FigEx: Aligned Extraction of Scientific Figures and Captions

Jifeng Song[1, 2] , Arun Das[2, 3] , Ge Cui[1], Yufei Huang[1, 2, 3*]
[1] Department of Electrical and Computer Engineering, University of Pittsburgh, USA
[2] Cancer Virology Program, UPMC Hillman Cancer Center, USA
[3] Department of Medicine, University of Pittsburgh, USA
{jis219, ard212, gec95, yuh119}@pitt.edu

Abstract

Automatic understanding of figures in scientific papers is challenging since they often contain subfigures and subcaptions in complex layouts. In this paper, we propose FigEx, a vision-language model to extract aligned pairs of subfigures and subcaptions from scientific papers. We also release BioSci-Fig, a curated dataset of 7,174 compound figures with annotated subfigure bounding boxes and aligned subcaptions. On BioSci-Fig, FigEx improves subfigure detection AP^b over Grounding DINO by 0.023 and boosts caption separation BLEU over Llama-2-13B by 0.465. The source code is available at https://github.com/Huang-AI4Medicine-Lab/FigEx.

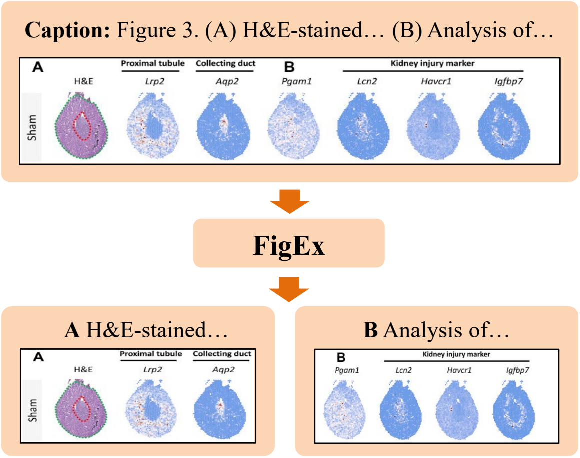


Figure 1: FigEx separates a compound figure and caption into aligned subfigures and subcaptions.

1 Introduction

Compound figure separation enables semantically consistent image-caption pairs for tasks like visual question answering and image captioning, yet vision-only pipelines still struggle with ambiguous panel borders and ignore caption text. Each year, millions of scientific publications are published that include figures with captions that summarize protocols, quantitative results, and key findings. In the biomedical domain, over 60% of figures contain microscopy images, charts, heatmaps and annotations (Meng et al., 2024). Consequently, treating a composite as a single image forces all subfigures to share one caption, violating the one-image-one-caption assumption and degrading downstream performance.

Captions for compound figures include both overall explanations and notes for each subfigure. To learn from these figures, we must separate each subfigure and align each panel with its corresponding caption segment. Simple rules from the ImageCLEF challenges used whitespace and grid layouts (De Herrera et al., 2016). CNN models improved robustness (Tsutsui and Crandall, 2017; Zou et al., 2020). Multi-step methods combined label detection with layout analysis (Jiang et al., 2021). Recent YOLO-OCR systems separate subfigures and read labels, and weakly supervised methods cut labeling effort (Yao et al., 2021). Vision-only methods still fail when separators are irregular or panels overlap, and without leveraging caption text, they miss panel-level semantics (Sun et al., 2024).

Vision-language methods offer a clear solution. Fusion models like Grounding DINO (Liu et al., 2024) integrate text features via a cross-modality decoder with language-guided queries rather than simply inserting tokens into the backbone. Region-matching models such as GLIP (Li et al., 2022b) unify object detection and phrase grounding through grounded language-image pre-training. Sequence models like Pix2Seq (Chen et al., 2021) cast object detection itself as language modeling. LLM-based systems including VisionLLM (Wang et al., 2023), LISA (Lai et al., 2024) and Lenna (Wei et al., 2025) combine detection and reasoning. However, these methods still require ex-

[*] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16558–16571
November 4-9, 2025 ©2025 Association for Computational Linguistics

plicit text queries or separate vision and language streams and fail to capture hidden links between subfigures and captions in compound figures.

Motivated by these insights, we introduce FigEx, a model guided by vision-language models (VLM) that separates compound figures and aligns each panel with its caption. This design first sends the image and full caption into the VLM to produce VLM-guided features and extract subcaptions. Next, it feeds the image and those features into a DeiT (Fang et al., 2021) to predict bounding boxes for each subfigure as in Figure 1. To connect these stages, FigEx injects a special <DET> token (Wei et al., 2025; Lai et al., 2024; Wang and Ke, 2024; Yan et al., 2024) as a supervised VLM output token. The hidden state of the <DET> token is used as a bridge to the detection module, which outputs boxes. FigEx also uses an attention mechanism to highlight the correct caption snippet for each panel.

Existing benchmarks to evaluate these methods, including benchmarks such as ImageCLEF biomedical (De Herrera et al., 2016), the MedICaT subfigure separation subset (Subramanian et al., 2020) and recent YOLO-OCR datasets (Meng et al., 2024), are restricted by either licenses or are too small to fully leverage large VLMs. They are also limited to specific scientific fields such as medicine, where the majority of the subfigures could be fixed shapes with unique colored backgrounds (rectangle and black, for example, in X-rays, CT scans, microscopy images, etc.) To address these gaps, we created BioSci-Fig, a scientific figure separation dataset, containing 7,174 compound figures with 43,183 manually annotated subfigure bounding boxes. Captions were split into fine-grained subcaptions using GPT-4.1 and then manually checked to ensure correct alignment, yielding an increase in publicly available compound figure separation data.

Our FigEx-7B model achieves strong subfigure detection performance compared to YOLO11 (Jocher and Qiu, 2024), YOLOS (Fang et al., 2021) and Grounding DINO (Liu et al., 2024) baselines, and delivers superior caption separation quality relative to Llama models (Touvron et al., 2023; Grattafiori et al., 2024).

Our contributions are summarized as follows:

- We propose **FigEx-7B**, a compact vision-language model that embeds a special <DET> token into a ViT backbone and employs hierarchical cross-attention for aligned extraction of scientific figures and captions.
- We curate **BioSci-Fig**, a dataset of 7,174 compound figures with meticulously annotated bounding boxes and aligned subcaptions, providing a benchmark setting for compound figure separation in scientific documents.
- We evaluate FigEx-7B on both MedICaT and BioSci-Fig, showing that it consistently outperforms vision-only and language-only baselines in subfigure detection and caption separation, and remains robust on long-tailed splits with a compact 7B model.

2 Related Work

2.1 Scientific Compound Figure Separation

Extracting subfigures from compound figures is a key preprocessing step for building image-caption datasets and enabling hypothesis-driven discovery (Zhou et al., 2024). The task emerged from ImageCLEF challenges (De Herrera et al., 2016), where early work relied on heuristics before shifting to CNNs. Tsutsui et al. (Tsutsui and Crandall, 2017) introduced a data-driven separation method, and Zou's unified framework (Zou et al., 2020) broadened generalization across figure styles. Hierarchical pipelines followed: Jiang et al. (Jiang et al., 2021) combined label detection with layout reasoning, while a recent YOLO-OCR system jointly splits figures and recognizes labels, releasing a new ImageCLEF-derived dataset. Weakly-supervised schemes such as Yao et al. (Yao et al., 2021) cut annotation costs via side losses and augmentation, achieving state-of-the-art ImageCLEF-2016 results.

Despite progress, current methods share three limitations: (i) basic CNN backbones struggle with intricate layouts, (ii) decomposition-centric designs overlook contextual links among subfigures, and (iii) purely visual cues ignore caption guidance, causing semantic fragmentation (Sun et al., 2024). Addressing contextual dependencies, caption-driven guidance, and multimodal fusion therefore remains essential for practical compound figure separation.

2.2 Vision-Language Enhanced Object Detection and Segmentation

Advances in vision-language research have produced three converging lines of work that progressively couple linguistic reasoning with open-vocabulary detection. First, architectural hybridization augments classic detectors with language-aware fusion. Grounding DINO (Liu et al., 2024) integrates image-text cross-attention and language-guided queries to enable open-set grounding, and DQ-DETR (Liu et al., 2023b) advances cross-modal grounding with dual queries. Later variants adopt related fusion modules and generally condition detection on text to improve zero-shot grounding. Second, task reformulation casts detection as sequence generation. Pix2Seq (Chen et al., 2021) tokenizes boxes and labels into an autoregressive stream, and multitask sequence-to-sequence frameworks such as OFA (Wang et al., 2022) handle captioning and grounding, with OFA also supporting detection within one model. Separately, GLIP (Li et al., 2022b) unifies phrase grounding and detection via grounded pre-training. These models are widely evaluated on referring-expression comprehension and typically rely on explicit prompts. Third, LLM-centric systems couple reasoning with perception modules. VisionLLM (Wang et al., 2023) connects an LLM to task-specific decoders via routing tokens and super-link queries, LISA (Lai et al., 2024) aligns language reasoning with segmentation, DetGPT (Pi et al., 2023) composes a VLM with an open-vocabulary detector to follow natural-language instructions, and Lenna (Wei et al., 2025) exposes a detection interface to an MLLM. CogVLM (Wang et al., 2024) introduces trainable visual-expert modules for deep fusion, and VLRM (Dzabraev et al., 2024) treats vision-language models as reward models to optimize generation.

2.3 Multimodal Vision-Language Models

Multimodality has rapidly matured in vision-language modeling (Wang et al., 2025; Zhao et al., 2024; Li et al., 2022a). Multimodal integration VLM has progressed through several design approaches. Connector-based models such as BLIP-2 (Chen et al., 2024b) and mPLUG-OwL (Ye et al., 2024) align vision and text by bridging frozen encoders and LLMs with lightweight adapters. They excel at captioning and retrieval but can struggle with multi-step reasoning. Instruction-tuned systems including LLaVA (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023) and Shikra (Chen et al., 2023) leverage GPT-4-style synthetic instruction data or spatial coordinate prompts, aligning pretrained vision encoders and LLMs via a projection or lightweight fine-tuning. MIMIC-IT (Li et al., 2023) provides large-scale multimodal instruction data supporting such tuning. Hybrid architectures preserve specialized vision processing: VisionLLM (Wang et al., 2023) routes tasks to external detectors, while DetGPT (Pi et al., 2023) converts detection outputs into language tokens, maintaining detector accuracy at the cost of latency. Yet balancing detection precision with deep reasoning remains an open challenge (Miyai et al., 2024; Dzabraev et al., 2024). Emerging work such as GSVA (Xia et al., 2024), LLMFormer (Shi et al., 2025), TaskCLIP (Chen et al., 2024a) and VLRM (Dzabraev et al., 2024) explores interfaces between perception modules and language models and introduces alignment or reinforcement mechanisms to better couple visual recognition with language.

3 Method

3.1 Architecture of FigEx

We introduce FigEx-7B, a multimodal model that splits compound figures into subfigures and matches each with its caption. Figure 2 shows the overall design (Wen et al., 2023). FigEx-7B builds on reasoning-based detection methods (Wei et al., 2025; Lai et al., 2024; Yan et al., 2024; Dai et al., 2025). It uses LLaVA-7B (Liu et al., 2023a) for language understanding and a DeiT-S backbone from YOLOS-S for vision. To bridge these components, we add a VLM-guided cross-attention module that steers visual feature extraction with the subcaptions and improves subfigure detection.

As shown in Figure 2, FigEx-7B first feeds the compound figure image x_{img} and full caption x_{txt} into the VLM module $\mathcal{F}$. The VLM outputs the detection-token hidden state h_{det} and subcaption features f_{subcap} as

$$h_{det}, f_{subcap} = \mathcal{F}(x_{img}, x_{txt}) \tag{1}$$

A vision encoder $\mathcal{F}_{enc}$ then processes the image to produce initial image features f_{img} as

$$f_{img} = \mathcal{F}_{enc}(x_{img}) \tag{2}$$

The VLM-guided cross-attention module produces a refined image feature as

$$\hat{f}_{img} = \mathcal{F}_{vgm}(f_{img}, h_{det}, f_{subcap}) \tag{3}$$

Finally, the MLP detection head $\mathcal{F}_{mlp}$ predicts bounding boxes $\hat{D}_{pred}$ from the refined feature as

$$\hat{D}_{pred} = \mathcal{F}_{mlp}(\hat{f}_{img}) \tag{4}$$

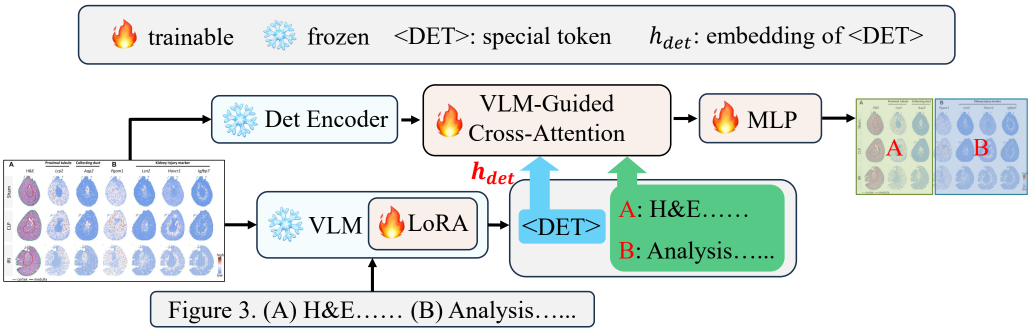


Figure 2: Overview of the FigEx architecture.

3.2 VLM-Guided Cross-Attention

As illustrated in Figure 3, we refine image features with two cross-attention modules.

First, the detection-text module $\mathcal{F}_{det\text{-}text}$ fuses the detection-token hidden states h_{det} with the sub-caption feature f_{subcap} as

$$f_{det\text{-}text} = \mathcal{F}_{det\text{-}text}(h_{det}, f_{subcap}) \tag{5}$$

Next, the cross-attention module $\mathcal{F}_{img}$ refines the encoded image feature f_{img} with $f_{det\text{-}text}$ to produce

$$\hat{f}_{img} = \mathcal{F}_{img}(f_{img}, f_{det\text{-}text}) \tag{6}$$

This step follows the standard query-key-value (Q, K, V) attention mechanism to highlight regions most relevant to subcaptions.

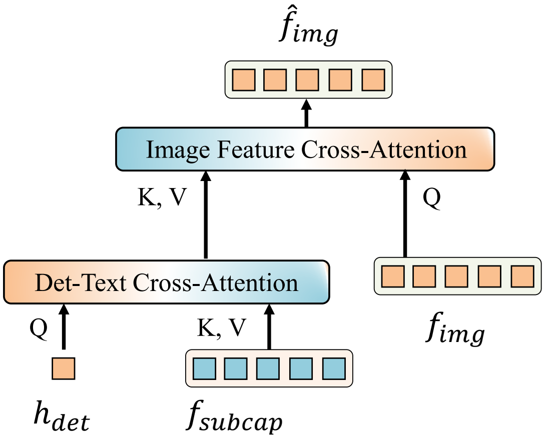


Figure 3: VLM-guided cross-attention.

3.3 Training Process

Our training pipeline, illustrated in Figure 4, consists of three sequential phases: an initial LoRA adaptation phase for tuning the language decoder, a detection head training phase for fine-tuning the vision encoder and MLP detection head, and a final joint fine-tuning phase that optimizes both token prediction and detection objectives together.

Stage 1: LoRA adaptation We freeze the vision encoder and MLP detection head and update only the LoRA layers in the VLM. The optimization target is the standard token-level cross-entropy loss as

$$\mathcal{L}_{tok} = -\frac{1}{L}\sum_{j=1}^{L} \log p(y_j \mid y_{<j}, x_{img}, x_{txt}) \tag{7}$$

where L is the length of the target subcaption, y_j is the j-th ground-truth token, $y_{<j}$ denotes all previous tokens, x_{img} denotes image features, and x_{txt} denotes the textual input embeddings.

Stage 2: Detection head training Freeze the VLM and fine-tune both the vision encoder and the MLP detection head, optimizing a loss combining classification cross-entropy, box regression and the Generalized Intersection over Union (GIoU) term as

$$\mathcal{L}_{det} = \lambda_{cls}\,\mathcal{L}_{cls} + \lambda_{bbox}\,\mathcal{L}_{bbox} + \lambda_{GIoU}\,\mathcal{L}_{GIoU} \tag{8}$$

Stage 3: Joint fine-tuning Unfreeze the LoRA adapters, the VLM-guided cross-attention module, and the MLP detection head, while the vision encoder remains frozen. Then jointly optimize the token-prediction and detection objectives.

$$\mathcal{L} = \lambda_{tok}\,\mathcal{L}_{tok} + \lambda_{det}\,\mathcal{L}_{det} \tag{9}$$

3.4 Dataset Formulation

BioSci-Fig is a large, manually curated dataset of 7,174 compound figures drawn from open-access scientific publications. A team of three annotators used Label Studio to place 43,183 precise

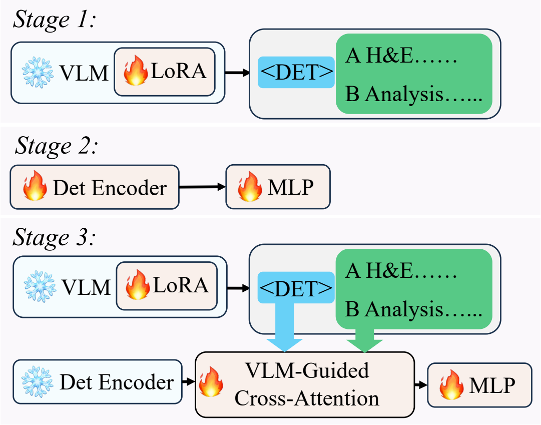


Figure 4: Three-stage training of FigEx.

bounding boxes, ensuring that each subfigure label (A–Z, a–z) falls entirely within its corresponding region. Annotation was especially challenging due to ambiguous panel borders, overlapping elements, and diverse layout styles, requiring rigorous cross-checking to maintain consistency and accuracy across all figures. This intensive effort yields high-quality ground truth for subfigure detection bounding boxes in complex scientific contexts.

In addition to BioSci-Fig, we train and evaluate our model on MedICaT, which provides 2,089 human-verified compound figures with annotated captions and bounding boxes. For BioSci-Fig, we split each joint caption into fine-grained subcaptions using GPT-4.1 with the prompt in Appendix A and manually verify every split to guarantee correct alignment. We then filter both datasets to retain only figures labeled A–Z or a–z and divide each into training, validation, and test sets in an 8:1:1 ratio as shown in Table 1.

Dataset	Train	Validation	Test
MedICaT	1,671	209	209
BioSci-Fig	5,740	717	717

Table 1: Number of compound figures per split (train, validation, test) for MedICaT and BioSci-Fig.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate subfigure detection by comparing predicted and ground-truth bounding boxes. Since the subfigure label counts in the MedICaT subset and BioSci-Fig follow a long-tailed distribution shown in Appendix C, we employ both standard Average Precision (AP) metrics and long-tail-aware metrics using the Detectron2 LVIS evaluation (Wu et al., 2019).

We report AP^b, AP^b_{50} and AP^b_{75} to evaluate overall detection quality. The metric AP^b averages precision across IoU thresholds from 0.5 to 0.95 in increments of 0.05. To assess performance under class imbalance, we use long-tailed-aware metrics AP_f, AP_c and AP_r. These metrics follow the same IoU averaging scheme as AP^b but compute precision separately for frequent labels, common labels and rare labels as described by Dong et al and Zhang et al (Dong et al., 2023; Zhang et al., 2023). In our setup, subfigure labels A through D are considered frequent, E through H are common and all others are rare.

To evaluate subcaption quality, we use BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) to measure lexical overlap between generated and reference text. We also include SPICE (Anderson et al., 2016) and BERTScore (Zhang et al., 2019) to assess semantic similarity. SPICE parses captions into scene graphs and compares semantic propositions, while BERTScore computes token-level cosine similarity between contextual embeddings.

4.2 Hyperparameters

We reference the hyperparameter settings from Lenna (Wei et al., 2025) and adjust them on our validation set.

In stage 1, we train the LoRA adapters for 10 epochs with a learning rate of 1×10^{-5} and batch size 1. In stage 2, we fine-tune the vision encoder and MLP for 50 epochs with a learning rate of 5×10^{-5}, batch size 1 and gradient accumulation of 8 steps. In stage 3, we jointly fine-tune the model for 20 epochs using a learning rate of 1×10^{-5} and batch size 1. We use AdamW as the optimizer and run all experiments on two A100 80 GB GPUs. The λ_{cls}, λ_{bbox}, λ_{GIoU}, λ_{tok} and λ_{det} are set to 1, 5, 2, 0.5, 0.5.

4.3 Comparison with Existing Methods

4.3.1 Evaluation of Subfigure Detection

We evaluate FigEx-7B against six established baselines shown in Table 2. YOLO11n and YOLO11l represent lightweight and large anchor-free detectors. YOLOS-Ti, YOLOS-S and YOLOS-B scale a vision transformer backbone from 6 M to 88 M parameters, providing a pure-vision transformer reference. Grounding DINO couples a Swin-T encoder with text prompts and therefore serves as a strong

Model	Vision Backbone	Dataset	AP^b	AP^b_{50}	AP^b_{75}	AP_r	AP_c	AP_f
YOLO11n	/	MedICaT	0.052	0.065	0.054	0	0	0.110
YOLO111	/		0.156	0.160	0.160	0	0	0.334
YOLOS-Ti	DeiT-Ti		0.155	0.183	0.168	0	0	0.332
YOLOS-S	DeiT-S		0.156	0.188	0.180	0	0.003	0.332
YOLOS-B	DeiT-B		0.150	0.170	0.157	0	0	0.321
Grounding DINO	Swin-T		0.165	0.169	0.167	0	0.021	0.332
FigEx-7B	DeiT-S		**0.175**	**0.203**	**0.199**	0	**0.024**	**0.351**
YOLO11n	/	BioSci-Fig	0.542	0.662	0.568	0.502	0.643	0.511
YOLO111	/		0.557	0.645	0.583	**0.519**	0.625	0.555
YOLOS-Ti	DeiT-Ti		0.370	0.486	0.416	0.176	0.417	0.663
YOLOS-S	DeiT-S		0.512	0.630	0.579	0.341	0.579	0.744
YOLOS-B	DeiT-B		0.537	0.649	0.599	0.347	0.643	0.763
Grounding DINO	Swin-T		0.572	0.601	0.600	0.437	**0.683**	0.697
FigEx-7B	DeiT-S		**0.595**	**0.753**	**0.687**	0.463	0.655	**0.767**

Table 2: Subfigure detection results on MedICaT and BioSci-Fig. We report AP^b, AP^b_{50}, AP^b_{75}, and AP_r, AP_c, AP_f (rare, common, frequent). Baselines (YOLO11, YOLOS, Grounding DINO) are compared with FigEx-7B. Bold indicates the best value in each column.

vision-language baseline. We compare FigEx-7B with baselines on the MedICaT and BioSci-Fig datasets.

On MedICaT, FigEx-7B achieves AP^b_{50} of 0.203, AP^b_{75} of 0.199 and AP^b of 0.175, surpassing Grounding DINO by 0.034 in AP^b_{50}, 0.032 in AP^b_{75} and 0.010 in AP^b. The long-tail metric AP_f for frequent subfigures rises from 0.332 to 0.351 while the AP_r remains at 0.000, reflecting the challenge of detecting rare labels in a small dataset. These results show that both FigEx-7B and Grounding DINO gain significant improvements on common subfigures and that FigEx-7B also leads on frequent subfigures. Although Grounding DINO outperforms vision-only detectors in overall AP^b, its advantage comes almost entirely from common samples and it offers no clear benefit on frequent subfigures, limiting its practical impact. In contrast, FigEx-7B outperforms all baselines consistently across both frequent and common subfigures.

On BioSci-Fig, FigEx-7B records AP^b_{50} of 0.753, AP^b_{75} of 0.687 and AP^b of 0.595, outperforming YOLO111 by 0.108 in AP^b_{50}, 0.104 in AP^b_{75} and 0.038 in AP^b. The AP_f increases from 0.555 to 0.767, while performance on rare categories remains comparable. Although YOLOS-B narrows the gap in AP_f when more training examples are available, it still falls behind on rare subfigures. This shows that text input and vision-language enhanced features give a clear advantage for detecting infrequent subfigures that are underrepresented in the data. We also find that transformer-based detectors outperform YOLO11 models on frequent subfigures while YOLO11 excels on rare subfigures, demonstrating the higher data demand of transformer architectures. In practical document analysis tasks where frequent subfigures are the most significant, FigEx-7B leads on frequent, is competitive on common, and improves rare over pure-vision transformers.

We further observe that the larger sample size in BioSci-Fig raises the baseline performance of pure-vision models, yet the gains from FigEx-7B remain robust across both dataset scales. The persistent advantage on rare subfigures underscores the value of language-guided features for edge cases and suggests that future work should explore targeted augmentation of rare classes and deeper text integration to close the remaining gaps.

4.3.2 Evaluation of Caption Separation

We compare FigEx-7B with zero-shot Llama models of various sizes including the text-only Llama-3.1-8B, the multimodal Llama-3.2-11B and the text-only Llama-2-13B on MedICaT and BioSci-Fig shown in Table 3.

On MedICaT, FigEx-7B leads across most metrics. ROUGE-1 climbs to 0.382, surpassing Llama-2-13B by 0.035 and Llama-3.1-8B by 0.060. ROUGE-2 improves from 0.285 to 0.340 and ROUGE-L from 0.324 to 0.366. BLEU reaches 0.157, closely matching the 0.160 of Llama-2-13B and clearly outpacing Llama-3.2-11B at 0.133. SPICE rises from 0.277 to 0.332, and BERTScore increases from 0.858 to 0.878. These gains illustrate that incorporating visual context through mul-

Model	Dataset	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	SPICE	BERTScore
Llama-3.1-8B	MedICaT	0.156	0.322	0.271	0.307	0.182	0.238	0.764
Llama-3.2-11B		0.133	0.311	0.260	0.298	0.156	0.284	0.857
Llama-2-13B		**0.160**	0.347	0.285	0.324	**0.274**	0.277	0.858
FigEx-7B		0.157	**0.382**	**0.340**	**0.366**	0.204	**0.332**	**0.878**
Llama-3.1-8B	BioSci-Fig	0.237	0.437	0.353	0.402	0.239	0.379	0.871
Llama-3.2-11B		0.183	0.351	0.274	0.319	0.156	0.267	0.849
Llama-2-13B		0.257	0.445	0.368	0.415	0.243	0.385	0.859
FigEx-7B		**0.722**	**0.809**	**0.777**	**0.798**	**0.492**	**0.779**	**0.958**

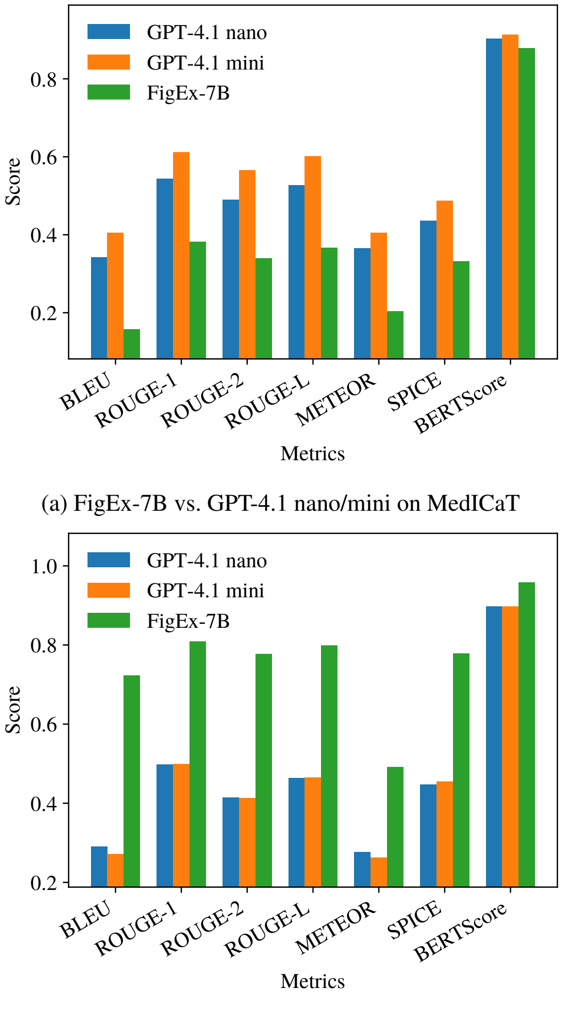
Table 3: Caption separation on MedICaT and BioSci-Fig. We report BLEU, ROUGE-1/2/L, METEOR, SPICE, and BERTScore. Bold marks the best value in each column.

timodal fine-tuning helps smaller models match or surpass larger, text-only baselines when training data are limited. The improvement in semantic metrics like SPICE and BERTScore further indicates that visual grounding contributes to more accurate and meaningful caption separation, not merely improved lexical matching.

On BioSci-Fig, the performance gap expands significantly. FigEx-7B attains a BLEU of 0.722, nearly triple the 0.257 achieved by Llama-2-13B, the best-performing baseline. ROUGE-1 reaches 0.809, ROUGE-2 climbs to 0.777, and ROUGE-L reaches 0.798, meaning each metric exceeds the baselines by a wide margin. Additionally, METEOR improves dramatically from 0.243 to 0.492, and SPICE nearly doubles, rising from 0.385 to 0.779. BERTScore, reflecting semantic precision, climbs to an impressive 0.958. The consistent lead across both lexical-overlap metrics like BLEU, ROUGE and METEOR, and semantic-focused metrics like SPICE and BERTScore underscores that FigEx-7B learns both precise textual alignment and deeper semantic understanding through visual grounding.

We further compare FigEx-7B with GPT-4.1 nano and GPT-4.1 mini on both datasets in Figure 5. On MedICaT, FigEx-7B underperforms compared to GPT-4.1 variants across all metrics, indicating limitations in its ability to match the language-generation capabilities of GPT-based models with limited training data. In particular, GPT-4.1 mini exhibits notable superiority, outperforming both GPT-4.1 nano and FigEx-7B significantly in BLEU, ROUGE, METEOR, SPICE, and BERTScore metrics.

Conversely, on the larger BioSci-Fig dataset, FigEx-7B notably surpasses both GPT-4.1 nano and GPT-4.1 mini across every evaluated metric. Specifically, FigEx-7B achieves substantial improvements in BLEU, ROUGE, METEOR, SPICE, and BERTScore, demonstrating robust performance gains in both lexical-overlap and semantics-oriented metrics. These findings suggest that mul-



(a) FigEx-7B vs. GPT-4.1 nano/mini on MedICaT

(b) FigEx-7B vs. GPT-4.1 nano/mini on BioSci-Fig

Figure 5: Caption separation on MedICaT (a) and BioSci-Fig (b). Bars compare GPT-4.1 nano, GPT-4.1 mini, and FigEx-7B over BLEU, ROUGE-1/2/L, METEOR, SPICE, and BERTScore.

Model	Dataset	AP^b	AP^b_{50}	AP^b_{75}	AP_r	AP_c	AP_f
FigEx-7B (VLM-guided)	MedICaT	0.165	0.197	0.192	0	0.013	0.341
FigEx-7B (Text)		0.166	0.198	0.189	0	0.041	0.316
FigEx-7B		**0.175**	**0.203**	**0.199**	0	**0.024**	**0.351**
FigEx-7B (VLM-guided)	BioSci-Fig	0.528	0.686	0.618	0.360	0.608	0.741
FigEx-7B (Text)		0.535	0.672	0.613	0.374	0.610	0.741
FigEx-7B		**0.595**	**0.753**	**0.687**	**0.463**	**0.655**	**0.767**

Table 4: Ablation of FigEx-7B variants on MedICaT and BioSci-Fig. We report AP^b, AP^b_{50}, AP^b_{75}, and AP_r, AP_c, AP_f (rare, common, frequent). Bold marks the best value in each column.

timodal fine-tuning enables FigEx-7B to significantly exceed the caption-separation performance of lightweight GPT-4.1 variants when training samples reach approximately 5,740 or more.

These observations highlight that under data-limited conditions, FigEx-7B's fine-tuned performance remains weaker than zero-shot GPT-based models. However, with an increased number of annotated examples, such as the 5,740 training samples available in BioSci-Fig, FigEx-7B clearly outperforms lightweight GPT-4.1 models, indicating that moderate-sized training datasets are sufficient for multimodal methods to surpass lightweight GPT-4.1 variants. Second, BioSci-Fig captions are substantially longer and richer than those in MedICaT, implying that FigEx-7B's strength may lie particularly in handling longer and more complex captions. Future studies should further investigate how caption length and complexity specifically influence the relative advantages of vision-language integration.

4.3.3 Ablation Study

Table 4 compares three variants of FigEx-7B to isolate the impact of language and vision features. The VLM-guided branch variant keeps only the VLM-guided detection token path, removing all text cross-attention while the text-only branch variant keeps only the caption cross-attention branch, discarding the VLM-guided detection token. On MedICaT, the full model achieves AP^b of 0.175 which is 0.010 higher than the VLM-guided branch variant at 0.165 and 0.009 higher than the text-only variant at 0.166. The AP_c for the full model is 0.024 which lies between the VLM-guided branch score of 0.013 and the text-only score of 0.041. These results show that combining VLM-guided features and textual cues preserves the precise localization of the complementary gains without losing the detection strength of the VLM stream.

On BioSci-Fig, the combined model records AP^b_{50} of 0.753 and AP^b_{75} of 0.687, exceeding the VLM-guided variant by 0.067 and 0.069 respectively, and AP^b of 0.595, which is 0.067 above VLM-guided and 0.060 above text-only. The AP_r rises to 0.463 compared with 0.360 for VLM-guided and 0.374 for text-only, while the AP_f climbs to 0.767. These gains reflect the complementary strengths of the two branches.

Our two-step attention first attends to associate each subcaption token with the VLM-guided spatial features, and then fuses the resulting semantic signal into the image feature map before detection. This simple fusion makes the detector focus on regions that are both visually salient and semantically meaningful, delivering steady improvements across datasets and effectively handling long-tail imbalances in scientific figure layouts.

5 Conclusion

In this paper, we propose FigEx-7B, a compact vision-language model for aligned extraction of scientific figures and captions. The architecture combines a lightweight ViT backbone with a vision-language model, showing that careful multimodal design can match or surpass much larger single-modality systems for object detection and text separation tasks. To foster progress in compound figure analysis, we release BioSci-Fig, a curated dataset of 7,174 scientific figures with precise bounding boxes and aligned subcaptions, providing a new benchmark for fine-grained document understanding. Our experiments reveal that VLM-guided spatial features and textual cues are complementary. In addition, moderate amounts of paired data enable multimodal fine-tuning to outperform larger models, and language guidance is especially valuable for rare or ambiguous panels.

Ethics Statement

BioSci-Fig figures were mined from open-access PubMed Central (PMC) articles in the biomedical literature and annotated by our team without involving any private or sensitive personal data. We note that public checkpoints may have been exposed to some evaluation samples during pre-training, which could slightly inflate baseline results.

Limitation

Our work is confined to subfigure detection and caption separation and does not assess open-ended multimodal tasks such as visual question answering or iterative subcaption refinement, and since all experiments are conducted on biomedical figures like MedICaT and BioSci-Fig, the model's ability to generalize to other domains, such as engineering diagrams, remains unverified.

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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

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.

Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Yezi Liu, Fei Wen, Alvaro Velasquez, Hugo Latapie, and Mohsen Imani. 2024a. Taskclip: Extend large vision-language model for task oriented object detection. *arXiv preprint arXiv:2403.08108*.

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

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*.

Wei Chen, Changyong Shi, Chuanxiang Ma, Wenhao Li, and Shulei Dong. 2024b. Depthblip-2: Leveraging language to guide blip-2 in understanding depth information. In *Proceedings of the Asian Conference on Computer Vision*, pages 2939–2953.

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

Siyuan Dai, Kai Ye, Guodong Liu, Haoteng Tang, and Liang Zhan. 2025. Zeus: Zero-shot llm instruction for union segmentation in multimodal medical imaging. *arXiv preprint arXiv:2504.07336*.

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*.

Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. 2023. Boosting long-tailed object detection via step-wise learning on smooth-tail data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6940–6949.

Maksim Dzabraev, Alexander Kunitsyn, and Andrei Ivaniuta. 2024. Vlrm: 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*.

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.

Glenn Jocher and Jing Qiu. 2024. Ultralytics yolo11.

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.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. 2023b. 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, 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.

Shuo Meng, Xinshuo Liang, Shuai Zhang, Leqi Lei, Hanbai Wu, Saira Iqbal, and Jinlian Hu. 2024. Yolo-ocr: End-to-end compound figure separation and label recognition of images in scientific publications. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 118–126. SIAM.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, and 1 others. 2024. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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*.

Hengcan Shi, Son Duy Dao, and Jianfei Cai. 2025. Llmformer: Large language model for open-vocabulary semantic segmentation. *International Journal of Computer Vision*, 133(2):742–759.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. 2024. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5034–5042.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Junchi Wang and Lei Ke. 2024. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774.

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.

Chen-Yueh Wen, Jui-Hu Hsiao, Yen-Dun Tony Tzeng, Renin Chang, Yi-Ling Tsang, Chen-Hsin Kuo, and Chia-Jung Li. 2023. Single-cell landscape and spatial transcriptomic analysis reveals macrophage infiltration and glycolytic metabolism in kidney renal clear cell carcinoma. *Aging (albany NY)*, 15(20):11298.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer.

Tianyuan Yao, Chang Qu, Quan Liu, Ruining Deng, Yuanhan Tian, Jiachen Xu, Aadarsh Jha, Shunxing Bao, Mengyang Zhao, Agnes B Fogo, and 1 others. 2021. Compound figure separation of biomedical images with side loss. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pages 173–183. Springer.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

Shaoyu Zhang, Chen Chen, and Silong Peng. 2023. Reconciling object-level and global-level objectives for long-tail detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18982–18992.

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. 2024. X-ray made simple: Lay radiology report generation and robust evaluation. *arXiv preprint arXiv:2406.17911*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Shijia Zhu, Naoto Kubota, Shidan Wang, Tao Wang, Guanghua Xiao, and Yujin Hoshida. 2024. Stie: Single-cell level deconvolution, convolution, and clustering in in situ capturing-based spatial transcriptomics. *Nature communications*, 15(1):7559.

Jie Zou, George Thoma, and Sameer Antani. 2020. Unified deep neural network for segmentation and labeling of multipanel biomedical figures. *Journal of the Association for Information Science and Technology*, 71(11):1327–1340.

A Prompt for Zero-Shot Caption Separation

The prompt is inspired by the PathAsst prompt (Sun et al., 2024).

> Given a caption of an image containing sub-images, please decompose the caption in accordance with each sub-image. Be sure to adhere to the following guidelines:
> 1. Preserve the original wording of the caption. Refrain from adding new information, summaries, or introductions.
> 2. Omit references to the index or number of the sub-images, such as xx), (xx), left, right, etc.
> 3. There might be a common caption shared among all sub-images; please incorporate it into each sub-image's caption.
> 4. The final output must use the following subcaption format:
> A: <original or combined caption text for sub-image A>
> B: <original or combined caption text for sub-image B>
> C: <...etc...>
> Input Caption: {caption}
> Output:

B Annotation

A postdoc and two graduate students manually annotated compound figures retrieved via the PMC API in Label Studio (Tkachenko et al., 2020-2025).

B.1 Annotation Instructions

Annotators should draw a bounding box around each subfigure and include the subfigure label.

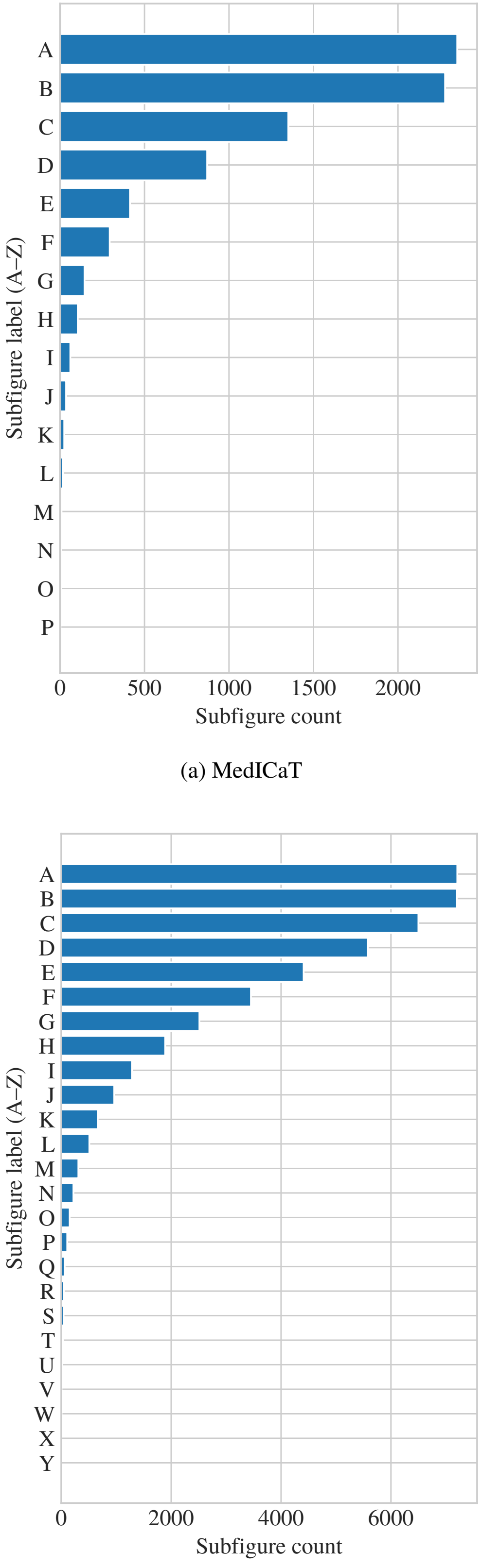
B.2 Inter-Annotator Agreement (IAA)

Three annotators labeled bounding boxes on 7,174 compound figures. Table 5 shows Cohen's kappa scores (Cohen, 1960) between annotator pairs.

	A	B	C
A	/	0.9500	0.9000
B	0.9221	/	0.8916
C	0.8911	0.8825	/

Table 5: Inter-annotator agreement (Cohen's κ).

C The Distribution of Subfigures



(a) MedICaT

(b) BioSci-Fig

Figure 6: Subfigure label distributions (A–Z) for MedICaT (a) and BioSci-Fig (b).

D Example

Caption: Single-cell level clustering in spatial transcriptomics by STIE. Cell type specific transcriptomic signature learning from 10X Visium mouse brain hippocampus FFPE (a) and 10X Visium human breast cancer FFPE (b). Spot-level clustering by K-means, SpaGCN, MUSE, subspot-level clustering by BayesSpace and single-cell-level clustering by STIE on 10X Visium FFPE mouse brain hippocampus (c), mouse brain cortex (e) and human breast cancer (g). Cell type deconvolution of spot-, subspot-, and single-cell-level clustering-derived CAGE in the mouse brain hippocampus (d), mouse brain cortex (f), and human breast cancer (h). For the mouse brain cortex, the cell types in the transcriptomic signature, which are not cortex layers and have small proportions, are not shown in the barplot. The box plot (h) represents the deconvoluted proportion of 9 cell types, where center line represents median, lower and upper hinges represent first and third quartiles, and whiskers extend from hinge to ±1.5 × IQR. The p-value was calculated based on one-sided Wilcoxon signed-rank test without adjustment for multiple comparisons. i The UMAP plot of human breast cancer scRNA-seq data from 26 primary tumors. The top panel is the original cell typing of 10,060 single cells, and the bottom panel is the subset of cells that are mapped to the six STIE clusters. Spot-level clustering by K-means (left), SpaGCN (middle), and single-cell-level clustering by STIE (right) on the simulated high-resolution spot spatial transcriptome data of the mouse brain hippocampus (j) and human breast cancer (m). Cell type deconvolution of spot- and single-cell-level clustering-derived CAGE in the mouse brain hippocampus (k) and human breast cancer (n). l, o The consistency table of single-cell clusters between the simulated high-resolution spot-based STIE clustering and the original low-resolution spot-based STIE clustering as ground truth of the mouse brain hippocampus (c) and human breast cancer (g). Source data are provided as a Source Data file. (Zhu et al., 2024)

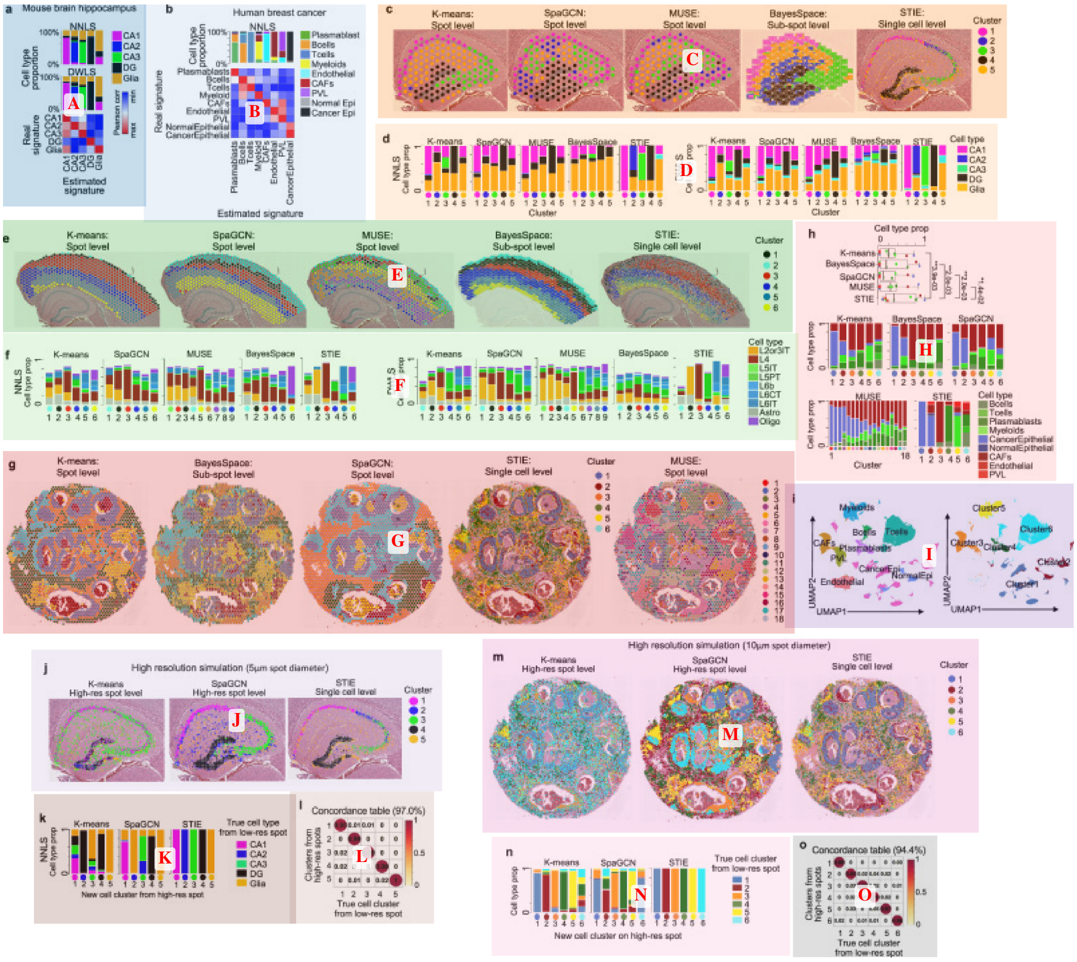


Figure 7: Example compound figure with human-verified subfigure boxes and labels (A–O).

Subcaptions:

- **A**: Cell-type signature matrix learned from 10X Visium mouse brain hippocampus FFPE.

- **B**: Cell-type signature matrix learned from 10X Visium human breast-cancer FFPE.

- **C**: Spot-level clustering by K-means, SpaGCN, MUSE; sub-spot-level clustering by BayesSpace; and single-cell-level clustering by STIE on 10X Visium FFPE mouse brain hippocampus.

- **D**: Cell-type deconvolution of spot-, sub-spot-, and single-cell-level clustering-derived CAGE in the mouse brain hippocampus.

- **E**: Spot-level clustering by K-means, SpaGCN, MUSE; sub-spot-level clustering by BayesSpace; and single-cell-level clustering by STIE on 10X Visium FFPE mouse brain cortex.

- **F**: Cell-type deconvolution of spot-, sub-spot-, and single-cell-level clustering-derived CAGE in the mouse brain cortex (rare non-cortical cell types omitted from barplot).

- **G**: Spot-level clustering by K-means and SpaGCN; sub-spot-level clustering by BayesSpace; and single-cell-level clustering by STIE on 10X Visium FFPE human breast cancer.

- **H**: Cell-type deconvolution of spot-, sub-spot-, and single-cell-level clustering-derived CAGE in human breast cancer. The box plot (h) represents the deconvoluted proportion of 9 cell types, where center line represents median, lower and upper hinges represent first and third quartiles, and whiskers extend from hinge to $\pm 1.5 \times$ IQR. The p-value was calculated based on one-sided Wilcoxon signed-rank test without adjustment for multiple comparisons.

- **I**: The UMAP plot of human breast cancer scRNA-seq data from 26 primary tumors. The top panel is the original cell typing of 10,060 single cells, and the bottom panel is the subset of cells that are mapped to the six STIE clusters.

- **J**: Spot-level clustering by K-means (left), SpaGCN (middle), and single-cell-level clustering by STIE (right) on the simulated high-resolution spot spatial transcriptome data of the mouse brain hippocampus.

- **K**: Cell-type deconvolution of spot- and single-cell-level clustering-derived CAGE in the simulated high-resolution mouse hippocampus dataset.

- **L**: The consistency table of single-cell clusters between the simulated high-resolution spot-based STIE clustering and the original low-resolution spot-based STIE clustering as ground truth of the mouse hippocampus.

- **M**: Spot-level clustering by K-means (left), SpaGCN (middle), and single-cell-level clustering by STIE (right) on the simulated high-resolution spot spatial transcriptome data of the human breast-cancer spatial transcriptomes.

- **N**: Cell-type deconvolution of spot- and single-cell-level clustering-derived CAGE in the simulated high-resolution human breast-cancer dataset.

- **O**: The consistency table of single-cell clusters between the simulated high-resolution spot-based STIE clustering and the original low-resolution spot-based STIE clustering as ground truth of human breast cancer.