processes visual tokens by adapting to the scale of the input image using its Native Dynamic Resolution method. As in the LLaVA-NeXT experiments, the region refinement stage only re-processes a cropped subset of the image. Therefore, during this re-encoding step, we restrict the vision processor's maximum processed pixels to 60% of the original setting, which effectively reduces the computational cost while still preserving sufficient resolution to capture fine-grained visual details.

B. Additional Experiments

B.1. Efficiency Analysis of Instruction-Region Alignment

We analyze the computational overhead introduced by the PinPoint module, which performs region selection and region refinement using a small set of learnable queries. Table A1 reports the FLOPs of PinPoint itself and the average total FLOPs across all VQA datasets [6, 10, 11, 14] when PinPoint is adopted. For LLaVA-NeXT [7], PinPoint adds only 1.67T FLOPs, corresponding to 7.14% of the total 23.37T FLOPs. For Qwen2-VL [16], the overhead is 5.68T FLOPs (22.3% of 25.44T), reflecting that its Native Dynamic Resolution scheme feeds more visual tokens into the ViT [5] backbone during region refinement than LLaVA-NeXT [7]. This result confirms that PinPoint effectively locates instruction-relevant regions with minimal additional computation, maintaining a significantly lower operational cost compared to vanilla model.

Table A1. **Computational Overhead of the PinPoint Module.** The table reports FLOPs of the frozen model (Vanilla), FLOPs (T) with PinPoint, and the resulting ratio.

Model	Vanilla	Ours		
	Total	PinPoint	Total	Ratio
LLaVA-NeXT-7B	40.30	1.67	23.37	7.14%
Qwen2-VL-7B	52.30	5.68	25.44	22.30%

Table A2. **Performance Analysis of Focused Region Selection for Qwen2-VL [16].** PinPoint effectively identifies and supplies answer-relevant regions, thereby improving performance on VQA tasks. Metric: ANLS. r: 0.3.

Model	InfoVQA	SPDocVQA	MPDocVQA	GQA
Vanilla [16]	0.7399	0.9359	0.7775	**0.7687**
Ours + Global	**0.7560**	**0.9380**	**0.8071**	0.7682

B.2. Quantitative Results

Effectiveness of Providing Pinpointed Regions to Qwen2-VL. Table A2 reports the results of augmenting Qwen2-VL [16] with both the full image (Global) and the answer-related crops identified by PinPoint. Unlike our main experiments, where only the PinPoint-selected regions are provided, here the model receives the global image together with the localized instruction-relevant regions. The results show substantial performance gains on fine-grained VQA datasets that require attending to small visual details (e.g., InfoVQA, MPDocVQA), confirming that PinPoint effectively localizes answer-related regions and that explicitly highlighting these regions enables Qwen2-VL to generate more accurate answers.

Table A3. **Performance comparison of training-based efficiency methods.** Our method outperforms Honeybee [3] across all evaluated datasets with lower computational cost. Model: LLaVA-NeXT [7]

Method	InfoVQA	SPDocVQA	MPDocVQA	GQA	Avg. FLOPs(T)
Vanilla [7]	0.2552	0.6628	0.3758	0.7598	42.77
Honeybee [3]	0.2287	0.4451	0.2395	0.7460	27.80
Ours	**0.3024**	**0.6472**	**0.3866**	**0.7608**	25.04

Comparison with Training Method. To ensure a fair comparison, we evaluate our method against HoneyBee [3], a state-of-the-art approach that enhances VLM efficiency through projector training—specifically utilizing the C-Abstractor architecture. Under identical training configu-

Q: Describe the given image.

(a) Vanilla

The image captures a cozy living room scene. Dominating the center of the room is a **green Christmas tree**, its branches reaching upwards towards a **white arched window**. The window, adorned with a **white curtain**, allows a soft light to filter into the room, illuminating the tree and casting a warm glow on the surroundings. To the left of the tree, a **black fireplace** stands tall, its mantel decorated with a **white vase** and a **black candle holder**. The fireplace adds a touch of elegance to the room. On the right side of the tree, a **black entertainment center** holds a **black TV** and a **black speaker**. The entertainment center provides a modern contrast to the traditional charm of the Christmas tree. The walls of the room are painted in a soothing shade of **beige**, complementing the overall warm and inviting atmosphere of the space. The floor is covered with a **beige carpet**, adding to the room's homely feel.

(b) FastV

A small Christmas tree in a living room.

(c) PyramidDrop

A small Christmas tree in a living room.

(d) PinPoint (Ours)

The image captures a cozy living room scene. Dominating the right side of the frame is a black flat-screen TV, its sleek surface reflecting the ambient light. To the left, a fireplace stands as a warm and inviting focal point. Above it, a framed picture adds a personal touch to the room. In the center of the image, a Christmas tree stands tall, its branches reaching upwards towards the ceiling. The tree is adorned with a string of lights that twinkle softly, casting a festive glow throughout the room. The room itself is painted in a soothing shade of beige, providing a neutral backdrop that allows the other elements to stand out. A window punctuates the wall on the left, its presence hinting at the world outside. Overall, the image paints a picture of a well-decorated living room, ready to welcome its occupants and their guests for a festive holiday season.

Figure A1. A qualitative comparison among LLaVA-NeXT [7], FastV [4], PyramidDrop [17], and our PinPoint shows that the vanilla model [7] produces rich but often hallucinated descriptions, whereas FastV [4] and PyramidDrop [17] generate overly short outputs that miss key details. In contrast, PinPoint yields coherent, detailed descriptions with fewer hallucinations by leveraging instruction-relevant regions. Note that blue marks the windows selected by PinPoint, and purple indicates the pinpointed answer-relevant regions. In addition, light blue indicates well-grounded keywords, whereas light red denotes hallucinated keywords.

Table A4. **Hallucination evaluation on POPE [8] (discriminative) and CHAIR [12] / AMBER [15] (generative).** PinPoint achieves the lowest hallucination rates across POPE [8], CHAIR [12], and AMBER [15], outperforming in both discriminative and generative settings while maintaining strong reliability. Model: LLaVA-NeXT [7].

Method	POPE			CHAIR		AMBER				Avg. FLOPs(T)
	Rand. ↑	Pop. ↑	Adv. ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	
Vanilla	88.1	86.8	85.6	26.1	7.8	8.7	**62.1**	48.9	4.6	34.32
Ours	**89.0**	**87.7**	**86.0**	**25.6**	**7.1**	**8.0**	53.1	**42.4**	**3.9**	**22.18**

Table A5. **Cross-dataset performance on general and real-world benchmarks.** Despite being trained exclusively on the GQA [6] dataset, PinPoint demonstrates superior generalization and robustness across challenging real-world tasks, including MMMU [18], MMMU-Pro [19] and TextVQA [13]. Our method significantly outperforms state-of-the-art train-free approaches while achieving the lowest computational cost (FLOPs).

Method	Training Data	MMMU	MMMU-Pro (standard 10)	TextVQA	Avg. FLOPs(T)
Vanilla [7]	-	0.3410 (100.0%)	0.1864 (100.0%)	0.5327 (100.0%)	39.92 (100.0%)
FastV [4]	*Train-free*	0.3356 (98.4%)	0.1927 (103.4%)	0.5102 (95.8%)	28.17 (70.6%)
PDrop [17]	*Train-free*	0.3411 (100.0%)	0.1871 (100.4%)	0.5240 (98.4%)	26.00 (65.1%)
PinPoint (Ours)	**GQA**	**0.3500 (102.6%)**	**0.1990 (106.8%)**	**0.7293 (136.9%)**	**24.39 (61.1%)**

rations (e.g., data and epochs), PinPoint consistently outperforms HoneyBee in both efficiency (FLOPs) and performance (ANLS), as summarized in Table A3. These results highlight our approach's superiority in extracting instruction-relevant, fine-grained features without the accuracy trade-offs typically associated with conventional token reduction techniques.

Evaluation on Hallucination Benchmark. To evaluate the robustness of our method against hallucination, we conduct experiments on three established hallucination benchmarks that span both discriminative and generative settings. For POPE [8], a discriminative benchmark constructed on the MSCOCO dataset [9], we follow the standard evaluation protocol and report F1 scores across all categories. For the generative benchmarks CHAIR [12] and AMBER [15], we evaluate on a representative subset by sampling 10% of the MSCOCO [9] validation split for CHAIR [9] (approximately 4K frames) and follow the standard AMBER [15] protocol. As shown in Table A4, PinPoint consistently reduces hallucinated predictions across both discriminative and generative benchmarks, demonstrating strong reliability in mitigating hallucination and

Table A6. **Effect of Encompass Supervision on Instruction–Region Alignment.** compare two supervision settings for instruction-relevant regions: (a) a single, tightly bounded answer box, and (b) encompass regions that aggregate multiple reasoning boxes covering all supporting visual evidence. Using encompass supervision (b) consistently yields higher ANLS and Region Accuracy across all datasets and base models, showing that including contextual supporting regions leads to better instruction–image alignment.

Model	Setting	InfoVQA		SPDocVQA		MPDocVQA	
		ANLS↑	Region Acc.↑	ANLS↑	Region Acc.↑	ANLS↑	Region Acc.↑
LLaVA-NeXT-7B	(a)	0.2980	82%	0.6457	97%	0.3836	86%
	(b)	**0.3024**	**84%**	**0.6472**	**98%**	**0.3866**	**87%**
Qwen2-VL-7B	(a)	0.7017	93%	0.8843	96%	0.6572	90%
	(b)	**0.7140**	**95%**	**0.8977**	**98%**	**0.6723**	**94%**

Table A7. **Ablation on Region Coverage Threshold r.** We vary the region coverage threshold r in the Region Selection stage and report answer accuracy (ANLS), region localization accuracy (Acc.), and region coverage (Cov.) for each dataset and base model. Across all settings, PinPoint maintains strong performance, with larger r values increasing coverage and generally yielding modest ANLS gains.

Model	r	InfoVQA			SPDocVQA			MPDocVQA			GQA		
		ANLS	Acc.	Cov.	ANLS	Acc.	Cov.	ANLS	Acc.	Cov.	ANLS	Acc.	Cov.
LLaVA-NeXT-7B	20%	0.2964	67%	36%	0.5724	83%	32%	**0.3891**	71%	36%	0.7219	85%	32%
	40%	0.3014	77%	55%	0.6270	93%	52%	0.3836	83%	56%	0.7508	93%	50%
	60%	**0.3024**	84%	71%	**0.6472**	98%	71%	0.3866	87%	72%	**0.7608**	98%	69%
Qwen2-VL-7B	20%	0.5882	69%	33%	0.7405	84%	30%	0.6094	73%	35%	0.7172	86%	30%
	40%	0.6649	87%	54%	0.8563	95%	54%	0.6578	86%	56%	0.7521	94%	51%
	60%	**0.7140**	95%	72%	**0.8977**	98%	72%	**0.6723**	94%	74%	**0.7624**	98%	71%

improving computational efficiency. We evaluate PinPoint trained on GQA [6], since it shares a similar real-world image distribution with MSCOCO [9]. The slight decrease in the Cover metric on AMBER [15] is attributed to PinPoint's focus on instruction-relevant regions rather than exhaustive object coverage, which aligns with our goal of grounding responses in the most semantically relevant visual evidence.

Comparison with Cross-Data Setting. To validate the generality and robustness of our approach, we evaluate PinPoint in a cross-dataset setting across expert-level reasoning tasks (MMMU [18, 19]) and real-world scene-text understanding (TextVQA [13]). Despite being optimized exclusively on the GQA [6] dataset, PinPoint consistently surpasses all state-of-the-art train-free baselines. Notably, on TextVQA [13], it achieves a remarkable 136.9% relative performance compared to the Vanilla baseline while reducing computational costs to only 61.1% (24.39 TFLOPs). This significant performance-efficiency gain on unseen benchmarks demonstrates that PinPoint effectively masters the identification of instruction-relevant regions rather than overfitting to its training distribution. These results confirm that our module provides a robust, general-purpose solution for enhancing VLM efficiency without sacrificing complex reasoning capabilities in diverse real-world scenarios.

B.3. Qualitative Results

Comparison for Caption Generation. We qualitatively compare LLaVA-NeXT [7], FastV [4], PyramidDrop [17], and our proposed PinPoint on the image captioning task (i.e., describing the given image), as shown in Figure A1. The vanilla LLaVA-NeXT [7] generates rich captions but often introduces hallucinated content. In contrast, FastV [4] and PyramidDrop [17] produce overly brief responses that miss important visual details. PinPoint generates detailed and coherent captions by effectively leveraging instruction-relevant regions, while also reducing hallucinations.

Additional Examples of PinPoint. We provide additional qualitative examples showing that PinPoint reliably identifies instruction-relevant regions across diverse image sizes and visual layouts. These results indicate that our method consistently focuses on the appropriate visual evidence and produces accurate responses, regardless of variations in image resolution or scene complexity. Additional qualitative results are shown in Figure A2 and A3 for InfoVQA [11], in Figure A4 for SPDocVQA [10], and in Figure A5 for GQA [6].

B.4. Additional Ablation

Impact of Contextual Supporting Regions. We investigate how using datasets that explicitly include contextual regions around the answer influences instruction–region alignment. Our PinPoint dataset annotates not only the answer-containing region but also an encompass region that includes surrounding contextual elements necessary to infer the answer. As shown in Table A6, training with only the tight answer box (setting (a)) yields consistently lower Region Accuracy and ANLS than training with encompass annotations that cover all supporting evidence (setting (b)).

Table A8. **Number of Bounding Box Annotations and QA Pairs per Dataset.**

Dataset	Split	Annotation Count			Image Count	Image Pages	QA Pairs
		answer	evidence	encompass			
InfoVQA	Train	1.06	1.64	1.0	4,162	1.0	17,887
	Validation	1.09	1.53	1.0	500	1.0	2,801
SPDocVQA	Train	1.00	0.30	1.0	10,194	1.0	39,463
	Validation	1.00	0.28	1.0	1,286	1.0	5,349
MPDocVQA	Train	1.00	1.05	1.0	5,131	9.1	36,230
	Validation	1.00	1.02	1.0	927	5.6	5,187

Table A9. **Distribution of InfoVQA [11] Questions by Processing Type.**

Category	Train (%)	Validation (%)
Visually Grounded	41.4	38.7
OCR-Extractable (Unique)	39.4	31.2
OCR-Extractable (Multiple)	12.6	21.9
Manual Annotation	6.6	8.2

Table A10. **Distribution of SPDocVQA [10] Questions by Processing Type.**

Category	Train (%)	Validation (%)
Visually Grounded	35.2	28.0
OCR-Extractable (Unique)	62.6	59.0
OCR-Extractable (Multiple)	2.2	12.9
Manual Annotation	0.0	0.1

Table A11. **Distribution of MPDocVQA [14] Questions by Processing Type.**

Category	Train (%)	Validation (%)
Visually Grounded	39.0	57.4
OCR-Extractable (Unique)	52.2	34.0
OCR-Extractable (Multiple)	8.6	7.9
Manual Annotation	0.2	0.7

This suggests that effective cross-modal alignment benefits from supervising not just the answer location itself, but the broader set of answer-related regions that ground the instruction semantically.

Sensitivity to the Region Coverage Threshold r. In the Region Selection stage, the number of selected regions k is adaptively determined by a predefined region coverage threshold r. Table A7 reports, for each value of r, the resulting region coverage (Cov.), region localization accuracy (Acc.), and ANLS of the generated answers. When using LLaVA-NeXT [7] as the base model, performance on fine-grained datasets that require focusing on small details (e.g., InfoVQA, MPDocVQA) remains relatively stable as r varies, whereas datasets that involve broader layouts (e.g., SPDocVQA, GQA) benefit from larger coverage and show improved ANLS with higher r. For Qwen2-VL [16], enlarging the covered area generally improves performance, indicating that this model gains from seeing more surrounding context. Notably, even at $r = 40\%$, where only about 40% of the image area is retained, PinPoint already achieves strong accuracy compared to competing methods (see main results), suggesting that our instruction–region alignment remains effective over a wide range of coverage thresholds.

C. Dataset

C.1. Dataset Generation Pipeline

We construct new annotated datasets for InfographicVQA (InfoVQA) [11], SinglePageDocVQA (SPDocVQA) [10], and MultiPageDocVQA (MPDocVQA) [14], in which we explicitly mark regions that are relevant to answering the question. These annotations go beyond a single answer box and include essential supporting evidence. For instance, for the question "What is the object the man is pointing at?", it is necessary not only to localize the pointed object but also to first locate the man; our annotations therefore include both the man and the target object as instruction-relevant regions.

To obtain these annotations, we design a unified dataset construction pipeline that is shared across all benchmarks, with minor adjustments to accommodate dataset-specific characteristics. The pipeline combines (i) an OCR engine (Amazon Textract), (ii) a vision-language model (Qwen2.5-VL [2]), and (iii) a large language model (GPT-4o [1]).

Relying on a single model to directly predict answer regions is often unreliable for information-dense (e.g., document layout) images, so we handle each sample using three complementary cases based on the OCR outcome. If the answer-related text can be found by OCR and appears exactly once, we directly use the corresponding bounding box as the instruction-relevant region. If multiple candidate boxes contain the answer text, we retain all candidates and use an large-language model to perform additional reasoning given the question, answer, OCR text, and candidate regions, selecting only the most relevant box. If the answer text cannot be recovered by OCR, indicating that complex visual reasoning is required, we instead use a vision-language model to infer and annotate the appropriate answer-related region.

While the overall pipeline structure is consistent, we introduced a specific enhancement for the challenging InfoVQA [11] dataset, which demands complex reasoning. For this dataset, an LLM generates rationale sentences from the input informations (question, answer, and OCR) to guide the localization of the corresponding reasoning-related regions. Finally, all datasets underwent a quality control step where instances with failed or invalid bounding box outputs were manually corrected via self-annotation, ensuring high-quality supervision.

C.2. Dataset Analysis

Table A8 summarizes the detailed statistics of our newly constructed annotations. For InfoVQA [11], the most challenging benchmark, answers often appear at multiple locations, and each question typically requires the largest number of evidence regions to be correctly resolved. Although both SPDocVQA [10] and MPDocVQA [14] are document-based datasets, MPDocVQA [14] consistently demands more supporting evidence per question, indicating a higher level of compositional and cross-page reasoning.

As described in Section C.1, our pipeline processes each sample through four dedicated cases to accurately identify answer-relevant regions. For InfoVQA [11], Table A9 shows that the Visually Grounded case accounts for nearly 40% of all samples, indicating that a large portion of questions cannot be resolved from OCR text alone and genuinely require visual reasoning. In contrast, for the document-layout datasets SPDocVQA [10] and MPDocVQA [14], Tables A10 and A11 shows that OCR-based cases constitute the majority, reflecting that most answers can be localized primarily via textual cues.

Across all datasets, a substantial fraction of samples still relies on joint text–image reasoning, which supports our decision to explicitly separate and handle these cases within the pipeline. Moreover, the proportion of manually annotated samples remains below 10% on every benchmark, demonstrating that our method significantly reduces human labeling effort while maintaining high annotation quality.

C.3. Dataset Examples

Figure A6 and A7 present qualitative examples of the annotations on InfoVQA [11] and SPDocVQA [10]. The encompass regions unify the answer and its supporting evidence into a single region, allowing the model to learn complete instruction-relevant cues. These examples illustrate that a substantial portion of document-understanding questions require multi-step reasoning across spatially dispersed content, and the encompass annotations effectively capture all necessary components for deriving the correct answer.

C.4. Prompt Design

We next describe the prompts used to construct the PinPoint dataset annotations. Figure A8 illustrates the prompt used when multiple candidate bounding boxes contain the answer text, where an LLM is asked to select the region most relevant to the question. Figures A9 and A10 show the prompts additionally used for InfoVQA [11]: the first elicits rationale sentences given the question, answer, and OCR text, and the second leverages these rationales to localize the corresponding reasoning-related regions. Finally, Figure A11 presents the prompt used when OCR fails to recover the answer text, in which a vision-language model performs visual reasoning to identify and annotate the answer-related region directly from the image.

Figure A2. Additional Qualitative Results for PinPoint on InfoVQA [11] large-size dataset. Red marks the ground-truth regions, blue shows the window selected by our method.

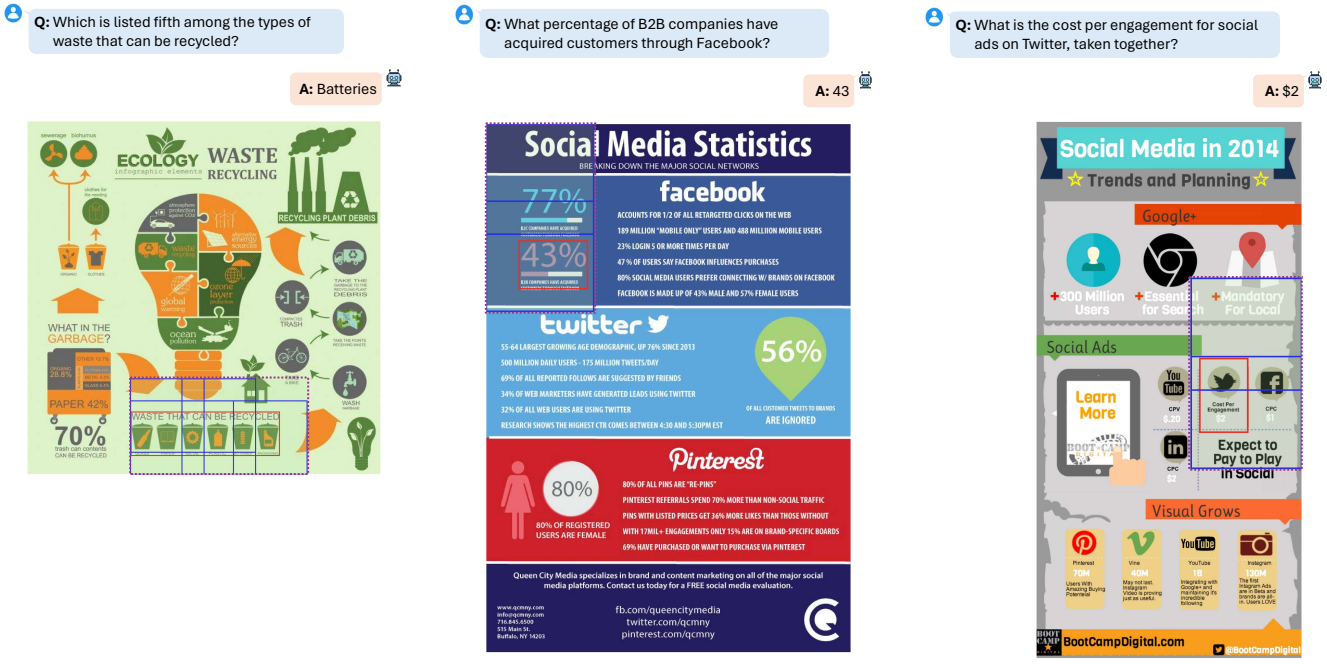


Figure A3. Additional Qualitative Results for PinPoint on InfoVQA [11] medium-size dataset. Red marks the ground-truth regions, blue shows the windows selected by our method, and purple highlights the pinpointed answer-relevant areas.

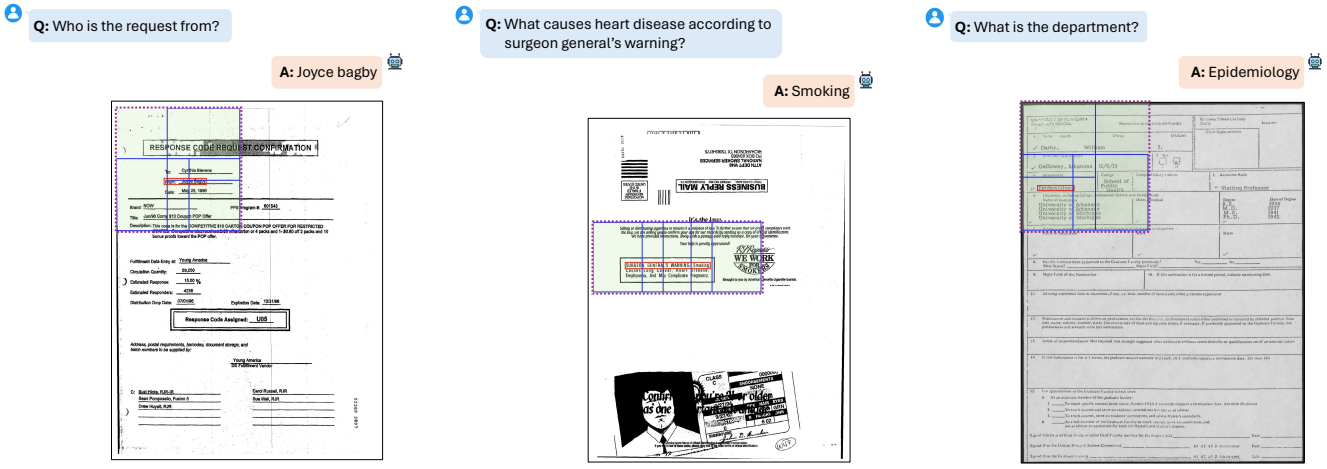


Figure A4. Additional Qualitative Results for PinPoint on SPDocVQA [10] dataset. Red marks the ground-truth regions, blue shows the windows selected by our method, and purple highlights the pinpointed answer-relevant areas.

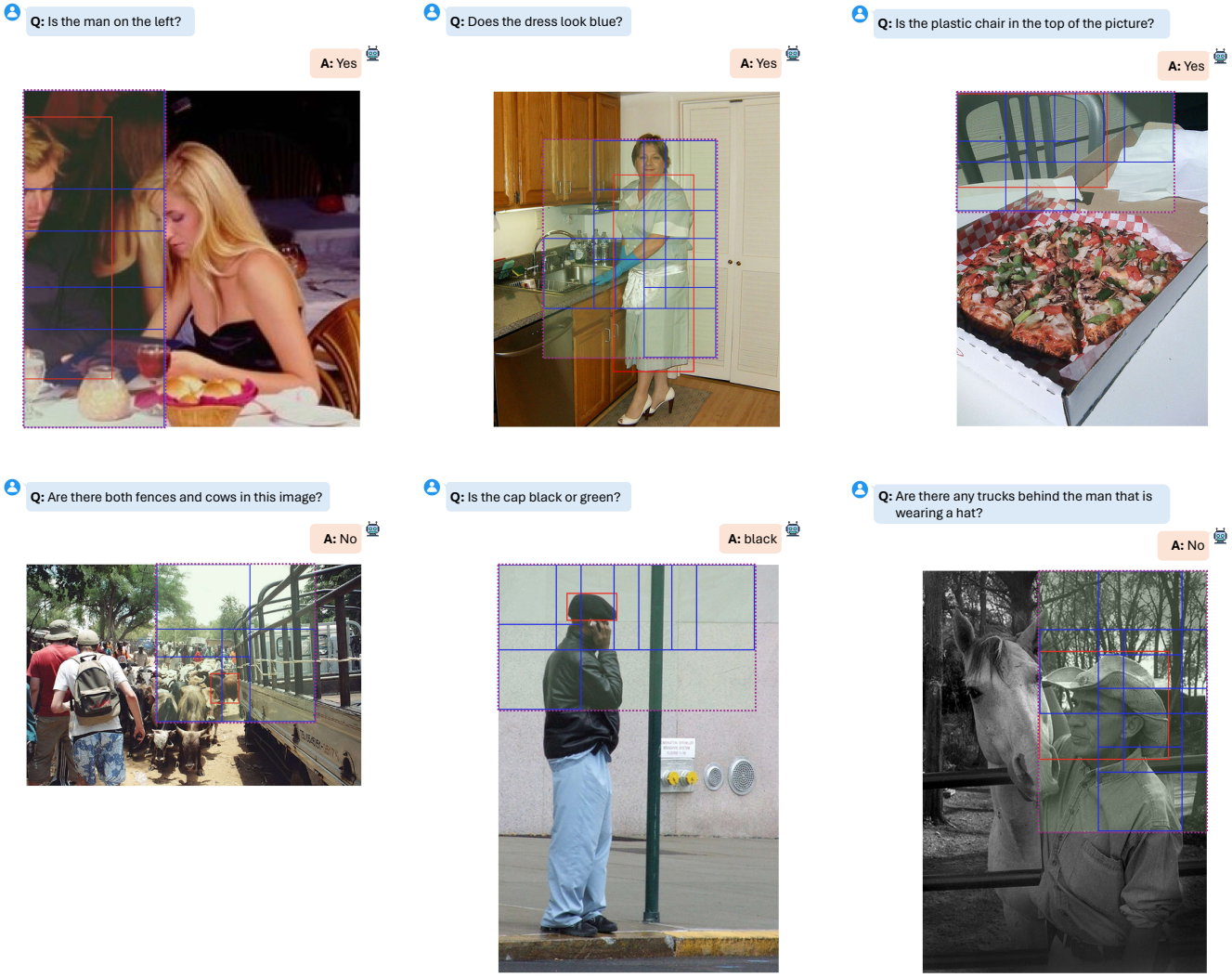
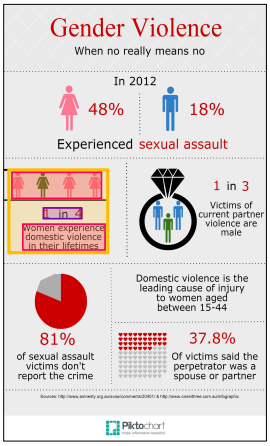
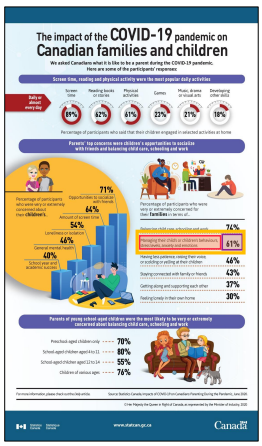


Figure A5. Additional Qualitative Results for PinPoint on GQA [6] dataset. Red marks the ground-truth regions, blue shows the windows selected by our method, and purple highlights the pinpointed answer-relevant areas.

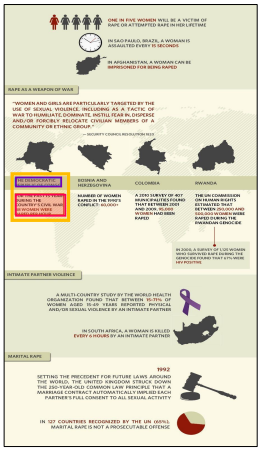


Q: What percentage of women among four women experience domestic violence in their lifetime?
GT Answer: 25%
Encompass_bbox: [66,1428,908,2094]

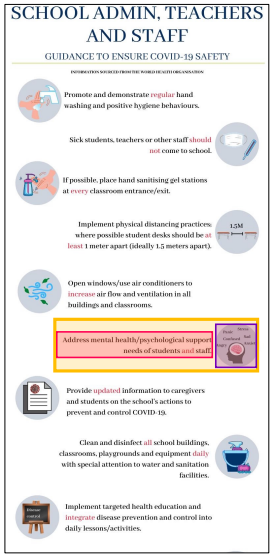
Q: What is the number of searches in Google?
GT Answer: 4.27 million
Encompass_bbox: [503,580,716,647]



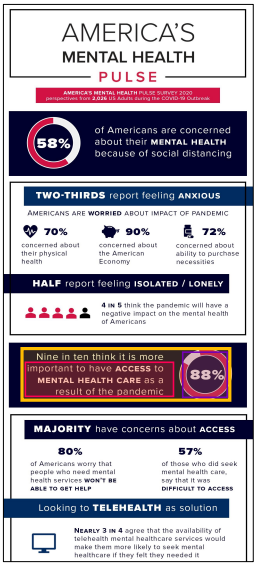
Q: What percentage of respondents said that their children has been engaged in physical activities daily or almost every day during the Covid-19 pandemic according to the survey?
GT Answer: 61%
Encompass_bbox: [522,347,608,478]



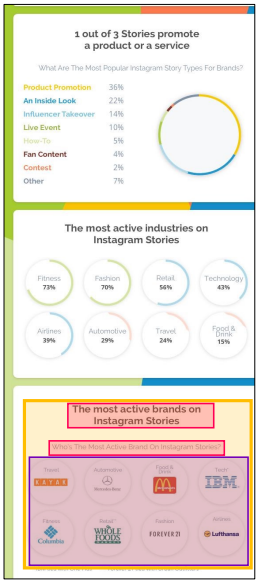
Q: Which country's civil war period reported 48 women raped every hour?
GT Answer: the democratic republic of congo
Encompass_bbox: [38,973,214,1134]



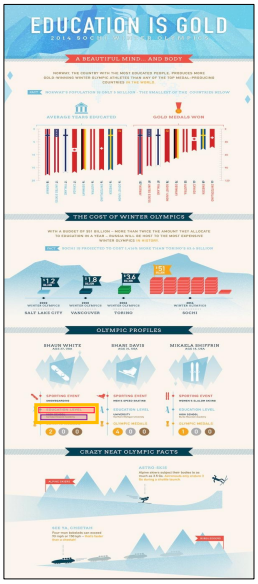
Q: How many symptoms are related to unsound mind?
GT Answer: 6
Encompass_bbox: [642,1114,784,1250]



Q: What percentage of people feel it is not important to have access to mental health care?
GT Answer: 12%
Encompass_bbox: [120,1850,1140,1992]

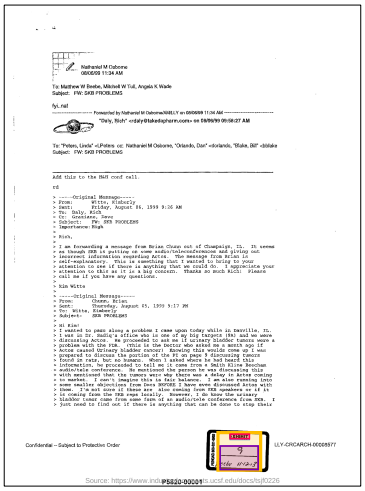


Q: How many active brands on Instagram stories?
GT Answer: 8
Encompass_bbox: [48,2768,532,2105]

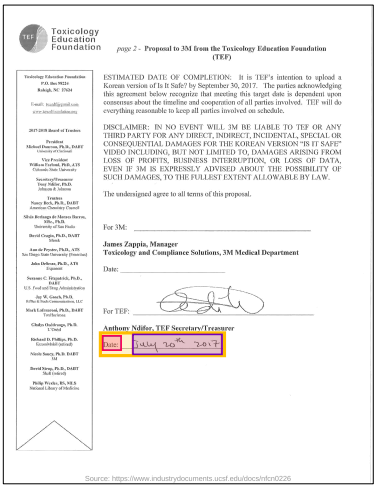


Q: Where did Shaun White complete his high school?
GT Answer: carlsbad seaside academy
Encompass_bbox: [100,1260,200,1284]

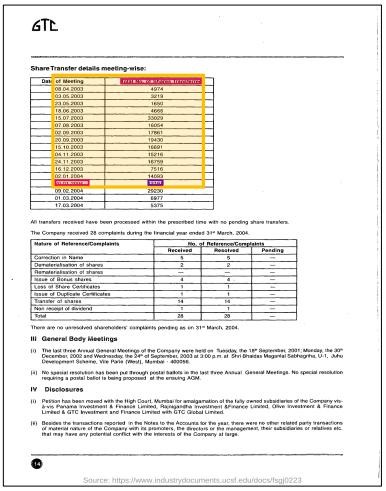
Figure A6. Qualitative examples of the PinPoint annotations on InfoVQA [11]. Purple boxes denote answer regions, pink boxes indicate supporting evidence regions, and yellow boxes represent encompass regions that jointly cover both answer and evidence, providing a more complete instruction-relevant area for supervision.



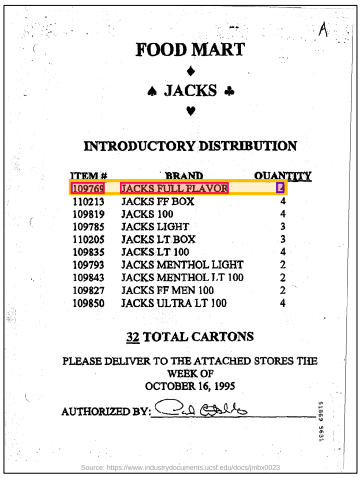
Q: What is the EXHIBIT number mentioned?
GT Answer: 9
Encompass_bbox:
[1027,1992,1211,2166]



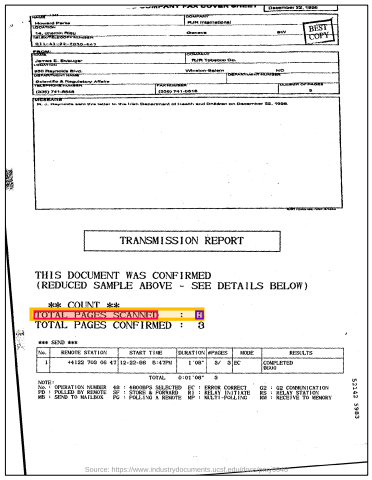
Q:What is the date mentioned at the bottom?
GT Answer: July 20th 2017
Encompass_bbox:
[458,1510,1004,1608]



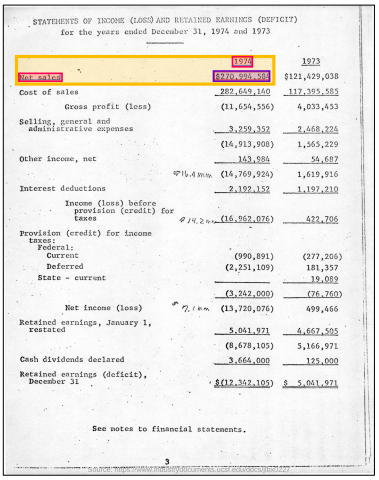
Q: what is the total no .of shares transferred on 20.01.2004
GT Answer: 12191
Encompass_bbox:
[222,798,686,820]



Q: What is the Quantity for Jacks Full Flavor?
GT Answer: 4
Encompass_bbox:
[326,852,1342,900]



Q:What are the Total Pages Scanned?
GT Answer: 3
Encompass_bbox:
[136,1456,922,1498]



Q: what is the amount of net sales in 1974 ?
GT Answer: $270,994,584
Encompass_bbox:
[72,314,1230,355]

Figure A7. Qualitative examples of the PinPoint annotations on SPDocVQA [10]. Purple boxes denote answer regions, pink boxes indicate supporting evidence regions, and yellow boxes represent encompass regions that jointly cover both answer and evidence, providing a more complete instruction-relevant area for supervision.

You are an OCR assistant.

Given a question, its correct answer, OCR lines (with center positions), and bounding box candidates,
select the **one best bounding box** that contains the correct answer.
Carefully read the question and answer. The question gives important context.
Use the (x, y) position of OCR lines to understand which content is spatially aligned with the answer.

Question: "{question}"
Answer: "{answer}"

Bounding Box Candidates: {bbox_candidates}

OCR entries (format: [Top, Left, Width, Height]): {ocr_entries}

Return format : "selected_answer_bbox": [x1, y1, x2, y2, x3, y3, x4, y4]

Only choose from the given candidates. Do not make up boxes.

Figure A8. Prompt for selecting a single bounding box among multiple candidates.

You are an OCR reasoning assistant.

Given a question, its answer, and OCR entries from an image (each includes text and bounding box),
explain briefly why the answer is correct based on the OCR content.

Question: {question}
Answer Text: {answer}

OCR entries (format: [Top, Left, Width, Height]): {ocr_entries}

Return format: "reasoning": "reasoning sentences"

Instructions:
- Use 3 short sentences.
- Focus on **why the answer logically follows from the question.**
- Highlight connections between question keywords and relevant OCR content.
- Do not invent or guess anything.

Figure A9. prompt for generating a reasoning sentence required to derive the correct answer.

You are an OCR reasoning assistant.

You are given:
- A question and its correct answer
- OCR entries extracted from the image
- A reasoning sentence that explains why the answer is correct (already inferred)
Only use the provided OCR entries to select bounding boxes. Do not modify or fabricate any text or boxes.

Question: {question}
Answer Text: {answer}
Reasoning: {reason sentence}

OCR entries (format: [Top, Left, Width, Height]): {ocr_entries}

Return format:
"answer": [[top, left, width, height],… [[top, left, width, height]]
"reasoning": [[top, left, width, height],..]

Your task is to return the bounding boxes of:
1. The answer
- If the answer consists of multiple words, search each word individually. Only combine words if:
  - They are spatially adjacent,
  - Appear in the correct reading order,
  - Answers are patially adjacent to reasons.
2. The supporting reason
- Select the most contextually appropriate one based on the question and layout.

Figure A10. Prompt for extracting and localizing the answer and supporting reasoning elements from OCR entries.

Please provide the bounding box of the elements necessary for reasoning to answer the given question.

Question: {question}
Answer: {answer}

1. Use the coordinate order `[x1, y1, x2, y2]`, where:
   - x1, y1 = top-left
   - x2, y2 = bottom-right

2. Provide **only the essential information**; keep verbosity to a minimum.
Example:
‘answer_bbox’: ‘[x1, y1, x2, y2]’
‘reasoning_bbox’: ‘[x3, y3, x4, y4]‘, ‘[x5, y5, x6, y6]’

Figure A11. Prompt for identifying instruction-relevant regions using a VLM

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[3] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *CVPR*, pages 13817–13827, 2024. 1

[4] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024. 2, 3

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

[6] 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. 1, 2, 3, 7

[7] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1, 2, 3, 4

[8] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2

[9] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3

[10] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *IEEE*, 2021. 1, 3, 4, 5, 7, 9

[11] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 1, 3, 4, 5, 6, 8

[12] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[13] 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. 2, 3

[14] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 2023. 1, 4, 5

[15] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 2, 3

[16] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 4

[17] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *CVPR*, 2025. 2, 3

[18] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 2, 3

[19] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, pages 15134–15186, 2025. 2, 3