

# FigEx2: Visual-Conditioned Panel Detection and Captioning for Scientific Compound Figures

Jifeng Song<sup>1,2</sup>, Arun Das<sup>2,3</sup>, Pan Wang<sup>1</sup>, Hui Ji<sup>4</sup>, Kun Zhao<sup>1</sup>, Yufei Huang<sup>1,2,3\*</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, University of Pittsburgh, USA

<sup>2</sup> Cancer Virology Program, UPMC Hillman Cancer Center, USA

<sup>3</sup> Department of Medicine, University of Pittsburgh, USA

<sup>4</sup> Department of Informatics and Networked Systems, University of Pittsburgh, USA  
{jis219, ard212, pan.wang, huj16, kun.zhao, yuh119}@pitt.edu

## Abstract

Scientific compound figures combine multiple labeled panels into a single image, and downstream pretraining and retrieval require panel-aligned visual-text pairs. However, in a PubMed Central (PMC)-scale crawl of 346,567 compound figures, 16.3% have no caption and are discarded by existing caption-decomposition pipelines. We propose FigEx2, a visual-conditioned framework that takes only a compound figure as input and jointly produces labeled panel boxes and panel-wise captions. FigEx2 introduces an Entity-Attention Kullback-Leibler (KL) regularizer that aligns the detector’s cross-attention with scientific entities annotated for each panel, providing a stable conditioning signal that also improves localization, and applies Group Relative Policy Optimization (GRPO) with a panel-level Entity-F1 reward to optimize scientific faithfulness. We curate BioSci-Fig-Cap for in-domain supervision and contribute physics and chemistry test suites for cross-disciplinary evaluation. FigEx2 achieves 0.751 mAP@0.5:0.95 on BioSci-Fig-Cap, and outperforms Qwen3-VL-8B by 6.80 Entity-F1 on MedICaT for captioning. It also transfers zero-shot to out-of-distribution domains. The source code is available at <https://github.com/Huang-AI4Medicine-Lab/FigEx2>.

## 1 Introduction

Scientific literature frequently employs compound figures (Lee et al., 2017; Subramanian et al., 2020) that aggregate multiple panels into a single image. Since each panel often depicts a distinct experiment, condition, or analysis, fine-grained understanding requires both localizing these panels and generating panel-specific captions. In practice, however, panel-level captions are frequently missing at scale. Our goal is to produce panel-aligned

\* Corresponding author

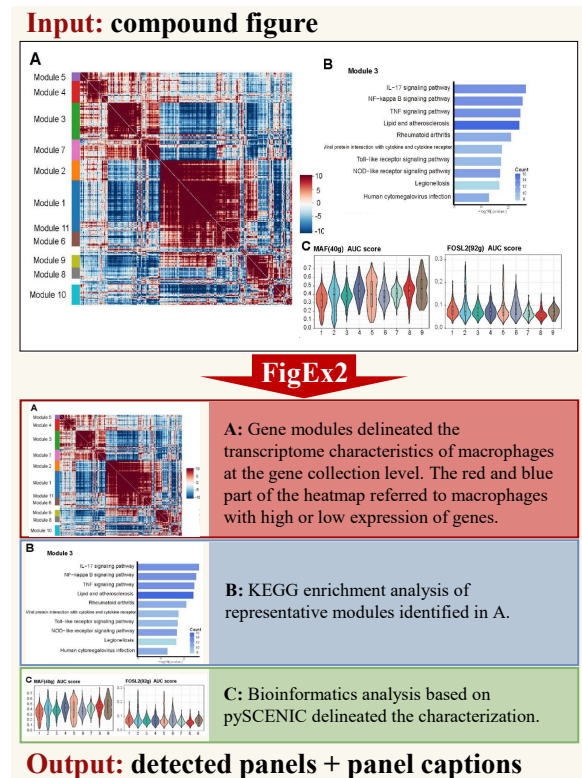


Figure 1: Task overview of FigEx2. Input is only the compound figure. FigEx2 detects labeled panels and generates panel-wise captions.

visual-text pairs from compound figures even in the absence of caption supervision, increasing effective dataset coverage for downstream representation learning.

Previous work (Song et al., 2025) addressed panel extraction by formulating it as a caption separation task: aligning segments of a panel-resolved caption to predicted visual boxes. However, this relies on a high-fidelity text-to-image mapping that is not always available in real-world pipelines (Clark and Divvala, 2016), and captions are often missing or limited to high-level summaries (Li et al., 2019) that fail to describe panels individually.

To resolve this, we propose a visual-conditioned

formulation that does not assume caption availability. As illustrated in Figure 1, the model takes only the compound figure as input, detects labeled panels, and generates a caption for each panel based on visual content, broadening the set of compound figures from which usable supervision can be extracted.

Based on this insight, we introduce FigEx2, a unified framework for visually grounded panel detection and captioning. Given only a compound figure, FigEx2 (i) localizes individual panels and (ii) generates panel-specific captions in a structured format that ends with a [DET] trigger (Wei et al., 2025) to link captioning with detection. A central design question is how the generated caption should condition the detector. Key biological entities such as genes, cell types, and diseases typically appear as printed annotations inside the corresponding panel; compared to the descriptive language that fills most of a caption, the caption tokens that match these on-panel entities provide a much more localized signal for detector queries.

We operationalize this idea through an Entity-Attention Kullback–Leibler (KL) regularizer that pulls the detector’s query-to-text cross-attention onto entity tokens during supervised training. After supervised fine-tuning, we further apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Liang, 2025) with a panel-level Entity-F1 reward, directly optimizing the scientific content that prior generic similarity or embedding-based evaluation signals across generation tasks (Rennie et al., 2017; Zhang et al., 2019; Hessel et al., 2021; Zhao et al., 2023, 2024) fail to target.

To support high-quality supervision, we introduce BioSci-Fig-Cap, a curated panel-caption benchmark derived from BioSci-Fig (Song et al., 2025), and contribute two out-of-domain test suites, PhysSci-Fig-Cap-Test (physics) and ChemSci-Fig-Cap-Test (chemistry), to benchmark cross-disciplinary transfer.

Our contributions are summarized as follows:

- We reformulate panel extraction from caption decomposition to visual-conditioned panel detection and captioning, where FigEx2 takes only a compound figure and produces labeled panel captions linked with detection via a [DET] trigger.
- We propose an Entity-Attention KL that aligns the detector’s query-to-text cross-attention with entity tokens in the generated caption,

providing a stable conditioning signal for the caption-to-detection interface that improves panel detection in addition to captioning. This contrasts with prior entity-grounded captioning, which uses entities only as RL rewards.

- We optimize panel captioning with GRPO using a panel-level Entity-F1 reward that directly targets scientific faithfulness, and report Entity-F1 as a new evaluation dimension complementary to lexical and embedding-based metrics.
- We curate BioSci-Fig-Cap for cleaner panel-level supervision and build two out-of-domain test suites in physics and chemistry to benchmark cross-domain transfer.

## 2 Related Work

### 2.1 Compound Figure Separation

Decomposing a scientific figure that contains multiple panels into labeled panels is essential for panel-level understanding and evaluation. This task has evolved from hand-crafted separator rules to pixel-level deep learning (ImageCLEF (De Herrera et al., 2016)), with CNNs (Tsutsui and Crandall, 2017) and label-guided pipelines (Jiang et al., 2021) significantly improving layout robustness. In the biomedical domain, SimCFS (Yao et al., 2022) further leverages synthetic composition to mitigate high annotation costs. However, existing methods focus primarily on geometric partitioning, often decoupling spatial separation from downstream semantic reasoning. Our work bridges this gap by integrating separation into a unified framework that links panel detection directly to generative reasoning.

### 2.2 Vision-Language Detection and Grounding

Large-scale Vision-Language Models (VLMs) have recently achieved remarkable success in multimodal understanding, demonstrating impressive capabilities in generating and interpreting content across visual and textual modalities (Wang et al., 2026b), with multimodal fusion and representation learning explored across tasks such as sentiment analysis (Wang et al., 2025, 2026a). Driven by these advancements, detection has evolved from contrastive region-text alignment (ViLD (Gu et al., 2021), GLIP (Li et al., 2022), DQ-DETR (Liu et al., 2023), Grounding DINO (Liu et al., 2024b)) to

generative grounding via coordinate tokenization (Pix2Seq (Chen et al., 2021), OFA (Wang et al., 2022)). Modern MLLMs extend this through external routing (DetGPT (Pi et al., 2023), Vision-LLM (Wang et al., 2023), LISA (Lai et al., 2024)) or native integration (Kosmos-2 (Peng et al., 2023), CogVLM (Wang et al., 2024), Lenna (Wei et al., 2025)). While foundation VLMs like Qwen2.5/3-VL (Bai et al., 2025b,a) show robust spatial reasoning (Li et al., 2025), RL-optimized localization (VLRM (Dzabraev et al., 2024)) remains nascent. Crucially, while these systems typically rely on textual priors, our work leverages DAB-DETR (Liu et al., 2022) to enable high-fidelity multi-instance panel localization in scientific compound figures, where localization must stay consistent with a structured, label-conditioned captioning output.

### 2.3 Entity-Grounded Caption Generation

Closely related to our setting, recent work in radiology report generation uses entity-level rewards to improve factual correctness. Building on the RadGraph dataset of clinical entities and relations (Jain et al., 2021), Delbrouck et al. (2022) introduce a RadGraph-based F1 reward that scores generated reports against reference reports on the set of clinical entities, and subsequent work extends this direction with semantic-graph combinations and multi-agent decomposition. These methods demonstrate that entity-level rewards substantially improve factual consistency over lexical or generic semantic rewards. Beyond reward design, radiology report generation has also been paired with more robust evaluation protocols (Zhao et al., 2026), underscoring that faithful generation and reliable evaluation are tightly coupled.

However, these methods use entities only at the reward level, a sequence-level signal applied during reinforcement learning, leaving the underlying representations unconstrained with respect to entity content. They also confine entity supervision to a single domain (chest X-ray), and only target caption faithfulness without considering how entity information could support spatial localization within the image.

## 3 Method

### 3.1 Overview

Given an input compound figure and a prompt, the captioning branch (a frozen vision-language model with trainable LoRA adapters) generates

panel-wise captions terminated by a [DET] token. The detection encoder independently encodes the same compound figure. Two caption-side features are then passed into the Fusion Module: the hidden state at [DET] ( $\mathbf{h}_{det}$ ), and the hidden states of the remaining generated caption tokens ( $\mathbf{H}_{cap}$ ). The Fusion Module integrates these with the visual features from the detection encoder, and the detection decoder predicts panel labels and bounding boxes.

### 3.2 Captioning to Detection Interface

Let the captioning branch produce a token sequence  $\hat{y}_{1:T}$  that includes a special trigger token [DET], with final-layer hidden states  $\mathbf{H} \in \mathbb{R}^{T \times d}$ , where  $T$  is the output length and  $d$  is the hidden size. We denote the hidden state at the [DET] position as  $\mathbf{h}_{det} \in \mathbb{R}^d$ .

We expose the detector to the full sequence of generated caption tokens (excluding [DET]). Let  $\mathbf{H}_{cap} \in \mathbb{R}^{T' \times d}$  denote the stacked hidden states of these  $T'$  tokens. We project these features into the detector’s text space:

$$\mathbf{F}_{txt} = \mathbf{H}_{cap} \mathbf{W}_{txt} \in \mathbb{R}^{T' \times d}, \quad (1)$$

where  $\mathbf{W}_{txt} \in \mathbb{R}^{d \times d}$  is a learned linear projection. Together with the [DET] hidden state  $\mathbf{h}_{det}$ ,  $\mathbf{F}_{txt}$  forms the caption-side input to the fusion module described next.

### 3.3 Fusion Module

The Fusion Module integrates the caption-side features  $\mathbf{h}_{det}$  and  $\mathbf{F}_{txt}$  into the detector’s object queries. Let  $\mathbf{Q} \in \mathbb{R}^{N_d \times d}$  be the  $N_d$  object queries initialized by the detection encoder, and let  $\mathbf{F}_{det} = \mathbf{h}_{det}^\top \in \mathbb{R}^{1 \times d}$  denote the detection-side interface token. We update the queries with two cross-attention steps:

$$\mathbf{Q}' = \text{ATTN}(\mathbf{Q}, \mathbf{F}_{det}), \quad (2)$$

$$\mathbf{Q}'' = \text{ATTN}(\mathbf{Q}', \mathbf{F}_{txt}), \quad (3)$$

where  $\text{ATTN}(\mathbf{Q}, \mathbf{F})$  is a standard multi-head attention block using  $\mathbf{Q}$  as queries and  $\mathbf{F}$  as keys/values. The detection decoder then consumes  $\mathbf{Q}''$  and predicts a set of panel bounding boxes with standard query-based decoding.

### 3.4 Entity-Attention Regularization

Feeding the detector caption-token hidden states gives the model maximum information, but provides no inductive bias for which tokens carry panel-specific scientific content. As illustrated in

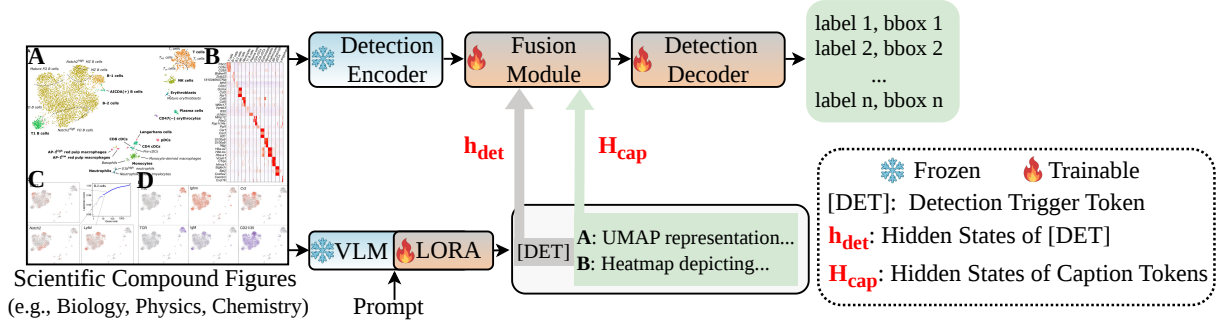


Figure 2: Overall framework of FigEx2. Given a compound figure, the captioning branch generates panel-wise captions ending with a [DET] token. The hidden state at [DET] ( $h_{det}$ ) and the hidden states of the remaining generated caption tokens ( $H_{cap}$ ) are passed to the Fusion Module, which integrates them with visual features from the detection encoder; the detection decoder then predicts panel labels and bounding boxes.

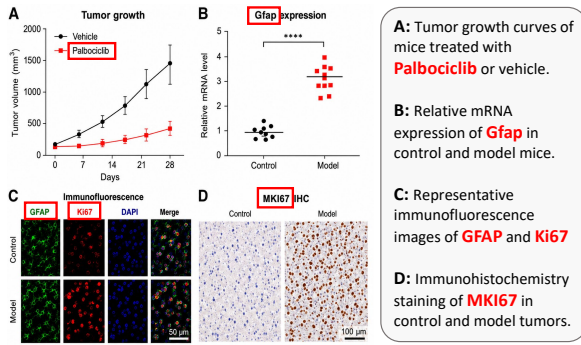


Figure 3: Scientific entities mentioned in panel captions (red) often appear as printed annotations inside the corresponding panel.

Figure 3, scientific entities often correspond to printed annotations inside the corresponding panel, making them natural visual anchors for detector queries. We therefore regularize the detector’s cross-attention to concentrate on entity tokens.

Let  $\alpha \in \mathbb{R}^{N_d \times T'}$  denote the softmax-normalized attention weights produced by Eq. (3), where  $\alpha_{i,t}$  is the attention from object query  $i$  to caption token  $t$ . To exploit this property, we introduce a binary Entity-Attention KL that explicitly directs cross-attention mass toward entity tokens. Let  $\bar{a}_t = \frac{1}{N_d} \sum_{i=1}^{N_d} \alpha_{i,t}$  denote the average attention from all object queries to caption token  $t$ , and define a target distribution  $w$  supported only on the entity subset  $\mathcal{T}_{ent}$ :

$$w_t = \begin{cases} 1/N_e & t \in \mathcal{T}_{ent}, \\ 0 & t \notin \mathcal{T}_{ent}. \end{cases} \quad (4)$$

That is, non-entity caption tokens receive zero target mass, while the  $N_e$  entity tokens share the mass uniformly. The regularizer is the KL divergence

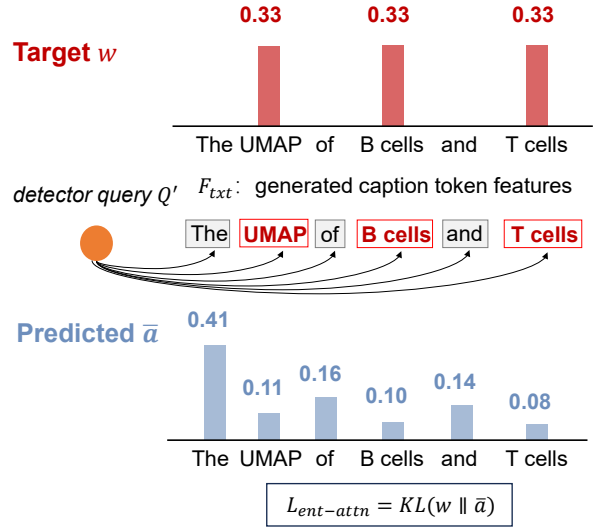


Figure 4: Entity-Attention KL on an example caption. The target  $w$  (top) places uniform mass on entity tokens and zero mass on non-entity tokens. The KL pulls the predicted attention  $\bar{a}$  (bottom) toward  $w$ .

from  $w$  to the predicted attention distribution:

$$\mathcal{L}_{ent-attn} = \text{KL}(w \parallel \bar{a}) = \sum_{t \in \mathcal{T}_{ent}} w_t \log \frac{w_t}{\bar{a}_t + \varepsilon}, \quad (5)$$

where  $\varepsilon$  is a small constant for numerical stability. Figure 4 illustrates the construction on an example caption.

### 3.5 Training Procedure

FigEx2 is trained to (i) generate structured panel captions and (ii) localize labeled panels. Given a compound figure  $x$ , the captioning branch generates a label-ordered target sequence  $y_{1:T}$  that ends with the trigger token [DET]; conditioned on the caption hidden states, the detector predicts a set of panel boxes and labels. We use a four-stage training schedule.

**Stage 1: Captioning pretraining.** We freeze the detector and update only the captioning branch with token-level cross entropy:

$$\mathcal{L}_{cap} = -\frac{1}{T} \sum_{t=1}^T \log p(y_t | y_{<t}, x), \quad (6)$$

where  $y_{1:T}$  is the structured ground-truth caption sequence of length  $T$ .

**Stage 2: Detection pretraining.** We freeze the captioning branch and train the detector with a standard set-based detection objective:

$$\mathcal{L}_{det} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{bbox} \mathcal{L}_{bbox} + \lambda_{giou} \mathcal{L}_{giou}, \quad (7)$$

where  $\mathcal{L}_{cls}$  is the classification loss,  $\mathcal{L}_{bbox}$  is the box regression loss,  $\mathcal{L}_{giou}$  is the generalized IoU loss, and  $\lambda_{cls}$ ,  $\lambda_{bbox}$ , and  $\lambda_{giou}$  are scalar weights.

**Stage 3: Joint supervised training with Entity-Attention KL.** Starting from the Stage 1 and Stage 2 checkpoints, we jointly optimize captioning, detection, and the Entity-Attention KL:

$$\mathcal{L}_{sft} = \lambda_{cap} \mathcal{L}_{cap} + \lambda_{det} \mathcal{L}_{det} + \lambda_{ent-attn} \mathcal{L}_{ent-attn}, \quad (8)$$

where  $\mathcal{L}_{cap}$ ,  $\mathcal{L}_{det}$ , and  $\mathcal{L}_{ent-attn}$  are the losses defined in Eqs. (6), (7), and (5), and  $\lambda_{cap}$ ,  $\lambda_{det}$ ,  $\lambda_{ent-attn}$  are scalar loss weights.

**Stage 4: Reward-augmented training with GRPO.** Starting from the Stage 3 checkpoint, we apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to the captioning branch with a LoRA adapter, while the detector remains frozen. For each input, we sample  $K$  candidates  $\{\hat{Y}_i\}_{i=1}^K$  from the behavior policy  $\pi_{\theta_{old}}$  (the captioning policy at the start of each GRPO update), compute their sequence-level rewards, and form a group-normalized advantage  $A_i$  by standardizing rewards within the group. The objective is a PPO-style clipped policy gradient with a KL penalty against the Stage 3 reference  $\pi_{ref}$ :

$$\mathcal{L}_{rl} = -\frac{1}{K} \sum_{i=1}^K \frac{1}{|\hat{Y}_i|} \sum_{t=1}^{|\hat{Y}_i|} \min(\rho_{i,t} A_i, \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon) A_i) + \beta \text{KL}(\pi_{\theta} \parallel \pi_{ref}), \quad (9)$$

where  $\hat{Y}_i$  is the  $i$ -th sampled caption with token length  $|\hat{Y}_i|$ ,  $\rho_{i,t} = \pi_{\theta}(\hat{y}_{i,t} | x, \hat{y}_{i,<t}) / \pi_{\theta_{old}}(\hat{y}_{i,t} | x, \hat{y}_{i,<t})$  is the per-token importance ratio between the current and behavior policies,  $\epsilon$  is the PPO clip range, and  $\beta$  controls the strength of the KL penalty.

| Dataset              | Split | Figures | Pairs |
|----------------------|-------|---------|-------|
| MedICaT              | Train | 1671    | 6359  |
|                      | Val   | 209     | 776   |
|                      | Test  | 209     | 840   |
| BioSci-Fig-Cap       | Train | 5290    | 30865 |
|                      | Val   | 652     | 3783  |
|                      | Test  | 658     | 3959  |
| PhysSci-Fig-Cap-Test | Test  | 200     | 778   |
| ChemSci-Fig-Cap-Test | Test  | 200     | 793   |

Table 1: Dataset splits after curation. We report the number of compound figures and panel-caption pairs.

**Entity-F1 reward.** The reward is panel-level Entity-F1. Each sampled caption is parsed into label-keyed subcaptions, and for each ground-truth panel label we run the dataset-specific entity extractor (Appendix A.3) on the predicted and reference subcaptions and compute set-level F1 between the resulting entity sets.

## 4 Benchmarks and Curation

We evaluate on four benchmarks. **BioSci-Fig-Cap** is a curated panel-caption benchmark we derive from BioSci-Fig with cleaner panel-level supervision; it is used for in-domain training and evaluation. **PhysSci-Fig-Cap-Test** (physics) and **ChemSci-Fig-Cap-Test** (chemistry) are two test-only benchmarks we curate to evaluate cross-disciplinary transfer. We also report results on **MedICaT** (Subramanian et al., 2020) for comparison with prior compound-figure protocols. BioSci-Fig-Cap is curated within each original BioSci-Fig subset, with its test set sourced only from the original test subset. Table 1 reports dataset sizes. Curation details and per-dataset entity extractors are provided in Appendices A and A.3.

## 5 Experiments

### 5.1 Evaluation Metrics

We report mAP@0.5:0.95 and mAP@0.5 for panel detection. For panel captioning, we use BLEU-4, ROUGE-L, METEOR, BERTScore, and Entity-F1. Entity-F1 is computed in the same way as the Stage 4 reward (Section 3.5), but is applied at evaluation time to compare predicted and reference panel captions. Predictions are aligned with ground-truth panels by label in reading order, with missing or extra pairs scored as 0; full details are in Appendix B.

| Model              | MedICaT     |              |              |              |              | BioSci-Fig-Cap |              |              |              |              |
|--------------------|-------------|--------------|--------------|--------------|--------------|----------------|--------------|--------------|--------------|--------------|
|                    | BLEU-4      | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    | BLEU-4         | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    |
| <b>Pre-trained</b> |             |              |              |              |              |                |              |              |              |              |
| Qwen2.5-VL-72B     | 2.08        | 15.91        | 9.95         | 83.67        | 9.95         | 1.71           | 23.62        | 14.74        | 82.23        | 21.36        |
| Llama3.2-90B       | 3.00        | 11.73        | 7.97         | 85.68        | 9.40         | 2.94           | 18.35        | 11.80        | 85.96        | 24.39        |
| gpt-5-nano         | 2.01        | 10.76        | 5.91         | 85.23        | 5.65         | 2.84           | 16.97        | 9.76         | 85.79        | 27.57        |
| gpt-5-mini         | 2.92        | 14.44        | 9.38         | 85.44        | 9.52         | 4.44           | 23.03        | 16.20        | 86.67        | 35.10        |
| gpt-5.2            | 2.43        | 13.25        | 7.44         | 85.54        | 7.15         | 4.07           | 23.85        | 14.63        | 87.12        | 35.56        |
| <b>Fine-tuned</b>  |             |              |              |              |              |                |              |              |              |              |
| LLaVA-NeXT         | 4.06        | 14.27        | 9.28         | 85.14        | 9.12         | 3.36           | 20.20        | 12.74        | 86.22        | 17.67        |
| LLaVA-Med          | 2.26        | 10.58        | 6.26         | 85.00        | 6.05         | 2.91           | 17.57        | 11.32        | 85.78        | 13.44        |
| Llama3.2-11B       | 4.80        | 16.38        | 11.39        | 85.62        | 8.43         | 3.88           | 21.44        | 15.79        | 86.33        | 22.58        |
| Qwen3-VL-8B        | 7.37        | 18.13        | 12.78        | 86.25        | 6.58         | 5.21           | 25.69        | 18.27        | 87.00        | 27.09        |
| FigEx2-8B          | <b>9.05</b> | <b>22.58</b> | <b>15.81</b> | <b>86.53</b> | <b>13.38</b> | <b>5.56</b>    | <b>26.63</b> | <b>19.05</b> | <b>87.38</b> | <b>36.36</b> |

Table 2: Panel captioning on MedICaT and BioSci-Fig-Cap.

## 5.2 Baselines

We compare against three baseline families: vision-only detectors (Fang et al., 2021) for panel detection, vision-language captioners (Liu et al., 2024a; Li et al., 2023; Grattafiori et al., 2024) prompted to produce labeled panel captions, and vision-language box generators prompted to output structured bounding boxes that we parse and evaluate as detections. Prompts are provided in Appendix C.

## 5.3 Implementation Details

FigEx2 combines Qwen3-VL-8B with a DAB-DETR detector and is trained with the four-stage procedure in Section 3.5. We train on 4 NVIDIA A100 80GB GPUs and run inference on NVIDIA L40S 48GB and NVIDIA A100 40GB GPUs. The hyperparameter settings are in Appendix D.

## 5.4 Comparison with Existing Methods

**Panel Captioning.** Table 2 reports panel captioning results. FigEx2 achieves the best overall performance on both datasets across all five metrics. On MedICaT, FigEx2 improves ROUGE-L from 18.13 to 22.58 and METEOR from 12.78 to 15.81 over Qwen3-VL-8B, the strongest fine-tuned baseline, while raising Entity-F1 from 6.58 to 13.38, more than doubling the Entity-F1 of Qwen3-VL-8B. On BioSci-Fig-Cap, FigEx2 again improves all five metrics over Qwen3-VL-8B, with the most pronounced gain on Entity-F1 (27.09 to 36.36). The disproportionately large gains on Entity-F1 indicate that FigEx2’s predictions recover scientific entities much more reliably than baseline captions of comparable lexical quality.

**Panel Detection.** Table 3 reports panel detection results under the LVIS evaluator (Gupta et al., 2019; Tran, 2025). On MedICaT, FigEx2 achieves

| Model                 | mAP@0.5:0.95 | mAP@0.5      |
|-----------------------|--------------|--------------|
| <b>MedICaT</b>        |              |              |
| YOLOS-Ti              | 0.155        | 0.183        |
| YOLOS-S               | 0.156        | 0.188        |
| YOLOS-B               | 0.150        | 0.170        |
| DAB-DETR              | 0.165        | 0.181        |
| Qwen3-VL-8B           | 0.230        | 0.309        |
| FigEx2-8B             | <b>0.360</b> | <b>0.405</b> |
| <b>BioSci-Fig-Cap</b> |              |              |
| YOLOS-Ti              | 0.100        | 0.129        |
| YOLOS-S               | 0.196        | 0.248        |
| YOLOS-B               | 0.448        | 0.534        |
| DAB-DETR              | 0.697        | 0.830        |
| Qwen3-VL-8B           | 0.439        | 0.578        |
| FigEx2-8B             | <b>0.751</b> | <b>0.906</b> |

Table 3: Panel detection on MedICaT and BioSci-Fig-Cap.

the highest mAP, improving mAP@0.5:0.95 from 0.230 to 0.360 over prompt-based box generation with Qwen3-VL-8B, and mAP@0.5 from 0.309 to 0.405. On BioSci-Fig-Cap, FigEx2 remains the top method, improving mAP@0.5:0.95 from 0.697 to 0.751 over DAB-DETR and mAP@0.5 from 0.830 to 0.906. The detection gains are notable given that DAB-DETR is already a strong specialized detector on BioSci-Fig-Cap (0.697 mAP@0.5:0.95); coupling the same detector backbone with caption-conditioned object queries lifts panel localization across both datasets.

## 5.5 Ablation Study

Table 4 reports our ablation across two design choices: the Entity-Attention KL on detector cross-attention, and the GRPO stage. Adding the Entity-Attention KL improves panel detection mAP@0.5:0.95 from 0.272 to 0.353 on MedICaT and from 0.722 to 0.740 on BioSci-Fig-Cap, and

| Variant                                          | Panel Detection |              | Panel Captioning |              |              |              |              |
|--------------------------------------------------|-----------------|--------------|------------------|--------------|--------------|--------------|--------------|
|                                                  | mAP@0.5:0.95    | mAP@0.5      | BLEU-4           | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    |
| <b>MedICaT</b>                                   |                 |              |                  |              |              |              |              |
| (1) SFT only, no Entity-Attention KL             | 0.272           | 0.301        | 8.11             | 21.17        | 14.58        | 86.45        | 7.45         |
| (2) SFT + Entity-Attention KL                    | 0.353           | 0.403        | 8.54             | 21.63        | 15.44        | <b>86.54</b> | 9.88         |
| (3) SFT + Entity-Attention KL + GRPO (Entity-F1) | <b>0.360</b>    | <b>0.405</b> | <b>9.05</b>      | <b>22.58</b> | <b>15.81</b> | 86.53        | <b>13.38</b> |
| <b>BioSci-Fig-Cap</b>                            |                 |              |                  |              |              |              |              |
| (1) SFT only, no Entity-Attention KL             | 0.722           | 0.865        | 5.35             | 25.75        | 18.25        | 87.15        | 31.85        |
| (2) SFT + Entity-Attention KL                    | 0.740           | 0.881        | 5.45             | 26.16        | 18.89        | 87.27        | 34.63        |
| (3) SFT + Entity-Attention KL + GRPO (Entity-F1) | <b>0.751</b>    | <b>0.906</b> | <b>5.56</b>      | <b>26.63</b> | <b>19.05</b> | <b>87.38</b> | <b>36.36</b> |

Table 4: Ablation of FigEx2 on MedICaT and BioSci-Fig-Cap. (1) SFT-only baseline; (2) adds the Entity-Attention KL; (3) further adds GRPO with the panel-level Entity-F1 reward.

| Model              | PhysSci-Fig-Cap-Test |              |              |              |              | ChemSci-Fig-Cap-Test |              |              |              |              |
|--------------------|----------------------|--------------|--------------|--------------|--------------|----------------------|--------------|--------------|--------------|--------------|
|                    | BLEU-4               | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    | BLEU-4               | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    |
| <b>Pre-trained</b> |                      |              |              |              |              |                      |              |              |              |              |
| Qwen2.5-VL-72B     | 1.48                 | 17.28        | 9.97         | 82.11        | 9.29         | 1.38                 | 17.68        | 10.37        | 81.64        | 8.38         |
| Llama3.2-90B       | 2.09                 | 12.84        | 8.05         | 84.76        | 11.85        | 2.21                 | 14.61        | 8.99         | 84.45        | 12.98        |
| gpt-5-nano         | 1.85                 | 11.56        | 6.33         | 84.37        | 9.75         | 1.73                 | 11.60        | 6.09         | 84.15        | 13.00        |
| gpt-5-mini         | <b>2.81</b>          | 16.11        | <b>10.56</b> | 84.81        | 14.55        | 2.82                 | 16.41        | 10.44        | 84.47        | 14.34        |
| gpt-5.2            | 2.77                 | 16.71        | 9.75         | 85.34        | 15.55        | 2.53                 | 17.16        | 9.57         | 85.12        | 15.03        |
| <b>Fine-tuned</b>  |                      |              |              |              |              |                      |              |              |              |              |
| LLaVA-NeXT         | 1.48                 | 11.25        | 6.07         | 84.45        | 6.03         | 1.93                 | 14.54        | 8.56         | 84.50        | 5.16         |
| LLaVA-Med          | 1.56                 | 10.72        | 6.04         | 84.67        | 1.64         | 1.57                 | 12.83        | 7.75         | 84.43        | 4.44         |
| Llama3.2-11B       | 2.17                 | 15.12        | 9.09         | 85.15        | 8.93         | 2.37                 | 16.56        | 10.29        | 84.84        | 13.10        |
| Qwen3-VL-8B        | 2.75                 | <b>17.75</b> | 9.98         | <b>85.57</b> | 12.28        | 3.05                 | 18.93        | 11.58        | 85.18        | 14.78        |
| FigEx2-8B          | 2.65                 | 16.90        | 9.39         | 85.50        | <b>16.38</b> | <b>3.05</b>          | <b>19.92</b> | <b>12.08</b> | <b>85.39</b> | <b>15.29</b> |

Table 5: Cross-domain panel captioning on PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test.

improves panel captioning across all metrics, with Entity-F1 rising from 7.45 to 9.88 on MedICaT and from 31.85 to 34.63 on BioSci-Fig-Cap. Notably the KL also benefits captioning even though it is applied only on the detector side: gradients flow back through the shared caption features, signaling to the captioning branch that entity tokens carry weight. Adding GRPO with the Entity-F1 reward further raises Entity-F1 to 13.38 on MedICaT and 36.36 on BioSci-Fig-Cap, with consistent gains on BLEU-4, ROUGE-L, and METEOR; detection improves as well, reaching 0.751 mAP@0.5:0.95 on BioSci-Fig-Cap, since richer entity content in the generated captions feeds cleaner anchors back to the detector. The two components therefore form a positive loop: the KL makes the detector rely on entity tokens, which pushes captioning to produce them more reliably, and GRPO directly reinforces this behavior. Captioning and detection co-improve rather than trade off.

## 5.6 Cross-domain Generalization

We evaluate robustness under domain shift by transferring models fine-tuned on BioSci-Fig-Cap to PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test

| Model                       | mAP@0.5:0.95 | mAP@0.5      |
|-----------------------------|--------------|--------------|
| <b>PhysSci-Fig-Cap-Test</b> |              |              |
| YOLOS-Ti                    | 0.087        | 0.107        |
| YOLOS-S                     | 0.136        | 0.164        |
| YOLOS-B                     | 0.237        | 0.280        |
| DAB-DETR                    | 0.372        | 0.418        |
| Qwen3-VL-8B                 | 0.374        | 0.447        |
| FigEx2-8B                   | <b>0.453</b> | <b>0.515</b> |
| <b>ChemSci-Fig-Cap-Test</b> |              |              |
| YOLOS-Ti                    | 0.109        | 0.144        |
| YOLOS-S                     | 0.185        | 0.231        |
| YOLOS-B                     | 0.306        | 0.376        |
| DAB-DETR                    | 0.357        | 0.411        |
| Qwen3-VL-8B                 | 0.322        | 0.431        |
| FigEx2-8B                   | <b>0.446</b> | <b>0.511</b> |

Table 6: Cross-domain panel detection on PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test.

with no additional training. All detection results are obtained with detectors fine-tuned on BioSci-Fig-Cap, while captioning compares (i) pre-trained vision-language models and (ii) models fine-tuned on BioSci-Fig-Cap.

**Cross-domain panel captioning.** Table 5 summarizes cross-domain panel captioning results. On ChemSci-Fig-Cap-Test, which shares substantial

| Few-shot           | PhysSci-Fig-Cap-Test |              |              |              |              | ChemSci-Fig-Cap-Test |              |              |              |              |
|--------------------|----------------------|--------------|--------------|--------------|--------------|----------------------|--------------|--------------|--------------|--------------|
|                    | BLEU-4               | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    | BLEU-4               | ROUGE-L      | METEOR       | BERTScore    | Entity-F1    |
| <b>Qwen3-VL-8B</b> |                      |              |              |              |              |                      |              |              |              |              |
| 1-shot             | 20.10                | 33.48        | 28.60        | 89.57        | 16.73        | 32.63                | 46.17        | 42.46        | 91.24        | 28.23        |
| 2-shot             | <b>39.84</b>         | <b>53.19</b> | <b>49.90</b> | 92.06        | 31.76        | 42.56                | 55.57        | 53.05        | 92.79        | 40.95        |
| <b>FigEx2-8B</b>   |                      |              |              |              |              |                      |              |              |              |              |
| 1-shot             | 19.97                | 32.83        | 27.94        | 89.76        | 19.72        | 32.48                | 45.58        | 41.91        | 91.68        | 27.12        |
| 2-shot             | 38.36                | 50.99        | 48.06        | <b>92.22</b> | <b>35.48</b> | <b>46.90</b>         | <b>59.96</b> | <b>56.90</b> | <b>93.39</b> | <b>43.81</b> |

Table 7: Few-shot prompting for cross-domain panel captioning on PhysSci-Fig-Cap-Test & ChemSci-Fig-Cap-Test.

domain knowledge with the biomedical training domain, FigEx2 achieves the best or tied-best result across all five metrics, improving ROUGE-L from 18.93 to 19.92, METEOR from 11.58 to 12.08, and Entity-F1 from 14.78 to 15.29 over Qwen3-VL-8B. On the more distant PhysSci-Fig-Cap-Test, FigEx2 trades a small amount of lexical overlap for a substantial gain in Entity-F1 from 12.28 to 16.38. The contrast suggests that domain proximity benefits lexical generation, but the Entity-Attention KL keeps the model focused on scientifically meaningful tokens even under stronger domain shift, so the entity-level advantage persists across both domains.

**Cross-domain panel detection.** Table 6 reports cross-domain detection performance. Among classical detectors, DAB-DETR is the strongest baseline, while FigEx2 further improves detection in both domains. On PhysSci-Fig-Cap-Test, FigEx2 improves mAP@0.5:0.95 from 0.372 to 0.453 over DAB-DETR; on ChemSci-Fig-Cap-Test, it improves mAP@0.5:0.95 from 0.357 to 0.446 over DAB-DETR. Two factors support this transfer: scientific compound figures share similar spatial layouts across disciplines, so the panel-localization skill carries over; and the attention-level inductive bias from the Entity-Attention KL transfers well, with the detector remaining focused on entity tokens even on out-of-domain captions.

**Few-shot prompting.** We evaluate few-shot prompting by providing ground-truth subcaptions for early panels as exemplars at inference time, using panel A for 1-shot and panels A–B for 2-shot, then evaluating only panels C and later. Table 7 shows that adding exemplars consistently improves captioning quality in both domains, and the gain is especially pronounced for FigEx2: a few example captions provide the domain vocabulary FigEx2 lacked under zero-shot transfer, lifting BERTScore above the Qwen3-VL-8B baseline at 2-shot (92.22

against 92.06 on PhysSci-Fig-Cap-Test) and indicating that few-shot prompting effectively alleviates the domain-knowledge gap. On ChemSci-Fig-Cap-Test, FigEx2-8B BLEU-4 increases to 46.90 in 2-shot against 42.56 for Qwen3-VL-8B, and 2-shot Entity-F1 reaches 43.81 against 40.95. On PhysSci-Fig-Cap-Test, FigEx2 achieves the best 2-shot Entity-F1 of 35.48 against 31.76. Overall, FigEx2 obtains the strongest entity-level few-shot performance in both domains.

**Discussion.** A consistent pattern holds across settings: FigEx2’s lexical scores scale with domain proximity, while its entity-level advantage persists on every benchmark. This is the intended effect of the Entity-Attention KL, which anchors the detector on entity tokens rather than generic descriptive language and thus preserves scientific specificity under domain shift.

## 6 Conclusion

We propose FigEx2, a visual-conditioned framework that jointly detects labeled panels and generates panel-wise captions directly from compound figures, converting otherwise unusable figures into aligned panel-text pairs for downstream pretraining and retrieval. FigEx2 links captioning to detection through a [DET] trigger and feeds the detector hidden states of generated caption tokens, while an Entity-Attention KL aligns the detector’s query-to-text attention with entity tokens, providing a stable conditioning signal that improves panel detection in addition to captioning. GRPO with a panel-level Entity-F1 reward then directly optimizes the caption policy for scientific faithfulness. On BioSci-Fig-Cap and our physics/chemistry test suites, FigEx2 improves both detection and captioning over strong fine-tuned baselines, with the largest gains on Entity-F1. Across domain shifts, detection gains hold uniformly thanks to shared spatial layouts and the transferable Entity-Attention

KL, while captioning gains scale with domain proximity on lexical metrics but the entity-level advantage persists everywhere. Few-shot prompting further alleviates the remaining domain-knowledge gap, showing that FigEx2 can adapt to new scientific conventions with minimal supervision.

## Ethics Statement

BioSci-Fig-Cap and our physics and chemistry test suites are derived from open-access sources and do not contain private or sensitive personal data. Curation and caption refinement were performed by trained annotators with domain expertise. While pretrained models may have seen parts of the scientific literature, we mitigate potential leakage by evaluating on held-out splits and reporting results transparently.

## Limitations

The detection branch relies on DAB-DETR, which predicts axis-aligned bounding boxes. While this suits grid-like layouts, it may be too rigid for non-rectangular regions or heavily overlapping panels common in some scientific domains; segmentation-based delineation is a natural next step. On the data side, our benchmarks cover biomedical, physics, and chemistry literature, but other scientific domains such as materials science, geosciences, and engineering remain uncovered; broadening BioSci-Fig-Cap-style curation to these domains is left to future work.

## Acknowledgments

This study was supported by grants from the National Institutes of Health U01CA279618 and R21GM155774 to Y. Huang and in part by the University of Pittsburgh Center for Research Computing, RRID:SCR\_022735. Specifically, this work used the HTC cluster, which is supported by S10OD028483.

## References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025a. *Qwen3-vl technical report*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu,

Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. *Qwen2.5-vl technical report*. *arXiv preprint arXiv:2502.13923*.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152.

Alba G Seco De Herrera, Stefano Bromuri, Roger Schaer, and Henning Müller. 2016. Overview of the medical tasks in imageclef 2016. *CLEF working notes*. Evora, Portugal.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Maksim Dzabaraev, Alexander Kunitsyn, and Andrei Ivaniuta. 2024. Vlm: Vision-language models act as reward models for image captioning. *arXiv preprint arXiv:2404.01911*.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren,

- Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
- Weixin Jiang, Eric Schwenker, Trevor Spreadbury, Nicola Ferrier, Maria KY Chan, and Oliver Cossairt. 2021. A two-stage framework for compound figure separation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1204–1208. IEEE.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Po-shen Lee, Jevin D West, and Bill Howe. 2017. Vizmetrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 4(1):117–129.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyu Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975.
- Pengyuan Li, Xiangying Jiang, and Hagit Shatkay. 2019. Figure and caption extraction from biomedical documents. *Bioinformatics*, 35(21):4381–4388.
- Yifan Li, Lichi Li, Anh Dao, Xinyu Zhou, Wenjun Huang, Tianyi Ma, Yicheng Qiao, Zheda Mai, Daeun Lee, Zichen Chen, and 1 others. 2025. Industry-nav: Exploring spatial reasoning of embodied agents in dynamic industrial navigation. *arXiv preprint arXiv:2511.17384*.
- Xu Liang. 2025. Group relative policy optimization for image captioning. *arXiv preprint arXiv:2503.01333*.
- Haotian Liu, Chunyu Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.
- Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1728–1736.
- Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and 1 others. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kueneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.
- Jifeng Song, Arun Das, Ge Cui, and Yufei Huang. 2025. Figex: Aligned extraction of scientific figures and captions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16558–16571.
- Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine Van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Mediat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120.
- Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. [Label Studio: Data labeling software](https://github.com/HumanSignal/label-studio). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Phi Vu Tran. 2025. Simltd: Simple supervised and semi-supervised long-tailed object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4672–4681.
- Satoshi Tsutsui and David J Crandall. 2017. A data driven approach for compound figure separation using convolutional neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 533–540. IEEE.
- Pan Wang, Lipeng Ke, Huajun Ying, Pritish Mohapatra, Rohan Sarkar, Suresh Lakhani, Sankar Venkataraman, and Jingtong Hu. 2026a. ConsMSA: Semantic distribution consistency learning for multimodal sentiment analysis. In *Proceedings of the 43rd International Conference on Machine Learning*.
- Pan Wang, Siwei Song, Hui Ji, Siqi Cao, Heng Yu, Zhijian Liu, Huanrui Yang, Yingyan Celine Lin, Beidi Chen, Mohit Bansal, and 1 others. 2026b. From models to systems: A comprehensive survey of efficient multimodal learning. *Transactions on Machine Learning Research*.
- Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 21180–21188.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.
- Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513.
- Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. 2025. Lenna: Language enhanced reasoning detection assistant. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Tianyuan Yao, Chang Qu, Jun Long, Quan Liu, Ruining Deng, Yuanhan Tian, Jiachen Xu, Aadarsh Jha, Zuhayr Asad, Shunxing Bao, and 1 others. 2022. Compound figure separation of biomedical images: Mining large datasets for self-supervised learning. *The journal of machine learning for biomedical imaging*, 1:025.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Kun Zhao, Chenghao Xiao, Sixing Yan, Haoteng Tang, William K Cheung, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2026. X-ray made simple: Lay radiology report generation and robust evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34583–34598.
- Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574.
- Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. 2024. Slide: A framework integrating small and large language models for open-domain dialogues evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15421–15435.

## A Additional Benchmark Details

This appendix provides curation and annotation details for the panel-caption benchmarks used in Section 4.

### A.1 Curation Details

BioSci-Fig-Cap is curated directly from BioSci-Fig to improve panel-level supervision for labeled, panel-specific captioning. We start from the panel captions and label annotations provided in BioSci-Fig, and apply additional filtering, cleaning, and targeted expert revision to reduce noise and improve terminology consistency and scientific plausibility.

**Starting point.** We inherit the BioSci-Fig panel boxes/labels and their associated panel captions as the initial supervision. We normalize alphabetic labels to uppercase (A–Z) and remove panels without a valid alphabetic label.

**Filtering.** We remove figures where any retained panel caption is extremely short (fewer than five tokens after tokenization), as such lines are typically underspecified. We also remove panel captions that

are clearly non-informative, including boilerplate-dominated text and generic references that do not convey panel-specific content.

**Cleaning.** We remove noisy or misaligned panel captions, including mislabeled entries and captions that do not correspond to the assigned panel content. We also remove duplicates and near-duplicates that arise from repeated or redundant panel descriptions.

**Expert revision.** The biology-related panel captions are reviewed by researchers with computational biology and machine learning backgrounds. Revisions focus on improving clarity, terminology consistency, and obvious scientific errors while preserving the intended meaning. We avoid edits that would introduce new claims not supported by the figure.

**Cross-domain test set construction.** PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test are constructed as test-only benchmarks using the FigEx caption decomposition and label association pipeline (Song et al., 2025). We select compound figures with high-quality full captions and apply the same FigEx-style procedure to derive label-conditioned panel lines and associate them with alphabetic panel labels.

## A.2 Annotation and Quality Control

We use Label Studio for manual review and quality control (Tkachenko et al., 2020-2025).

**Annotation instructions.** Annotators draw a bounding box around each panel and record its alphabetic label. If a figure contains multiple panels with the same label, annotators reject the compound figure.

**Overall accuracy of curated benchmarks.** We assess curated benchmark quality with a sample-level overall accuracy: for each benchmark, an annotator inspects 200 samples (the full test set for PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test; 200 randomly sampled figures for BioSci-Fig-Cap) and marks each as correct only if its panel boxes and panel captions are judged acceptable. Table 8 reports the resulting overall accuracy.

## A.3 Entity Extractors and Matching Policy

Each benchmark ships with per-panel entity tokens used by the Entity-Attention KL (Section 3.4),

| Dataset              | Overall Accuracy (%) |
|----------------------|----------------------|
| BioSci-Fig-Cap       | 97.0                 |
| PhysSci-Fig-Cap-Test | 95.5                 |
| ChemSci-Fig-Cap-Test | 94.5                 |

Table 8: Overall accuracy on the curated benchmarks.

the Entity-F1 reward (Section 3.5), and the Entity-F1 evaluation metric (Section 5.1). The benchmarks use different domain-specialized NER models. For the biomedical benchmarks (BioSci-Fig-Cap and MedICaT), entities are extracted with BERN2 (Sung et al., 2022), and Entity-F1 is computed by matching entities on BERN2’s normalized identifiers, which is robust to surface variation (e.g., synonyms, abbreviations) within a single normalized concept. For ChemSci-Fig-Cap-Test we use alvaroalon2/biobert\_chemical\_ner, a community fine-tuned BioBERT (Lee et al., 2020) chemical-entity tagger, and for PhysSci-Fig-Cap-Test we use pranav-s/PolymerNER (Shetty et al., 2023), a MaterialsBERT-based model for polymer and materials entities. Since neither provides a normalized identifier scheme, Entity-F1 on these two benchmarks matches on the surface mention strings produced by the extractor, lower-cased and whitespace-normalized.

## B Evaluation Protocol Details

**Occurrence-level label alignment.** Ground-truth panels are ordered by the ground-truth reading order. For each label  $\ell$ , let  $\mathcal{R}_\ell = [r_{\ell,1}, \dots, r_{\ell,n_\ell}]$  be reference captions for all occurrences of  $\ell$ , and let  $\mathcal{P}_\ell = [p_{\ell,1}, \dots, p_{\ell,m_\ell}]$  be predicted captions with label  $\ell$ . We match occurrences by index:  $r_{\ell,k}$  pairs with  $p_{\ell,k}$  if  $k \leq m_\ell$ , otherwise the prediction for this occurrence is missing; if  $m_\ell > n_\ell$ , the extra predicted occurrences are treated as extra pairs.

**Union pairing and zero scoring.** We form evaluation pairs over the union of ground-truth and predicted labeled lines. A missing pair or an extra pair receives a score of 0 for all caption metrics, including Entity-F1.

**Entity-F1 computation.** For each evaluation pair  $(\hat{c}, c)$ , we run the dataset-specific entity extractor (Appendix A.3) on both sides to obtain entity sets  $E(\hat{c})$  and  $E(c)$ , and compute set-level F1 under that dataset’s matching policy (ID match for biomedical benchmarks; mention match for chemistry and physics). Empty entity sets on both sides

give an F1 of 1; empty on only one side gives 0.

**Aggregation.** For each figure, we compute caption metrics on all evaluation pairs and average within the figure as  $\text{Score}(\text{figure}) = \frac{1}{N} \sum_{i=1}^N s(p_i, r_i)$ , then report the dataset score as the mean over figures.

## C Prompts

### C.1 Prompt Used for Panel Captioning

#### PROMPT\_CAPTIONING

You are given a scientific compound figure.

Task: detect subfigures and, for each detected subfigure that shows a visible alphabetic label A to Z or a to z, write exactly one short scientific caption.

Formatting rules:

- 1) Output one line per subfigure in ascending label order A, B, C, ...
- 2) Use uppercase labels and the exact format: "A: <caption>".
- 3) After listing all subfigure captions, output a single [DET] token on a NEW line. Return ONLY the caption lines followed by the final [DET]; no extra text.

### C.2 Prompt Used for Panel Detection

#### PROMPT\_DETECTION

You are given a scientific compound figure containing multiple sub-panels A, B, C, ...

Detect all sub-panels and output ONLY a JSON array.

Each element must be an object with fields:

- "class": an integer in [0, 25] where A maps to 0, B maps to 1, ..., Z maps to 25
- "bbox\_2d": [x\_min, y\_min, x\_max, y\_max] using normalized coordinates in range [0, 1000]

Rules:

- Do not output any text outside the JSON array.
- bbox\_2d must satisfy  $x_{\min} < x_{\max}$  and  $y_{\min} < y_{\max}$ .
- Include every detected panel; multiple boxes may share the same class.

## D Additional Implementation Details

We use LoRA with rank 16, alpha 32, dropout 0.05, and target modules q\_proj, k\_proj, v\_proj, o\_proj, gate\_proj, up\_proj, down\_proj as a parameter-efficient adaptation method. We train with batch size 1 and gradient accumulation 16, using a maximum sequence length of 8192 tokens throughout training and inference.

We report BERTScore as  $100 \times \text{F1}$ .

## E Examples

Figures 5–7 show qualitative examples with the compound image and the panel boxes and captions.

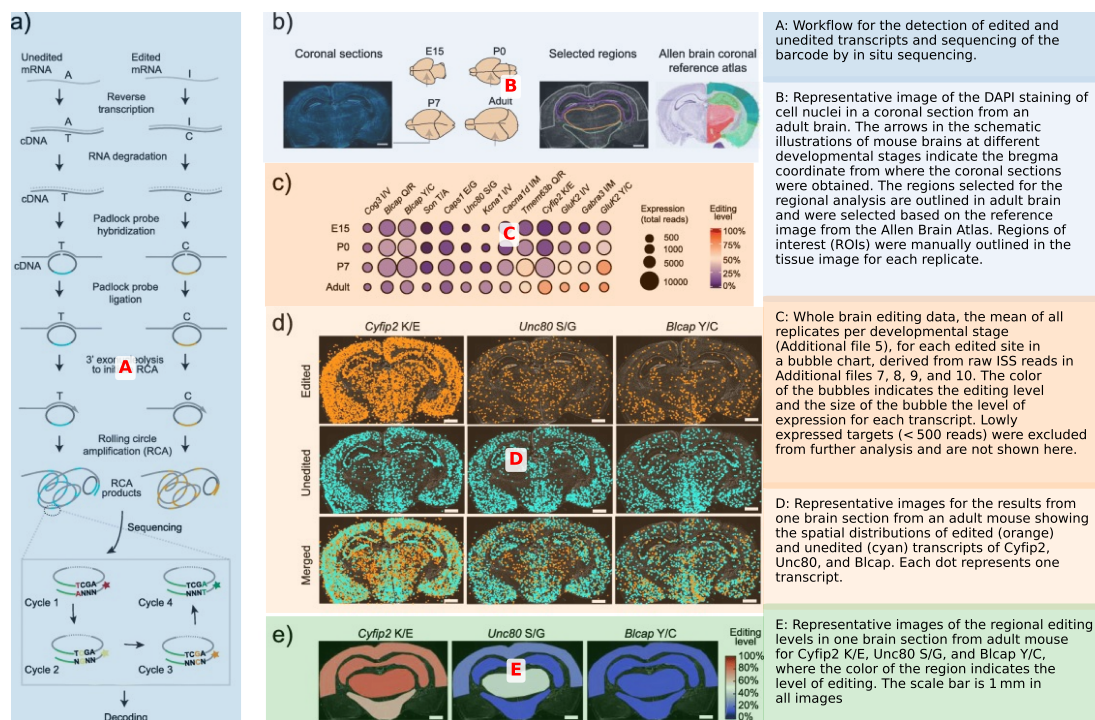


Figure 5: Compound figure with panel boxes with aligned subcaptions (example 1).

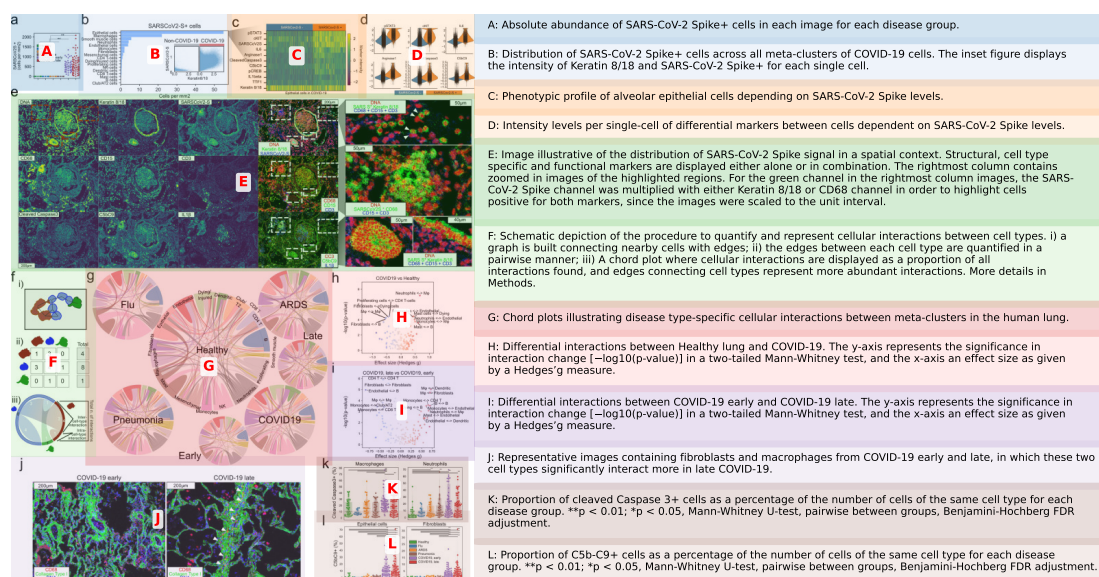


Figure 6: Compound figure with panel boxes with aligned subcaptions (example 2).

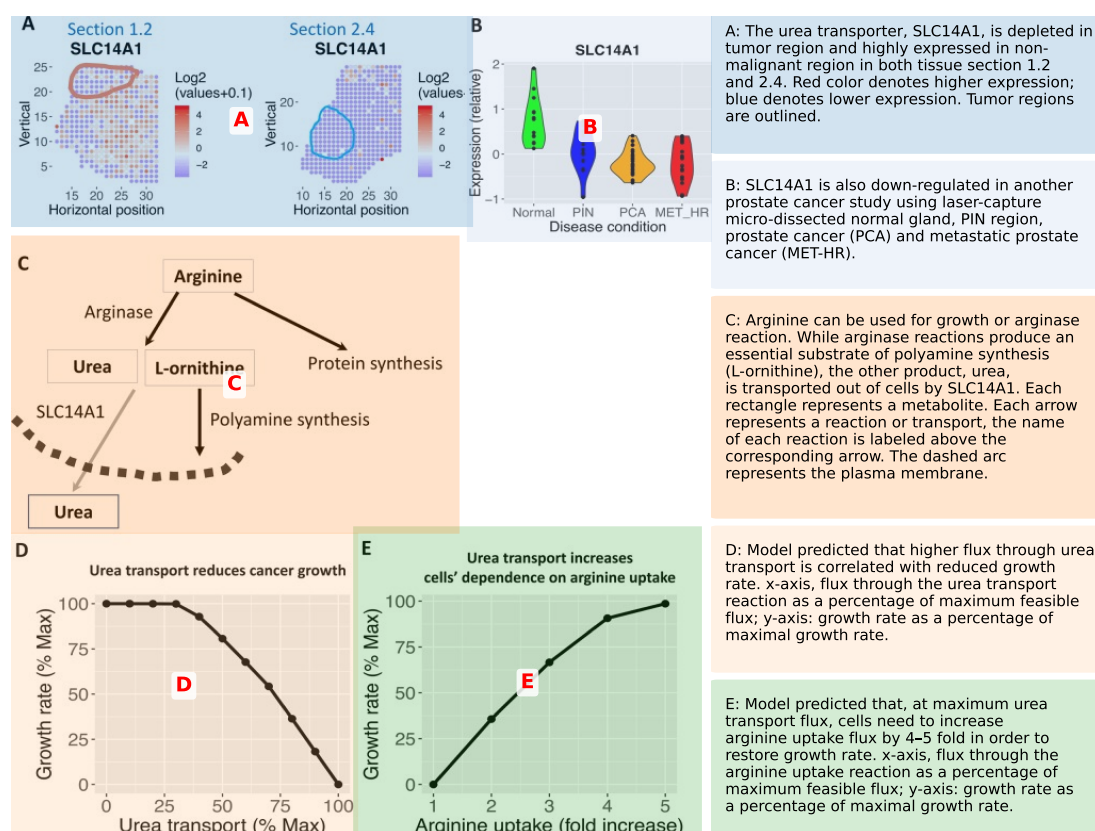


Figure 7: Compound figure with panel boxes with aligned subcaptions (example 3).