



FigEx: Aligned Extraction of Scientific Figures and Captions

Jifeng Song^{1,2}, Arun Das^{2,3}, Ge Cui¹, Yufei Huang^{1,2,3}

¹ Department of Electrical and Computer Engineering, University of Pittsburgh; ² Hillman Cancer Center, University of Pittsburgh Medical Center; ³ Department of Medicine, University of Pittsburgh

Motivation & Contribution

- Motivation:** Automatic understanding of figures in scientific papers is challenging since they often contain subfigures and subcaptions in complex layouts.

Contribution:

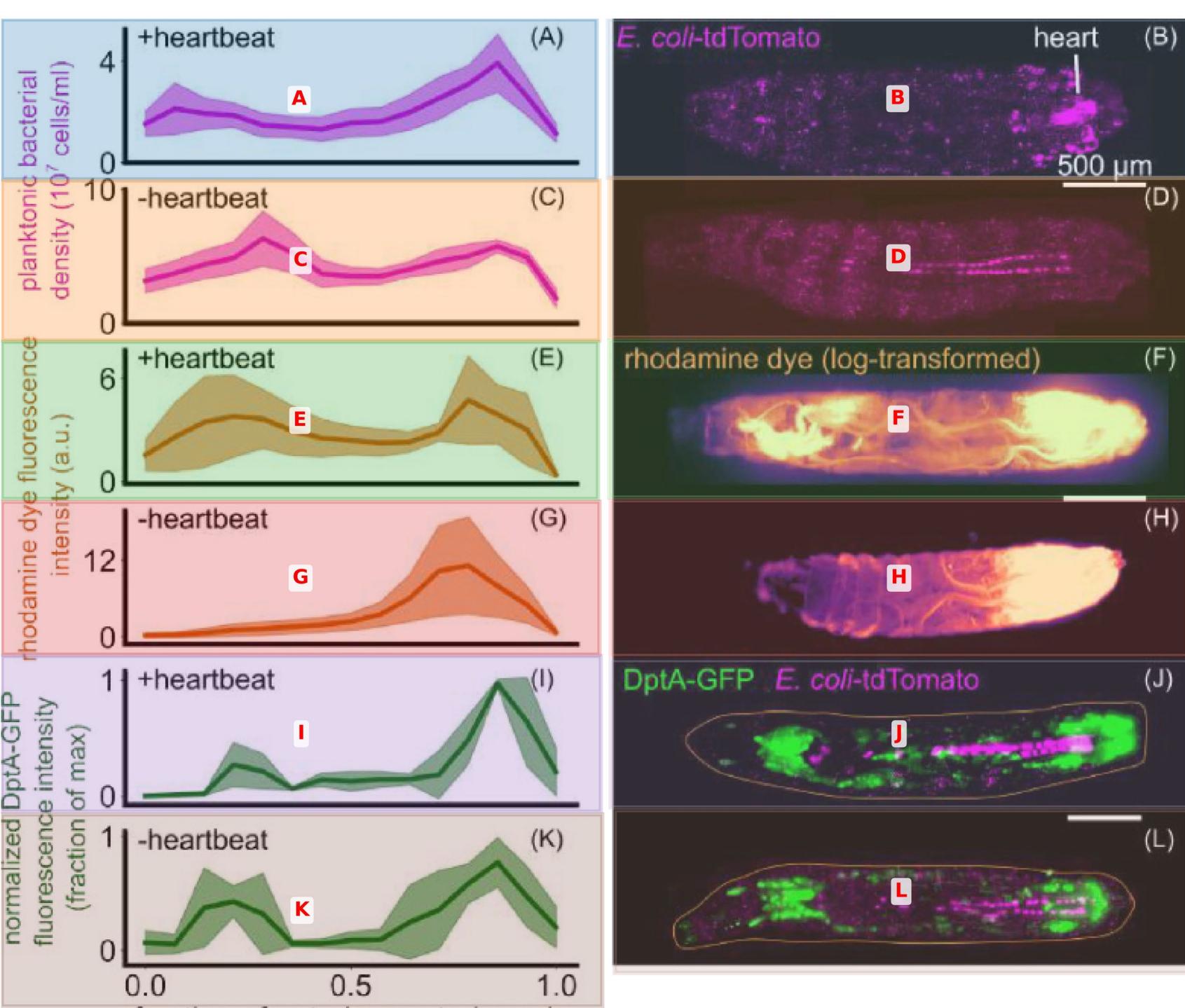
- We propose **FigEx-7B**, a compact vision-language model for aligned extraction of scientific figures and captions.
- We curate **BioSci-Fig**, a dataset of 7,174 compound figures with meticulously annotated bounding boxes and aligned subcaptions, providing a benchmark setting for compound figure separation 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.

BioSci-Fig: Curated Dataset

- BioSci-Fig:** A dataset of **7,174** compound figures with **43,183** annotated bounding boxes and aligned subcaptions, providing a standardized benchmark for compound figure separation in scientific documents.

Example

Caption: Heartbeat-induced fluid flows pattern bacteria and dye but are not required for patterning of DptA. Each row shows quantification (left, mean and standard deviation) and a representative image (right) of various quantities. (A)-(D) E. coli 3 hours post injection with and without a heartbeat (larvae per group). In the quantification, to avoid counting fluorescence internalized by host cells, planktonic bacteria freely suspended in the hemolymph were computationally identified and only these cells were counted (Methods). The heartbeat was eliminated by myosin knockdown in the heart using NP1029-Gal4 x UAS-Mhc-RNAi. (E)-(H) Rhodamine dye injected in the posterior and imaged 5 minutes after injection, with and without a heartbeat (larvae per group). (I)-(L) DptA-GFP 6 hours post injection in animals with and without a heartbeat (larvae per group). All scale bars are 500 μ m. In (J) and (L), the approximate outline of the larva is marked as an orange line. Images in (B), (D), (J), and (L) are maximum intensity projections of 3D light sheet images stacks. Images in (F) and (H) are single-plane widefield images.

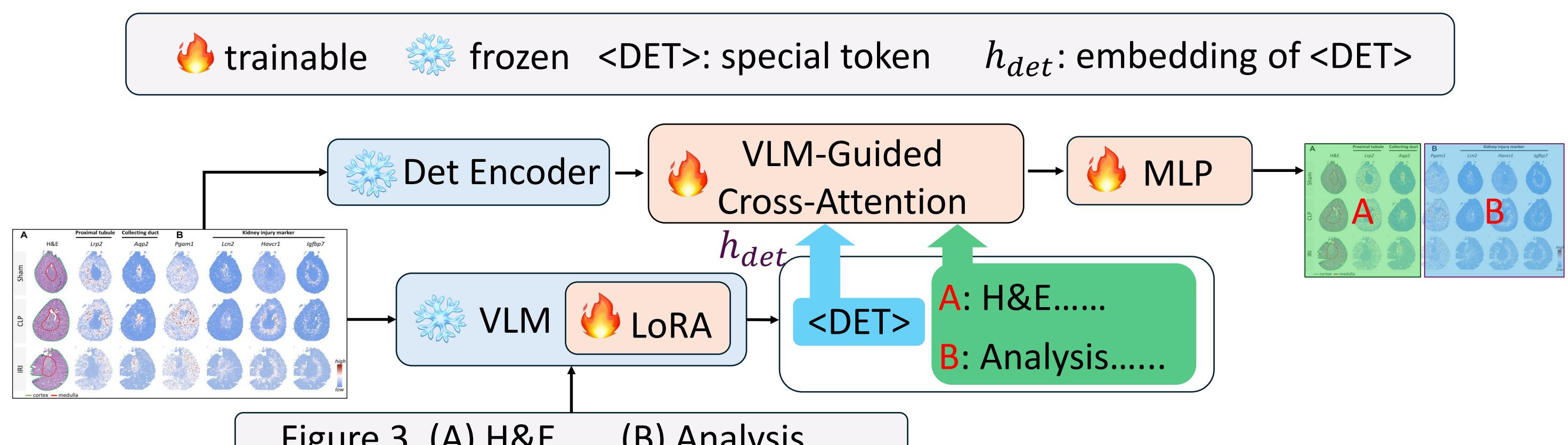


Subcaptions:
A: Spatial profile of planktonic E. coli density along the larval anterior–posterior axis 3 h after injection in heartbeat-positive animals (magenta curve, mean \pm SD).
B: Representative light-sheet projection showing freely suspended E. coli (magenta) in a control larva with active heartbeat; heart location indicated.
C: Spatial profile of planktonic E. coli density in heartbeat-ablated larvae (myosin knock-down), revealing reduced anterior transport.
D: Representative projection of heartbeat-negative larva showing posteriorly restricted planktonic bacteria.
E: Spatial profile of injected rhodamine dye fluorescence 5 min post-injection in heartbeat-positive larvae, illustrating anterior advection.
F: Single-plane widefield image of rhodamine distribution in a beating-heart larva (log-intensity).
G: Spatial profile of rhodamine fluorescence in heartbeat-negative larvae, displaying minimal anterior spread.
H: Widefield image of dye confinement to posterior in a larva lacking heartbeat.
I: Spatial profile of innate-immune reporter DptA-GFP intensity 6 h post injection in heartbeat-positive larvae, peaking at anterior and posterior ends.
J: Projection image showing DptA-GFP (green) and E. coli (magenta) in a control larva; larval outline traced in orange.
K: Spatial profile of DptA-GFP intensity in heartbeat-negative larvae, demonstrating a similar biphasic pattern despite absent flow.
L: Projection of DptA-GFP and E. coli in a heartbeat-ablated larva, confirming flow-independent immune patterning.

Method

Overview of the FigEx architecture

FigEx separates a compound figure and caption into subfigures and subcaptions with a special bridge token <DET> between VLM and detection backbone.



Three-stage training of FigEx

Stage 1: LoRA adaptation

We update only the LoRA layers in the VLM.

Stage 2: Detection head training

Fine-tune vision encoder and MLP detection head.

Stage 3: Joint fine-tuning

Fine-tune LoRA adapters, the VLM-guided cross-attention, and the MLP detection head.

VLM-guided cross-attention

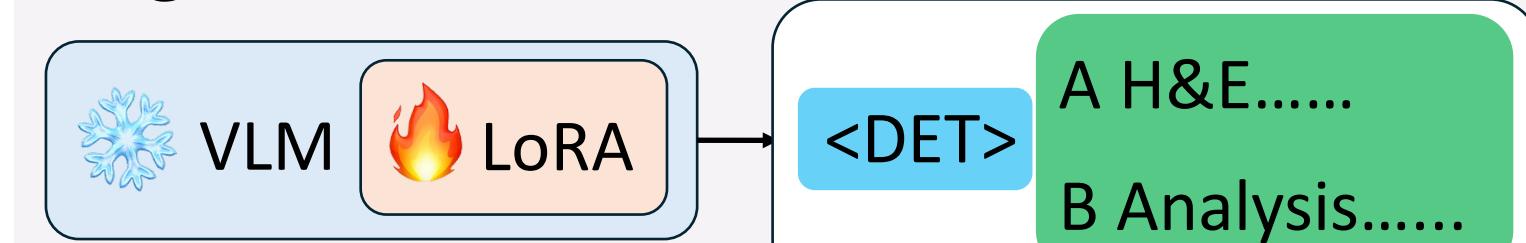
Det-Text Cross-Attention

We refine image features with two cross-attention modules. First, the detection-text module $\mathcal{F}_{det-text}$ fuses the detection-token hidden states h_{det} with the sub-caption feature f_{subcap} .

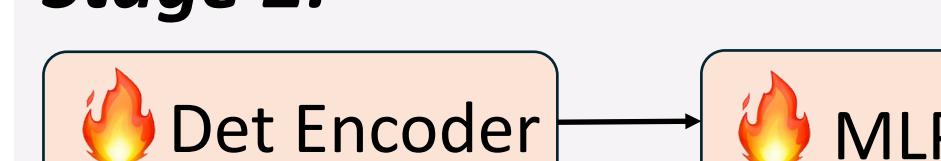
Det-Text Cross-Attention

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

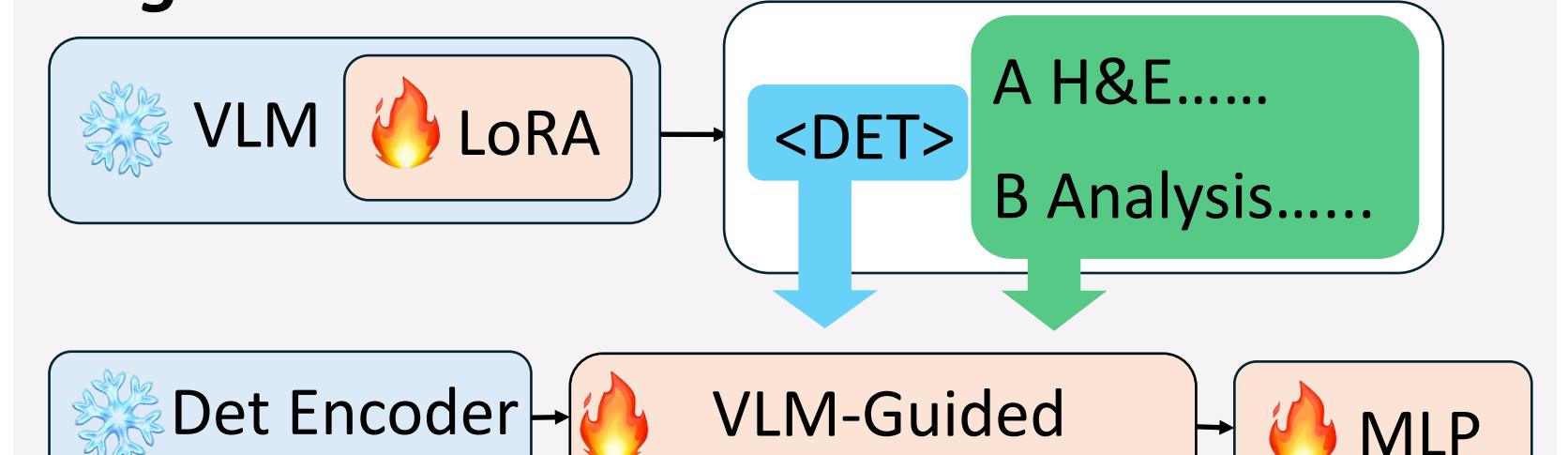
Stage 1:



Stage 2:



Stage 3:

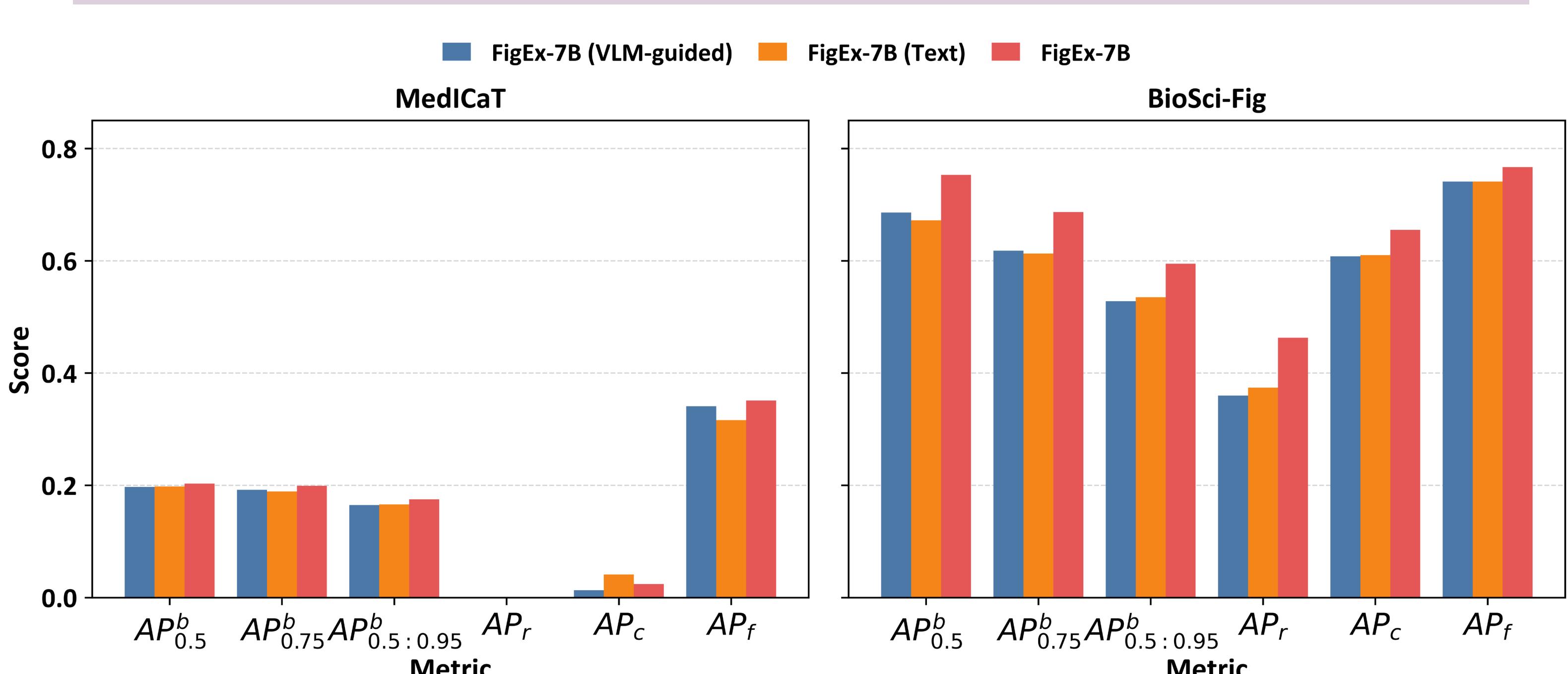


Experiments

Evaluation of subfigure detection

Model	Vision Backbone	Dataset	$AP_{0.5}^b$	$AP_{0.75}^b$	$AP_{0.5:0.95}^b$	AP_r	AP_c	AP_f
YOLO11n	/	MedIcaT	0.052	0.065	0.054	0	0	0.110
YOLO11l	/		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
YOLO11l	/		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

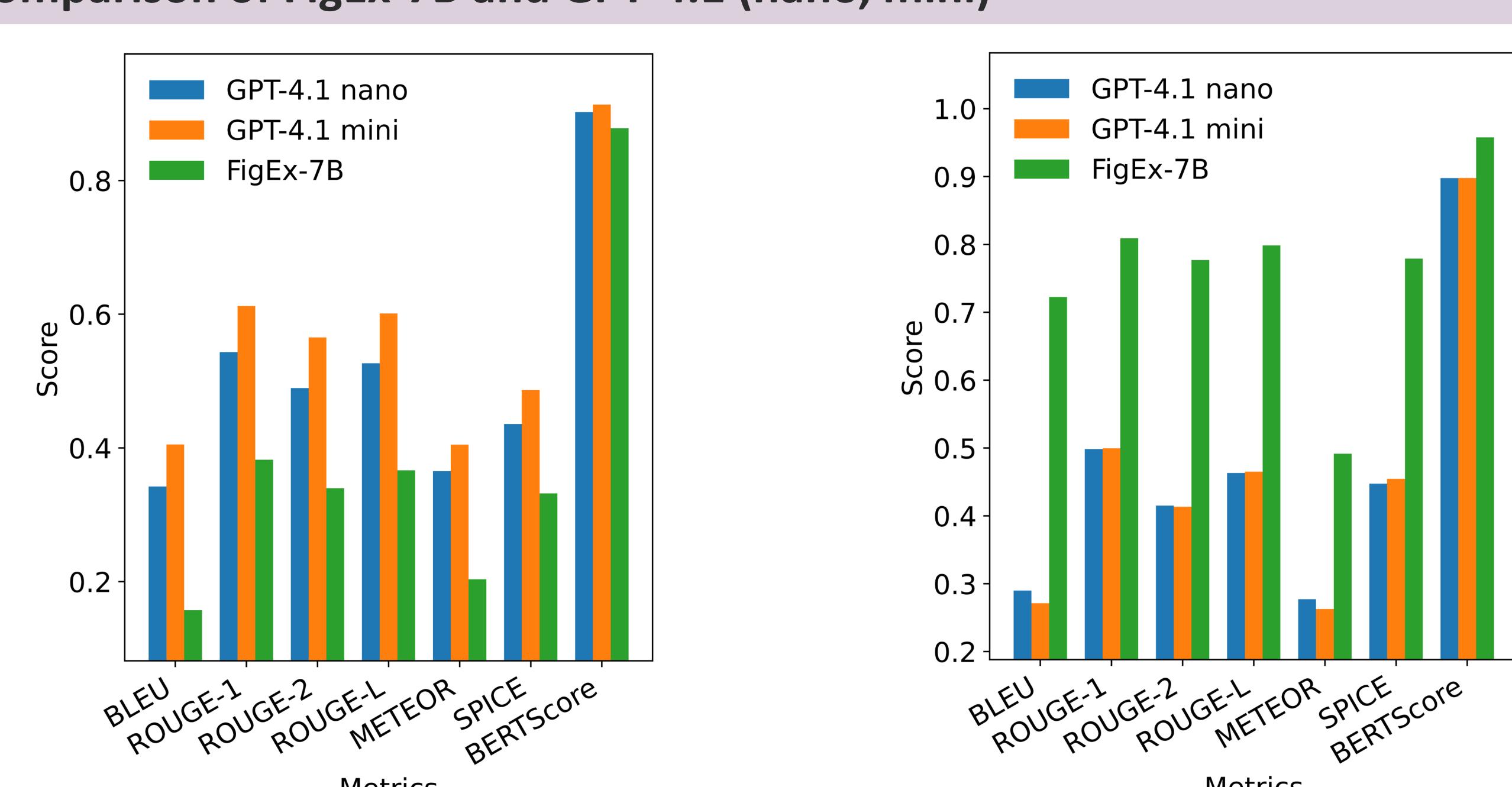
Ablation Study



Evaluation of caption separation

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

Comparison of FigEx-7B and GPT-4.1 (nano, mini)



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.