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

Anonymous ACL submission

Abstract

Scientific compound figures combine multiple labeled panels into a single image. However, in a PMC-scale crawl of 346,567 compound figures, 16.3% have no caption and 1.8% only have captions shorter than ten words, causing them to be discarded by existing caption-decomposition pipelines. We propose FigEx2, a visual-conditioned framework that localizes panels and generates panel-wise captions directly from the image, converting otherwise unusable figures into aligned panel-text pairs for downstream pretraining and retrieval. To mitigate linguistic variance in open-ended captioning, we introduce a noise-aware gated fusion module that adaptively controls how caption features condition the detection query space, and employ a staged SFT+RL strategy with CLIP-based alignment and BERTScore-based semantic rewards. To support high-quality supervision, we curate BioSci-Fig-Cap, a refined benchmark for panel-level grounding, alongside cross-disciplinary test suites in physics and chemistry. FigEx2 achieves 0.728 mAP@0.5:0.95 for detection, outperforms Qwen3-VL-8B by 0.44 in METEOR and 0.22 in BERTScore, and transfers zero-shot to out-of-distribution scientific domains without fine-tuning.

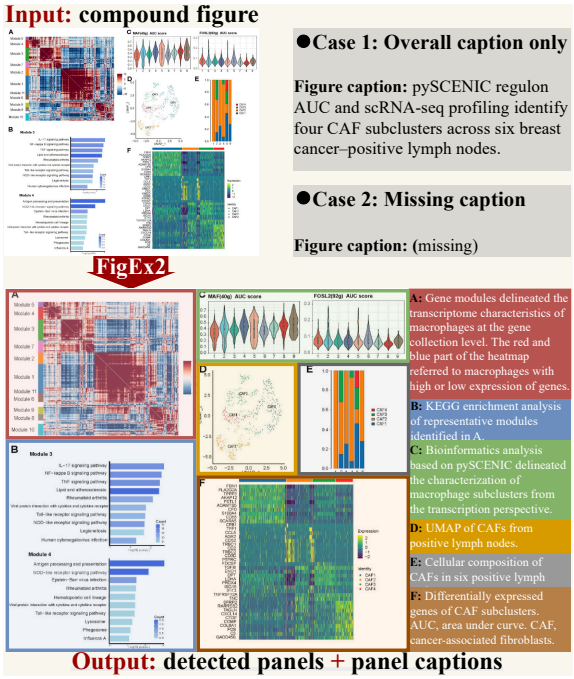


Figure 1: Task overview of FigEx2. Input is only the compound figure. FigEx2 detects labeled panels and generates panel-wise captions. We highlight two practical challenges: figures with an overall caption only and figures with missing captions. The output is a set of labeled panel boxes paired with corresponding panel captions (bottom).

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, achieving fine-grained understanding requires both localizing these panels and generating panel-specific captions. In practice, however, panel-level captions are frequently missing or uninformative at scale: in a PMC-scale crawl of 346,567 compound figures, 16.3% have no caption and 1.8% have captions shorter than ten words. These figures are simply discarded by existing pipelines. Our goal is therefore to produce panel-aligned visual-text pairs from compound figures even in the absence of reliable 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 detailed, panel-resolved caption to predicted visual boxes. However, this approach relies on a high-fidelity text-to-image mapping that is not always available in real-world extraction pipelines (Clark and Divvala, 2016). In practice, captions are often miss-

ing or limited to high-level figure summaries (Li et al., 2019) that fail to describe panels individually. In these text-sparse scenarios, text-conditioned decomposition is unreliable because the available metadata provides no guidance; a model must be capable of autonomously generating panel-level captions to restore the semantic details lost at the figure level.

To resolve this, we propose a visual-conditioned formulation that does not assume caption availability. The model takes only the compound figure as input, detects labeled panels, and generates a caption for each panel based on visual content, as illustrated in Figure 1. This converts otherwise unusable figures into aligned panel-text pairs for pretraining and retrieval.

Based on this insight, we introduce FigEx2, a unified framework for visually grounded panel captioning and detection. 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. However, transitioning from simple caption separation to open-ended visual captioning introduces significant linguistic variance, which can destabilize the language-conditioning signal for the detector. To mitigate this, FigEx2 introduces a noise-aware gated fusion module to transfer token-level caption features into the detector's query space, adaptively suppressing noisy channels to ensure robust localization.

Furthermore, since purely supervised training is often insufficient to ensure semantic alignment between generated captions and localized panels under noisy supervision, FigEx2 further incorporates reinforcement learning with multimodal rewards (Rennie et al., 2017; Zhang et al., 2019; Hessel et al., 2021) that directly encourage vision-text consistency in the learned captions.

Finally, we address the lack of high-quality supervision by introducing BioSci-Fig-Cap, a carefully curated panel-caption benchmark derived from BioSci-Fig. We refine existing benchmarks (Song et al., 2025) by filtering misaligned labels and enforcing scientific terminology consistency. To test true generalization, we contribute two out-of-domain test suites, PhysSci-Fig-Cap-Test in physics and ChemSci-Fig-Cap-Test in chemistry, to benchmark the model's ability to transfer its grounding capabilities across scientific disciplines.

Our contributions are summarized as follows:

- We reformulate panel extraction from caption decomposition to visual-conditioned panel captioning, where FigEx2 takes only a compound figure and produces labeled panel captions linked with detection via a `[DET]` trigger.
- 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.
- We develop a gated fusion module that stabilizes how caption signals conditionally facilitate detection under variable generation.
- We optimize joint captioning and detection with a staged SFT+RL recipe, using multimodal rewards to score panel-caption alignment with CLIP and semantic faithfulness using BERTScore.

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 on complex datasets like MedICaT.

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., 2025). 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., 2024)) 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), VisionLLM (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, 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 Reinforcement Learning for Captioning

Reinforcement learning (RL) optimizes non-differentiable sequence-level objectives, building on REINFORCE (Williams, 1992) and early policy-update recipes (Ranzato et al., 2015). Self-critical sequence training (SCST) (Rennie et al., 2017) stabilized this paradigm by using greedy baselines to optimize lexical metrics like METEOR (Banerjee and Lavie, 2005). To capture deeper semantics, rewards evolved from BERTScore (Zhang et al., 2019) and visual-semantic embeddings (Ren et al., 2017) to CLIP-style similarity (Radford et al., 2021; Hessel et al., 2021), which facilitates more distinctive descriptions (Cho et al., 2022). Recently, multi-objective formulations such as MOCHa (Ben-Kish et al., 2023) leverage RL to balance adequacy and faithfulness while mitigating hallucinations.

3 Method

3.1 Overview

As shown in Figure 2, FigEx2 is a unified framework for panel captioning and panel detection in scientific compound figures. Given only a compound figure x, a vision-language captioning branch generates panel-wise captions in a structured, label-ordered format and terminates with a special trigger token `[DET]`. The hidden state at `[DET]` provides a direct interface to the detector, which predicts the corresponding set of panel bounding boxes and labels in the same forward pass. To better link caption generation with box prediction, FigEx2 adds a cross-branch gated fusion module that feeds caption-token features into the detector queries, so the detector can use caption cues while keeping stable query decoding.

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 $\mathbf{h}_{det} \in \mathbb{R}^d$, which serves as the interface feature that triggers the detection branch. Rather than compressing language guidance into a single pooled vector, we expose the detector to the full set of caption-token hidden states. Let $\mathbf{H}_{cap} \in \mathbb{R}^{N_t\times d}$ denote the hidden states of the generated panel-caption tokens (excluding `[DET]`), where N_t is the number of caption tokens. We use these token features as text-side conditioning tokens:

$$\mathbf{F}_{txt} = \mathbf{H}_{cap}\mathbf{W}_{txt} \in \mathbb{R}^{N_t\times d}, \tag{1}$$

where $\mathbf{W}_{txt} \in \mathbb{R}^{d\times d}$ is a learned linear projection.

3.3 Gated Fusion Module

As shown in Figure 3, the gated fusion module updates the detector queries with two cross-attention steps and then applies a learnable gate to control how strongly caption features influence detection.

Stage 1-1/1-2: Cross-attention query updates. Let $\mathbf{Q} \in \mathbb{R}^{N_d\times d}$ be the N_d object queries of the detection decoder, where d is the hidden size. We condition these queries on (i) a detection-side interface token $\mathbf{F}_{det} \in \mathbb{R}^{1\times d}$ and (ii) caption-token features $\mathbf{F}_{txt} \in \mathbb{R}^{N_t\times d}$. We set $\mathbf{F}_{det}$ to the `[DET]` hidden state, i.e., $\mathbf{F}_{det} = \mathbf{h}_{det}^{\top}$. We apply two lightweight cross-attention updates:

$$\mathbf{Q}' = \textsc{Attn}(\mathbf{Q}, \mathbf{F}_{det}), \tag{2}$$

$$\mathbf{Q}'' = \textsc{Attn}(\mathbf{Q}', \mathbf{F}_{txt}), \tag{3}$$

where $\textsc{Attn}(\mathbf{Q}, \mathbf{F})$ is a standard multi-head attention block (with residual connection and layer normalization) that uses $\mathbf{Q}$ as queries and $\mathbf{F}$ as keys/values.

Stage 2: Gated query modulation. We then modulate the updated queries to control the conditioning strength:

$$[\mathbf{b},\, \mathbf{g}_{raw}] = \mathbf{Q}''\mathbf{W}_{bg} + \mathbf{c}_{bg}, \quad \mathbf{g} = \tanh(\mathbf{g}_{raw}), \tag{4}$$

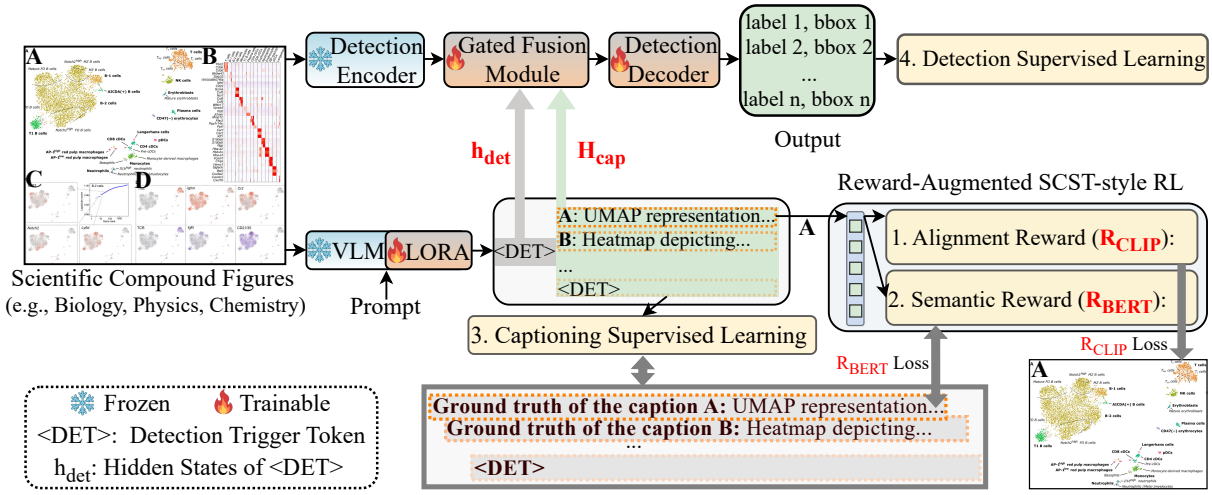


Figure 2: Overall framework of FigEx2, containing two branches: Detection (top) and Captioning (bottom). Given an input compound figure, the captioning branch first generates structured panel-wise captions and outputs a [DET] token; the detector is then conditioned on the final-layer hidden state h_{det} to predict panel boxes and labels. FigEx2 is trained with supervised caption/detection objectives and an SCST-style reward-augmented stage using CLIP-based panel-caption alignment computed from ground-truth crops and BERTScore-based semantic rewards.

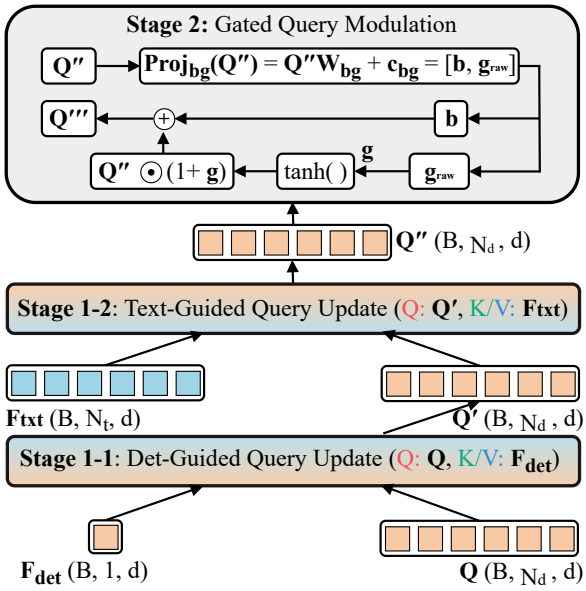


Figure 3: Gated fusion module for conditioning detector queries on caption features with cross-attention and gated modulation. B denotes batch size.

where $\mathbf{W}_{bg} \in \mathbb{R}^{d\times 2d}$ and $\mathbf{c}_{bg} \in \mathbb{R}^{2d}$ are learnable parameters, and $\mathbf{b}, \mathbf{g} \in \mathbb{R}^{N_d\times d}$ are the query-wise bias and gate. The final modulated queries are

$$\mathbf{Q}''' = \mathbf{Q}'' \odot (1 + \mathbf{g}) + \mathbf{b}, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Finally, the detection decoder uses $\mathbf{Q}'''$ to predict panel boxes and labels $\hat{D} = \{(\hat{\mathbf{b}}_i, \hat{\ell}_i)\}_{i=1}^{\hat{N}}$ with its standard query-based decoding, where $\hat{N}$ is the number of predicted panels.

3.4 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 $y_{1:T}$ and ends with a trigger token [DET]; conditioned on the corresponding hidden state, 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 (LoRA parameters) with token-level cross entropy:

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

where $y_{1:T}$ is the structured caption sequence and the input includes a fixed prompt.

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

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

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 λ_{cls}, λ_{bbox}, and λ_{giou} are scalar weights.

Stage 3: Joint supervised training (SFT). Starting from the Stage 1 and Stage 2 checkpoints, we jointly optimize captioning and detection:

$$\mathcal{L}_{sft} = \lambda_{cap}\mathcal{L}_{cap} + \lambda_{det}\mathcal{L}_{det}. \tag{8}$$

This stage enables the caption-to-detection interface and trains the gated fusion module, where λ_{cap} and λ_{det} are scalar weights.

Stage 4: Reward-augmented joint training. Starting from the Stage 3 checkpoint, we apply self-critical sequence training (SCST) to improve panel-wise caption quality and panel-caption alignment. The SCST loss is:

$$\mathcal{L}_{rl} = -\big(R(\hat{Y}^s) - R(\hat{Y}^g)\big) \sum_t \log p\big(\hat{y}^s_t \mid \hat{y}^s_{<t}, x\big), \tag{9}$$

where $\hat{Y}^s = (\hat{y}^s_{1:T})$ is a sampled structured output and let $\hat{Y}^g$ be the greedy baseline, where $\hat{y}^s_t$ denotes the t-th sampled token and $\hat{y}^s_{<t}$ is its prefix, and $R(\cdot)$ is a sequence-level reward computed from the parsed structured output.

We parse the structured output into label-keyed captions $\{\hat{c}_\ell\}$ and compute rewards over the set of ground-truth labels $\mathcal{L}$ present in the image. For each $\ell \in \mathcal{L}$, let c_ℓ and b_ℓ be the ground-truth sub-caption and box, respectively. We use a semantic reward based on BERTScore:

$$R_{\text{BERT}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{BERTSCORE}(\hat{c}_\ell, c_\ell), \tag{10}$$

and an alignment reward based on CLIP-style cosine similarity between the ground-truth crop and the generated caption:

$$R_{\text{CLIP}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \cos\big(\phi_{img}(\text{CROP}(x, b_\ell)),\ \phi_{txt}(\hat{c}_\ell)\big), \tag{11}$$

where $\phi_{img}(\cdot)$ and $\phi_{txt}(\cdot)$ are the CLIP image and text encoders, and using ground-truth crops avoids coupling the reward to localization errors early in training. We combine rewards as

$$R = \alpha R_{\text{BERT}} + \beta R_{\text{CLIP}}, \tag{12}$$

where α and β are scalar weights.

During Stage 4, we continue optimizing detection with $\mathcal{L}_{det}$ and update captioning with a hybrid supervised + RL objective:

$$\mathcal{L} = \lambda_{cap}\mathcal{L}_{cap} + \lambda_{det}\mathcal{L}_{det} + \lambda_{rl}\mathcal{L}_{rl}, \tag{13}$$

where λ_{rl} is a scalar weight.

4 Benchmarks and Curation

We evaluate visual-conditioned panel captioning on curated benchmarks spanning in-domain training and cross-domain generalization. Starting from BioSci-Fig, we construct **BioSci-Fig-Cap** with cleaner panel-level supervision for the target task. To assess transfer, we further curate two *test-only* out-of-domain benchmarks, **PhysSci-Fig-Cap-Test** and **ChemSci-Fig-Cap-Test**. In addition to these task-specific benchmarks, we also report results on MedICaT for comparison with prior compound-figure protocols.

Dataset	Split	Figures	Pairs
MedICaT	Train	1671	6359
	Val	209	776
	Test	209	840
BioSci-Fig-Cap	Train	5290	30 865
	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.

BioSci-Fig-Cap. BioSci-Fig-Cap is a curated panel-caption benchmark derived from BioSci-Fig, designed to provide higher-quality labeled panel supervision for learning panel-specific captions.

PhysSci-Fig-Cap-Test & ChemSci-Fig-Cap-Test. To evaluate cross-disciplinary generalization, we curate two out-of-domain test sets in physics and chemistry, representing scientific fields with distinct visual patterns and terminology.

Splits and statistics. MedICaT follows an 8:1:1 train/validation/test split. BioSci-Fig-Cap is curated within each original BioSci-Fig subset, and its test set is sourced only from the original BioSci-Fig test subset. The cross-domain benchmarks are used for evaluation only. Table 1 reports dataset sizes. Curation details are provided in Appendix A.

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, and BERTScore (reported as 100×F1). FigEx2 outputs labeled lines; we align by ground-truth label with occurrence-level matching in reading order, score missing or extra pairs as 0, and average within each figure and then over figures. Details are in Appendix B.

Model	MedICaT				BioSci-Fig-Cap			
	BLEU4	ROUGE-L	METEOR	BERTScore	BLEU4	ROUGE-L	METEOR	BERTScore
Pre-trained								
Qwen2.5-VL-72B	2.08	15.91	9.95	83.67	1.71	23.62	14.74	82.23
Llama3.2-90B	3.00	11.73	7.97	85.68	2.94	18.35	11.80	85.96
gpt-5-nano	2.01	10.76	5.91	85.23	2.84	16.97	9.76	85.79
gpt-5-mini	2.92	14.44	9.38	85.44	4.44	23.03	16.20	86.67
gpt-5.2	2.43	13.25	7.44	85.54	4.07	23.85	14.63	87.12
Fine-tuned								
LLaVA-NeXT	4.06	14.27	9.28	85.14	3.36	20.20	12.74	86.22
LLaVA-Med	2.26	10.58	6.26	85.00	2.91	17.57	11.32	85.78
Llama3.2-11B	4.80	16.38	11.39	85.62	3.88	21.44	15.79	86.33
Qwen3-VL-8B	7.37	18.13	12.78	86.25	5.21	25.69	18.27	87.00
FigEx2-8B	**7.83**	**19.74**	**13.89**	**86.49**	**5.43**	**26.28**	**18.71**	**87.22**

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

5.2 Baselines

We compare against three baseline families: vision-only detectors for panel detection, vision-language captioners 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 3.4. We train on 4 NVIDIA A100 80GB GPUs and run inference on NVIDIA L40S 48GB and NVIDIA A100 40GB GPUs. All hyperparameter settings are in Appendix D.

5.4 Comparison with Existing Methods

Panel Captioning. Table 2 reports panel captioning results. Fine-tuned models consistently outperform pre-trained models in this structured panel-wise setting, indicating that in-domain supervision is important for scientific panel captioning. FigEx2 achieves the best overall performance on both datasets. On MedICaT, FigEx2 improves ROUGE-L from 18.13 to 19.74 and METEOR from 12.78 to 13.89 over Qwen3-VL-8B. It also increases BLEU4 from 7.37 to 7.83 and BERTScore from 86.25 to 86.49. On BioSci-Fig-Cap, FigEx2 improves ROUGE-L from 25.69 to 26.28 and METEOR from 18.27 to 18.71 over Qwen3-VL-8B. It further increases BLEU4 from 5.21 to 5.43 and BERTScore from 87.00 to 87.22. These advantages suggest that our FigEx2 is effective, especially our reward-augmented SCST for structured panel captioning, improving semantic faithfulness.

Model	mAP@0.5:0.95	mAP@0.5
MedICaT		
YOLOS-Ti	0.155	0.183
YOLOS-S	0.156	0.188
YOLOS-B	0.150	0.170
DAB-DETR	0.165	0.181
Qwen3VL-8B	0.230	0.309
FigEx2-8B	**0.291**	**0.314**
BioSci-Fig-Cap		
YOLOS-Ti	0.100	0.129
YOLOS-S	0.196	0.248
YOLOS-B	0.448	0.534
DAB-DETR	0.697	0.830
Qwen3VL-8B	0.439	0.578
FigEx2-8B	**0.728**	**0.871**

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

Panel Detection. Table 3 reports panel detection results under the LVIS evaluator. On MedICaT, FigEx2 achieves the highest mAP, improving mAP@0.5:0.95 from 0.230 to 0.291 over prompt-based box generation with Qwen3VL-8B, while also slightly increasing mAP@0.5 from 0.309 to 0.314. On BioSci-Fig-Cap, FigEx2 remains the top method, improving mAP@0.5:0.95 from 0.697 to 0.728 over DAB-DETR and mAP@0.5 from 0.830 to 0.871, indicating more accurate detection across benchmarks. This improvement is driven by our gated fusion module and reward-augmented SCST, which stabilize caption-conditioned detector queries and yield more reliable panel localization under variable generations.

5.5 Ablation Study

Table 4 summarizes our ablations over key components of FigEx2. In Stage 4 under SCST, **Sup.**

Variant	Stage 4 (SCST)			Panel Detection		Panel Captioning			
	Sup.	Bert	CLIP	mAP@0.5:0.95	mAP@0.5	BLEU4	ROUGE-L	METEOR	BERTScore
MedICaT									
FigEx2 (S3, no gate)	–	–	–	0.272	0.301	7.52	19.64	13.53	86.45
FigEx2 (S3)	–	–	–	0.289	0.311	7.72	**19.75**	13.83	86.46
FigEx2 (SCST, Bert+CLIP)	–	✓	✓	0.286	0.309	7.70	19.41	13.78	86.43
FigEx2 (SCST, sup.+CLIP)	✓	–	✓	0.288	0.311	7.72	19.50	13.71	86.42
FigEx2 (SCST, sup.+Bert)	✓	✓	–	0.289	0.310	7.74	19.67	13.87	86.48
FigEx2 (SCST, sup.+Bert+CLIP)	✓	✓	✓	**0.291**	**0.314**	**7.83**	19.74	**13.89**	**86.49**
BioSci-Fig-Cap									
FigEx2 (S3, no gate)	–	–	–	0.722	0.865	5.35	25.75	18.25	87.15
FigEx2 (S3)	–	–	–	0.726	0.867	5.45	**26.33**	18.58	87.21
FigEx2 (SCST, Bert+CLIP)	–	✓	✓	0.723	0.859	5.44	26.22	18.52	87.22
FigEx2 (SCST, sup.+CLIP)	✓	–	✓	0.724	0.860	5.40	25.78	18.30	87.15
FigEx2 (SCST, sup.+Bert)	✓	✓	–	0.726	0.865	**5.49**	26.25	18.59	87.22
FigEx2 (SCST, sup.+Bert+CLIP)	✓	✓	✓	**0.728**	**0.871**	5.43	26.28	**18.71**	**87.22**

Table 4: Ablations of FigEx2 variants and Stage 4 (SCST) components on MedICaT and BioSci-Fig-Cap. (S3) denotes the Stage 3 checkpoint before SCST. **Sup.** keeps supervised losses during SCST; **Bert** and **CLIP** add the corresponding rewards. **no gate** disables gated query modulation in the fusion module.

Model	PhysSci-Fig-Cap-Test				ChemSci-Fig-Cap-Test			
	BLEU4	ROUGE-L	METEOR	BERTScore	BLEU4	ROUGE-L	METEOR	BERTScore
Pre-trained								
Qwen2.5-VL-72B	1.48	17.28	9.97	82.11	1.38	17.68	10.37	81.64
Llama3.2-90B	2.09	12.84	8.05	84.76	2.21	14.61	8.99	84.45
gpt-5-nano	1.85	11.56	6.33	84.37	1.73	11.60	6.09	84.15
gpt-5-mini	2.81	16.11	**10.56**	84.81	2.82	16.41	10.44	84.47
gpt-5.2	2.77	16.71	9.75	85.34	2.53	17.16	9.57	85.12
Fine-tuned								
LLaVA-NeXT	1.48	11.25	6.07	84.45	1.93	14.54	8.56	84.50
LLaVA-Med	1.56	10.72	6.04	84.67	1.57	12.83	7.75	84.43
Llama3.2-11B	2.17	15.12	9.09	85.15	2.37	16.56	10.29	84.84
Qwen3-VL-8B	2.75	17.75	9.98	85.57	3.05	18.93	11.58	85.18
FigEx2-8B	**2.89**	**18.24**	10.52	**85.69**	**3.21**	**19.93**	**12.21**	**85.48**

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

keeps supervised losses, while **Bert** and **CLIP** add a BERTScore-based semantic reward and a CLIP-similarity alignment reward, respectively. Comparing the Stage 3 jointly supervised checkpoint to its no gate counterpart shows the importance of gated query modulation: it improves detection by a clear margin, increasing mAP@0.5:0.95 from 0.272 to 0.289 on MedICaT and from 0.722 to 0.726 on BioSci-Fig-Cap, with consistent captioning gains as well.

In Stage 4, the full configuration that retains supervised losses and combines both rewards achieves the strongest overall results, reaching 0.291 mAP@0.5:0.95 on MedICaT and 0.728 on BioSci-Fig-Cap. The combined rewards yield METEOR gains of +0.06 on MedICaT and +0.13 on BioSci-Fig-Cap over the Stage 3 checkpoint, confirming that reward-augmented training strengthens semantic faithfulness. BLEU4 and ROUGE-L remain stable across SCST variants, as our rewards target semantic and alignment objectives rather than lexical overlap. Their combination under supervised SCST delivers the most consistent gains across detection and captioning.

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 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. Pre-

Few-shot	PhysSci-Fig-Cap-Test				ChemSci-Fig-Cap-Test			
	BLEU4	ROUGE-L	METEOR	BERTScore	BLEU4	ROUGE-L	METEOR	BERTScore
Qwen3-VL-8B								
1-shot	19.83	33.04	28.22	89.50	32.63	46.17	42.46	91.20
2-shot	39.41	52.62	49.37	91.99	42.56	55.57	53.05	92.77
FigEx2-8B								
1-shot	21.07	34.28	29.64	89.84	31.02	44.09	41.14	91.31
2-shot	**41.05**	**53.99**	**50.92**	**92.48**	**48.00**	**60.86**	**57.96**	**93.42**

Table 6: Few-shot prompting for cross-domain panel captioning on PhysSci-Fig-Cap-Test & ChemSci-Fig-Cap-Test.

Model	mAP@0.5:0.95	mAP@0.5
PhysSci-Fig-Cap-Test		
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	**0.423**	**0.482**
ChemSci-Fig-Cap-Test		
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	**0.394**	**0.465**

Table 7: Cross-domain panel detection on PhysSci-Fig-Cap-Test and ChemSci-Fig-Cap-Test.

trained models vary in performance across domains, while fine-tuned models are more reliable in this structured setting. FigEx2 achieves the strongest overall captioning quality across both domains. Compared with Qwen3-VL-8B, FigEx2 improves PhysSci-Fig-Cap-Test ROUGE-L from 17.75 to 18.24 and METEOR from 9.98 to 10.52, and improves ChemSci-Fig-Cap-Test ROUGE-L from 18.93 to 19.93 and METEOR from 11.58 to 12.21. These gains suggest that jointly modeling detection and captioning helps maintain generation quality under distribution shift.

Cross-domain panel detection. Table 7 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 0.423 over DAB-DETR and from 0.374 to 0.423 over prompt-based box generation with Qwen3-VL-8B. On ChemSci-Fig-Cap-Test, FigEx2 improves mAP@0.5:0.95 from 0.357 to 0.394 over DAB-DETR and achieves the best overall results on both mAP@0.5:0.95 and mAP@0.5, indicating that our joint interface remains effective under domain shift.

Few-shot prompting. We evaluate few-shot prompting by providing ground-truth subcaptions for early panels as exemplars at inference time. We consider 1-shot using the caption of panel A and 2-shot using the captions of panels A and B, and evaluate only panels C and later. Table 6 shows that increasing the number of exemplars consistently improves captioning quality in both domains, and FigEx2 benefits more from additional exemplars than the baseline. For example, on PhysSci-Fig-Cap-Test, FigEx2-8B BLEU4 increases from 21.07 to 41.05 when moving from 1-shot to 2-shot, while Qwen3-VL-8B increases from 19.83 to 39.41. On ChemSci-Fig-Cap-Test, FigEx2-8B increases from 31.02 to 48.00, compared with 32.63 to 42.56 for Qwen3-VL-8B, indicating stronger few-shot adaptation of FigEx2 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 a gated fusion module that transfers token-level caption signals into detector queries for robust localization. We curate BioSci-Fig-Cap and establish test-only benchmarks in physics and chemistry to evaluate cross-domain transfer. FigEx2 improves both detection and captioning, and reward-augmented training further strengthens panel-level faithfulness and alignment. Few-shot prompting with in-context panel captions provides consistent additional gains across disciplines, showing that FigEx2 can adapt to new scientific conventions with minimal supervision.

Ethics Statement

BioSci-Fig-Cap is derived from open-access scientific articles and does not contain private or sensitive personal data. Dataset curation and caption refinement were performed by trained annotators with domain expertise. While pretrained models may have been exposed to parts of the scientific literature during pretraining, 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, training is drawn primarily from biomedical literature, and the model may rely on domain-specific visual priors that do not transfer well to abstract diagram families such as node-link graphs; broadening coverage is mainly limited by annotation cost. Additionally, visual-conditioned captioning without access to figure captions or surrounding text can miss experimental conditions stated only in prose. Our rewards also rely on automatic metrics that may not fully capture scientific correctness, and human evaluation of panel-level captions across diverse domains remains prohibitively expensive at scale.

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, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2023. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2.

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*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

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

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*.

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.

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.

Po-shen Lee, Jevin D West, and Bill Howe. 2017. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 4(1):117–129.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan 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.

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. 2024. 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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298.

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.

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. Medicat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

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, Siwei Song, Hui Ji, Siqi Cao, Heng Yu, Zhijian Liu, Huanrui Yang, Yingyan (Celine) Lin, Beidi Chen, Mohit Bansal, Xiaoming Liu, Pengfei Zhou, Ming-Hsuan Yang, Tianlong Chen, and Jingtong Hu. 2025. From models to systems: A comprehensive survey of efficient multimodal learning. *Authorea Preprints*.

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.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

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*.

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.

	A	B	C
A	/	0.9164	0.9113
B	0.9164	/	0.9043
C	0.9113	0.9043	/

Table 8: Inter-annotator agreement for panel bounding boxes, measured by pairwise Cohen's kappa.

	A	B	C
A	/	0.8892	0.8916
B	0.8892	/	0.9325
C	0.8916	0.9325	/

Table 9: Inter-annotator agreement for panel captioning, measured by pairwise Cohen's kappa.

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).

A.2.1 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.

A.2.2 Inter-Annotator Agreement

We assess annotation consistency with pairwise Cohen's kappa (Cohen, 1960) on two subtasks: panel bounding box annotation and panel caption quality verification. Annotation and verification were conducted by a three-person team consisting of two PhD-level annotators and one PhD student, with complementary backgrounds spanning AI/machine learning, computational biology/bioinformatics, and electrical engineering. For bounding boxes, annotators verify, correct, and where necessary draw new panel boxes across all benchmarks. For caption verification, annotators judge whether each panel caption accurately describes the corresponding panel content. Table 8 and Table 9 report the pairwise kappa scores for each subtask.

A.3 Top Keywords by Dataset

Keyword	Count
expression	10112
genes	7686
spatial	6131
gene	5931
tissue	2849
clusters	2841
tumor	2687
umap	2650
cluster	2498
spots	2382
mice	2288
marker	2033
heatmap	1944
correlation	1817
enrichment	1501

Table 10: Top-15 keywords in BioSci-Fig-Cap.

Keyword	Count
optical	108
pqds	94
sensor	92
bilayer	88
density	86
phase	86
sers	76
electron	76
energy	72
piezoelectric	66
characterization	62
fabrication	60
fiber-based	58
fiber	58
perovskite	54

Table 11: Top-15 keywords in PhysSci-Fig-Cap-Test.

Keyword	Count
chemical	70
synthesis	57
binding	50
structures	50
structure	45
surface	43
mechanism	41
catalyst	36
residues	36
derivatives	33
chemistry	31
particles	29
catalytic	26
ions	24
spectra	23

Table 12: Top-15 keywords in ChemSci-Fig-Cap-Test.

B Evaluation Protocol Details

Occurrence-level label alignment. Ground-truth panels are ordered by the ground-truth reading order. For each label ℓ, let $\mathcal{R}_\ell = [r_{\ell,1}, \ldots, r_{\ell,n_\ell}]$ be reference captions for all occurrences of ℓ, and let $\mathcal{P}_\ell = [p_{\ell,1}, \ldots, p_{\ell,m_\ell}]$ be predicted captions with label ℓ. 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.

Aggregation. For each figure, we compute caption metrics on all evaluation pairs and average within the figure as $\mathrm{Score}(\text{figure}) = \frac{1}{N}\sum_{i=1}^{N} s(p_i, r_i)$, 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. Stage 3 uses joint supervised training with loss weights `lambda-cap` 0.2 and `lambda-det` 1.0. Stage 4 performs reward-augmented training with `lambda-rl` 1.0 and reward weights `rl-w-bert` 1.0 and `rl-w-clip` 0.5, with seed 42. RL rollouts use nucleus sampling with top-p 0.8, temperature 0.7, and `max-new-tokens-rl` 8192. We compute the CLIP-based alignment reward using BiomedCLIP and report BERTScore as $100\times$F1.

E Examples

Figures 4–6 show qualitative examples of FigEx2 predictions. Each figure contains the input compound image with the panel boxes and labels, along with the corresponding panel captions displayed on the right.

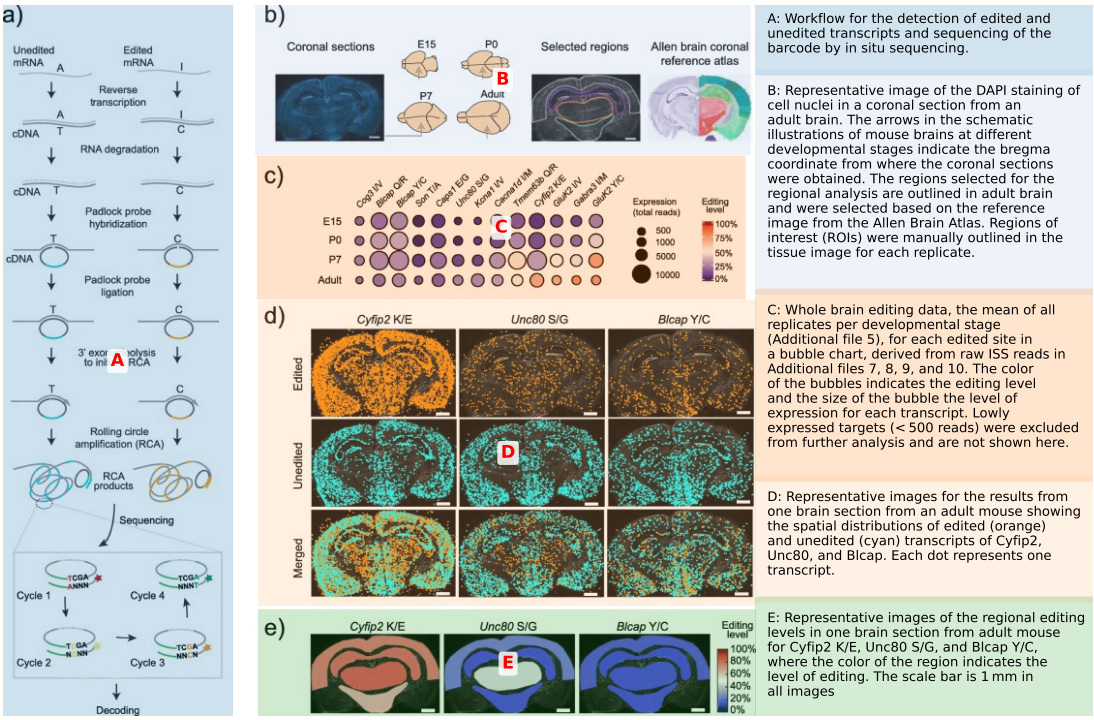


Figure 4: Compound figure with panel boxes with aligned subcaptions.

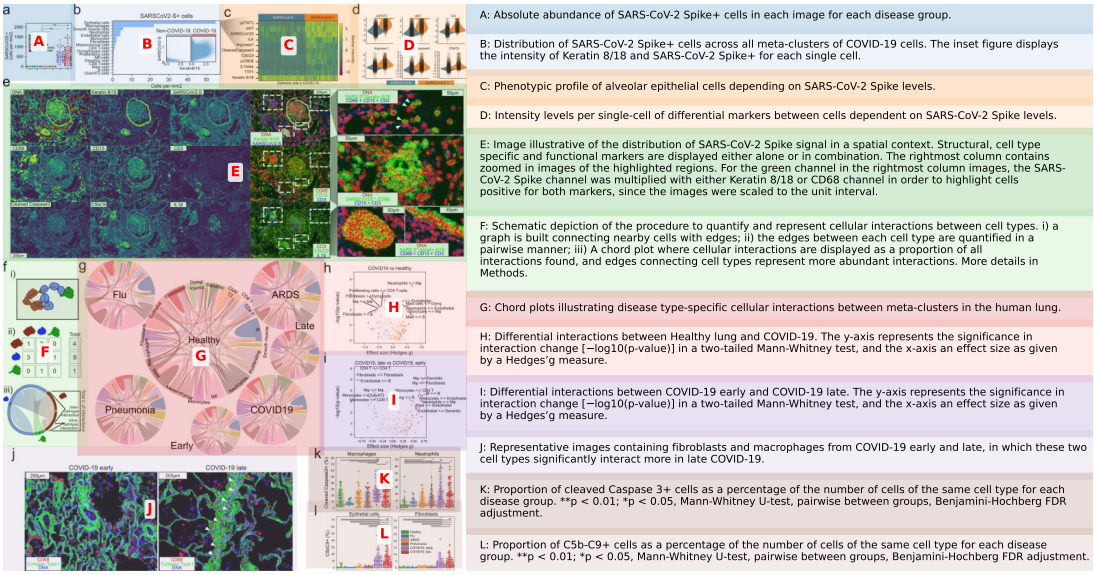
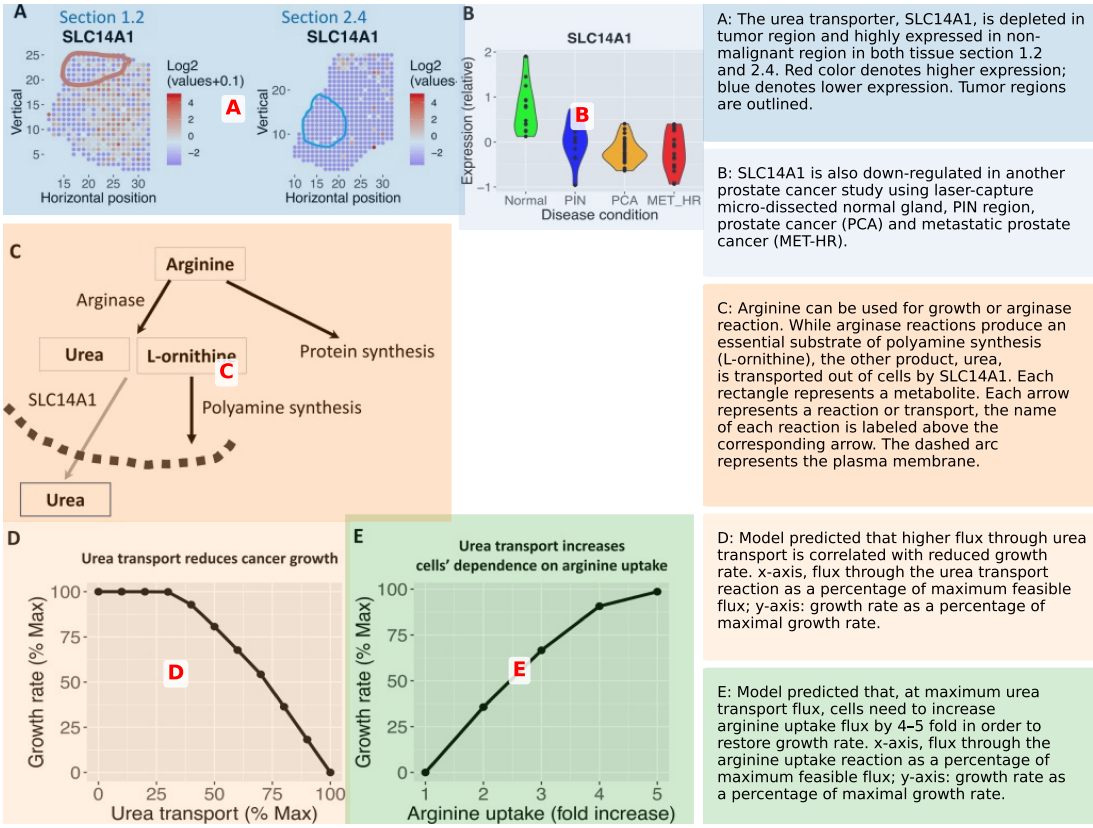


Figure 5: Compound figure with panel boxes with aligned subcaptions.



A: The urea transporter, SLC14A1, is depleted in tumor region and highly expressed in non-malignant region in both tissue section 1.2 and 2.4. Red color denotes higher expression; blue denotes lower expression. Tumor regions are outlined.

B: SLC14A1 is also down-regulated in another prostate cancer study using laser-capture micro-dissected normal gland, PIN region, prostate cancer (PCA) and metastatic prostate cancer (MET-HR).

C: Arginine can be used for growth or arginase reaction. While arginase reactions produce an essential substrate of polyamine synthesis (L-ornithine), the other product, urea, is transported out of cells by SLC14A1. Each rectangle represents a metabolite. Each arrow represents a reaction or transport, the name of each reaction is labeled above the corresponding arrow. The dashed arc represents the plasma membrane.

D: Model predicted that higher flux through urea transport is correlated with reduced growth rate. x-axis, flux through the urea transport reaction as a percentage of maximum feasible flux; y-axis: growth rate as a percentage of maximal growth rate.

E: Model predicted that, at maximum urea transport flux, cells need to increase arginine uptake flux by 4–5 fold in order to restore growth rate. x-axis, flux through the arginine uptake reaction as a percentage of maximum feasible flux; y-axis: growth rate as a percentage of maximal growth rate.

Figure 6: Compound figure with panel boxes with aligned subcaptions.