Supplementary Material

1. Panel2Patch Details

Here we provide additional implementation and prompt-design details for our proposed Panel2Patch pipeline, which converts complex biomedical figures into a set of localized, semantically grounded patches. The pipeline consists of four main stages: (i) SoM-guided panel decomposition, (ii) marker-guided region mining, (iii) caption-box mining, and (iv) bounding-box post-processing. Unless otherwise stated, all stages use the same LVLM configuration described in Sec. 1.3.

1.1. Prompts for SoM-Guided Panel Decomposition

As shown in Fig. 1, we design a structured prompt that instructs the LVLM to first identify the global layout of a multi-panel figure and then enumerate all constituent panels together with their spatial extents. The prompt explicitly asks the model to:

1. Determine whether the input is a single-panel or multi-panel figure.
2. If multi-panel, list all panels in reading order (e.g., A, B, C, . . .).
3. For each panel, return an axis-aligned bounding box in normalized coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ and a short description summarizing its content.

To make parsing more robust, we embed SoM-style few-shot exemplars in the prompt that illustrate typical biomedical layouts, including grids of microscopy images, plots with shared axes, and schematic diagrams. These examples help the LVLM distinguish true content panels from auxiliary elements such as legends, color bars, or scale bars, which should not be treated as separate panels.

1.2. Prompts for Marker-Guided Region Mining and Captioning

Marker Boxes. We further exploit fine-grained visual cues such as arrows, circles, stars, and other graphical markers that highlight regions of interest in biomedical figures (e.g., lesions, cells, or anatomical structures). As illustrated in Fig. 2, we prompt the LVLM to:

1. Identify all visible visual markers in the image.
2. For each marker, infer its semantic role (e.g., "lesion", "tumor boundary", "positive cells") using the figure caption and article title as context.
3. Return the bounding box of the *target region* being highlighted, rather than the marker glyph itself.

By grounding markers in the regions they point to, Panel2Patch captures subtle but clinically relevant structures that may occupy only a small fraction of the panel.

Caption Boxes. In addition to explicit markers, many figures contain local captions, labels, or inset annotations that implicitly define regions of interest (e.g., "Zoomed-in view of region A", "Tumor core", "Control"). As shown in Fig. 3, we design a complementary prompt that:

1. Parses the caption text to extract candidate objects and regions explicitly mentioned in the figure description.
2. Verifies that each candidate object is visibly present in the image.
3. Returns tight bounding boxes and short descriptions for objects that are both mentioned in the caption and clearly visible.

This caption-guided mining step focuses on "contextual objects" that are grounded in both text and image, complementing the marker-guided regions that are purely visually highlighted.

Combining marker-guided and caption-guided mining yields a richer set of semantically grounded regions, going beyond coarse panel crops to capture both visually emphasized and textually emphasized structures.

1.3. LVLM Configs

We run Qwen2.5-VL-72B in **float32** precision using vLLM with tensor parallel size 4, GPU memory utilization 0.8, maximum sequence length 8192, and up to 512 concurrent sequences. Decoding uses nucleus sampling with temperature 0.2, top-p 0.9, top-k 50, repetition penalty 1.05, and a maximum of 128 generated tokens. Unless otherwise specified, these settings are used consistently for panel parsing, caption decomposition, marker-guided region mining, and region captioning.

1.4. Bounding-Box Processing Details

After obtaining raw bounding boxes from panel decomposition, marker mining, and caption mining, we perform several post-processing steps before constructing the final training set. Concretely, we:

1. Convert all coordinates to absolute pixel values and clip them to the image boundaries.
2. Discard degenerate boxes with extremely small area or aspect ratios outside a valid range.
3. Apply class-agnostic non-maximum suppression (NMS) within each figure using an IoU threshold of 0.7 to remove near-duplicate boxes produced by different prompts.
4. Merge overlapping boxes that share highly similar textual descriptions, measured by a sentence-embedding cosine similarity above 0.9.
5. Normalize and store the final boxes as image–bbox–text triplets.

Prompt for SoM-Guided Panel Decomposition

```
PROMPT = (
    "Subtask 1: Identify and describe panels in the image.\n\n"
    "You are given the image along with structured metadata including the article title, abstract, and caption.\n"
    "Use this context to understand the image and describe the panels meaningfully.\n\n"
    "Instructions:\n"
    "- If the image contains multiple panels (e.g., labeled A, B, C...), segment the image into distinct panels.\n"
    "- If the image shows only a single panel, treat the whole image as one panel.\n"
    "- Each panel must include:\n"
    "  - panel_id: an id (e.g., A, B, C, 1, 2, 3) or 'A' if single-panel\n"
    "  - bbox: bounding box in [x1, y1, x2, y2] format, where x1 < x2 and y1 < y2\n"
    "  - caption: the exact subcaption text from the original caption that corresponds to this panel (do not rewrite)\n"
    "  - description: a 1-2 sentence explanation of what the panel illustrates, using context from title, abstract, and caption\n\n"
    "Output Format:\n"
    "Return a single valid JSON object with key 'panels'.\n"
    "The JSON must be strictly formatted like this:\n"
    "{\n"
    '  "panels": [\n'
    "    {\n"
    '      "panel_id": "A",\n'
    "      \"bbox\": [100, 200, 400, 600],\n"
    '      "caption": "Panel A shows laparoscopic view of cyst resection.",\n'
    '      "description": "This panel depicts the surgical exposure of a ruptured cyst guided by the laparoscopic grasper, as
discussed in the caption and abstract."\n'
    "    },\n"
    "    ...\n"
    "  ]\n"
    "}\n\n"
    "Rules:\n"
    "- The 'caption' field must always be directly taken from the original caption text, not paraphrased.\n"
    "- Return only a valid JSON object.\n"
    "- Do not include any extra text or explanation.\n"
    "- If JSON is malformed, the system will fail to parse your output.\n"
```

Figure 1. **Prompt for SoM-guided panel decomposition.** We show an example prompt and LVLM response for decomposing a multi-panel biomedical figure into individual panels, each with a panel ID, bounding box, and short description. The prompt enforces a strict JSON schema, which facilitates reliable downstream parsing.

This processing significantly improves the spatial precision and diversity of regions while avoiding excessive redundancy, as reflected by the visual examples in Fig. 4 and Fig. 6.

2. Dataset Details

2.1. Data Sampling

We build our pretraining corpus on top of the Biomedica figure–caption dataset [1]. Starting from all available figures, we sample approximately 350 k multimodal figures and retain those that:

1. provide an associated figure caption in English,
2. are released under a license compatible with research usage,
3. are annotated as scientific figures (excluding logos, advertisements, and non-scientific graphics),
4. contribute to a reasonably balanced distribution of modalities.

We then apply Panel2Patch to each sampled figure to obtain panel-level crops and region-level boxes, as detailed in Sec. 2.2.

2.2. Dataset Statistics

Our final pretraining dataset contains:

- 364,216 figure-level image–caption pairs,
- 1,303,950 panel-level image–caption pairs after panel decomposition,

Prompt for SoM-Guided Panel Decomposition

```
PROMPT = (
    "Subtask 2: Detect and describe visual markers in the image.\n\n"
    "Visual markers include elements such as arrows, circles, boxes, stars, highlights, or other graphical annotations that point to or emphasize specific regions or objects in the image.\n\n"
    "You are given the image along with contextual meta data (article title and image caption). Use this context to understand the function of each marker.\n\n"
    "Instructions:\n"
    "- Identify all visible visual markers in the image.\n"
    "- For each marker, return:\n"
    "  - marker_id: a string identifier like '1', '2', etc.\n"
    "  - target_bbox: bounding box of the object/region being pointed to, in [x1, y1, x2, y2] format\n"
    "  - caption: a short label or title for the marker target (e.g., 'arrow pointing to tumor')\n"
    "  - description: a one-sentence explanation of what the marker highlights or emphasizes, based on both the image and context\n\n"
    "Output Format:\n"
    "Return a single valid JSON object with key 'markers'.\n"
    "The JSON must be strictly formatted like this:\n"
    "{\n"
    '  "markers": [\n'
    "    {\n"
    '      "marker_id": "1",\n'
    '      "target_bbox": [150, 120, 300, 250],\n'
    '      "caption": "Arrow indicating liver mass",\n'
    '      "description": "This arrow highlights a suspected hepatic lesion located in the upper right quadrant, as described in the caption."\n'
    "    },\n"
    "    ...\n"
    "  ]\n"
    "}\n\n"
    "Rules:\n"
    "- Only return a valid JSON object.\n"
    "- Do not include explanations or markdown outside the JSON.\n"
    '- If no markers are present, return an empty list: { "markers": [] }\n'
```

Figure 2. **Prompts for marker-guided region mining.** Example prompt and LVLM output for detecting visual markers (arrows, stars, etc.) and producing bounding boxes for the corresponding target regions. Each entry includes a marker ID, target box, short label, and a one-sentence description.

- 619,424 image–bbox–text triplets obtained from marker-guided region generation, and
- 1,030,194 bounding boxes generated using the LVLM's internal knowledge (e.g., inferred object regions without explicit markers).

We take the union of marker-guided and LVLM-inferred boxes and apply the bounding-box processing procedure from Sec. 1.4 and Fig. 4 before using them for pretraining.

The 364,216 figure-level image–text pairs are drawn from Biomedica. We crawl the associated metadata and classify each figure into a primary and (optionally) secondary label corresponding to its dominant visual type (e.g., plots, microscopy, chemical structures). The distribution over secondary labels is long-tailed and reflects the diversity of real-world biomedical figures. As summarized in Tab. 1 and Tab. 2, common categories include plots, bar plots, scientific illustrations, microscopy, and chemical diagrams, while many specialized categories (e.g., laryngoscopy, karyotype) appear at much lower frequencies. This diversity is beneficial for learning robust visual representations that generalize across modalities and scales.

2.3. Quality of Annotations from Panel2Patch

LVLM for Panel Parsing. We empirically observe that stronger LVLMs lead to more accurate and consistent panel decompositions. As shown qualitatively in Fig. 5, Qwen2.5-VL-72B produces more precise panel boundaries, avoids merging unrelated subpanels, and yields more informative

Prompt for SoM-Guided Panel Decomposition

```
PROMPT =
    "Task: Find contextual objects — items that are BOTH visible in the image AND explicitly mentioned in the provided
text.\n\n"
    "You are given the text:\n"
    "- caption: the figure caption\n\n"
    "Procedure:\n"
    "1) Read the title and caption. Extract candidate object mentions (noun phrases) that appear verbatim or as clear
synonyms.\n"
    "2) For each candidate, check the image. Keep it only if the object is clearly visible.\n"
    "3) For each kept object, return a tight bounding box and a short description grounded in the image evidence.\n"
    "Requirements:\n"
    "- Return ONLY a valid JSON object with key 'contextual_objects'.\n"
    "- Use double quotes for all strings.\n"
    "- Coordinates are pixel indices with origin at top-left; bbox format is [x1, y1, x2, y2] with x1<x2, y1<y2 (integers).\n"
    "- Boxes must be tight around the visible object.\n"
    "- Do not infer objects that are not visibly present.\n"
    "- If nothing qualifies, return { \"contextual_objects\": [] }.\n\n"
    "Output schema:\n"
    "{\n"
    "  \"contextual_objects\": [\n"
    "    {\n"
    "      \"type\": \"<object label>\",\n"
    "      \"bbox\": [x1, y1, x2, y2],\n"
    "      \"description\": \"<brief, image-grounded description>\"\n"
    "    }\n"
    "  ]\n"
    "}\n\n"
    "Rules:\n"
    "- Return only a valid JSON object.\n"
    "- Use double quotes for all strings.\n"
    "- If no contextual objects are found, return:\n"
    '{ "contextual_objects": [] }\n'
    "- Do not include explanations or formatting outside the JSON output.\n"
```

Figure 3. **Prompts for caption-guided region mining.** Example prompt and LVLM response for extracting bounding boxes and descriptions of contextual objects that are jointly grounded in the figure caption and the image content. Only objects that are explicitly mentioned and visually present are retained.

textual summaries compared to a smaller variant (e.g., 32B). In particular, the 72B model better respects subtle layout cues such as shared axes, legends, and insets, which are common in biomedical figures.

LVLM for Biomedical Region Mining. For region-level mining, we find that LVLMs can reliably localize clinically meaningful structures when guided by explicit prompts about markers and captions. Fig. 6 shows qualitative examples where the model correctly highlights lesions, cell clusters, and anatomical regions indicated by arrows or local labels. Combined with the post-processing described in Sec. 1.4, this yields high-quality region annotations suitable for training fine-grained biomedical vision–language models.

3. Implementation Details

We perform continual pretraining by initializing from a strong Biomedica-trained CLIP backbone (ViT-L-14) and adapting it on the Panel2Patch corpus. Training is run with distributed data parallelism over 4 GPUs using a per-GPU batch size of 40 (effective batch size 160) for 20 epochs in `fp32` precision. We jointly optimize global figure-level objectives and region-level objectives for panels, fine regions, and coarse regions.

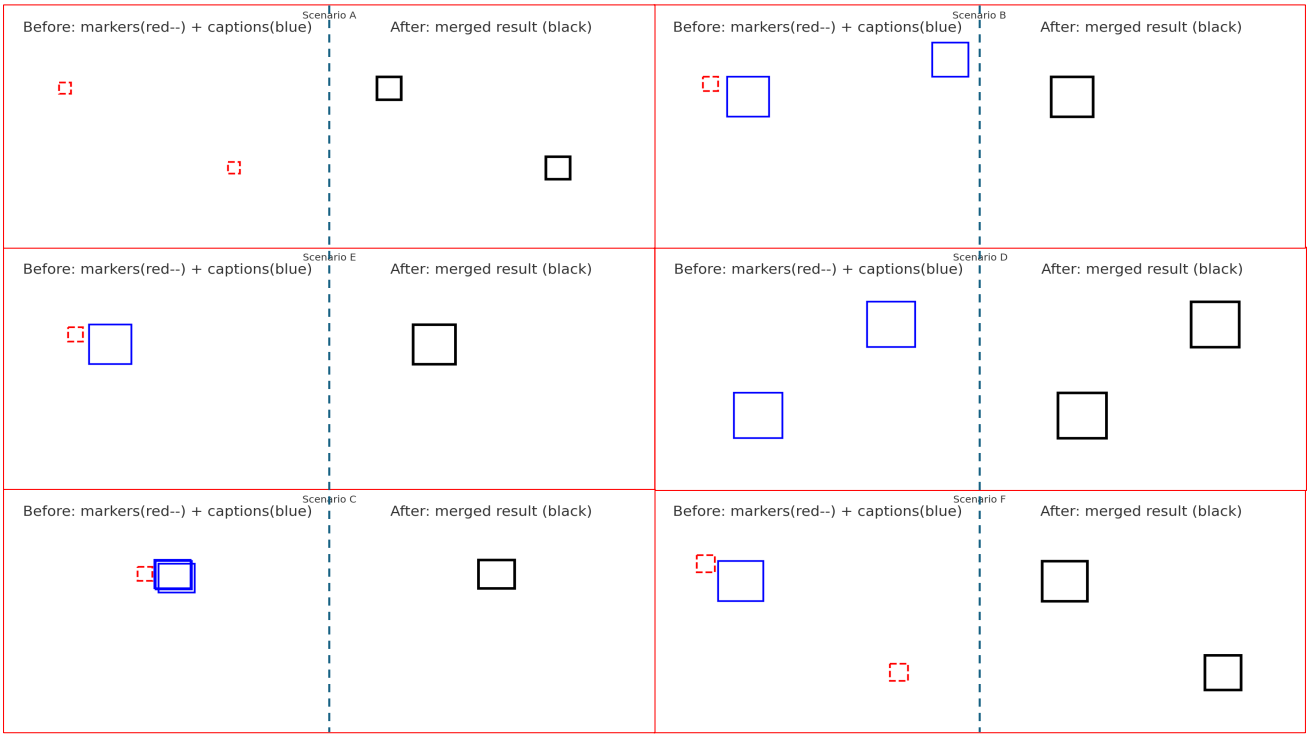

Figure 4. **Bounding-box post-processing pipeline.** Illustration of how our box post-processing method handles bounding boxes from different sources. We simulate different scenarios to demonstrate the method's robustness and precision.

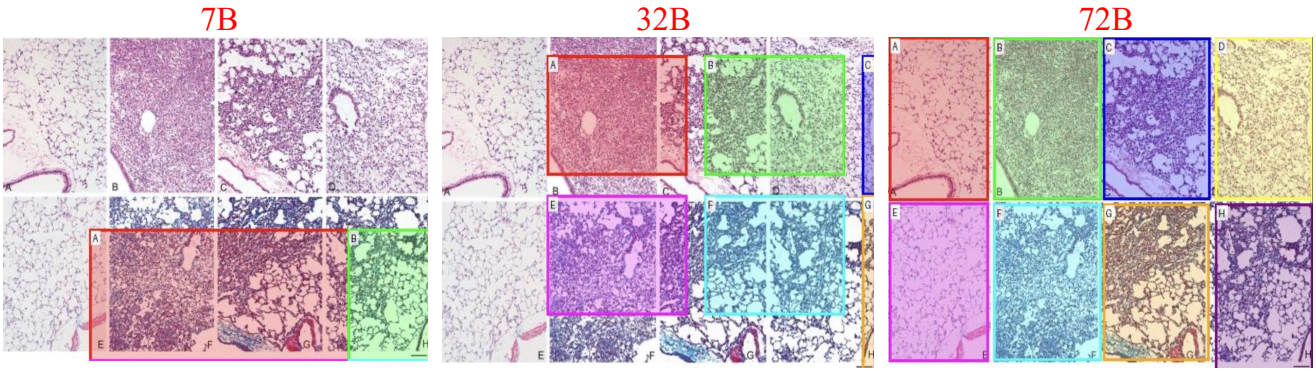

Figure 5. **Qualitative comparison of panel parsing.** Visualization of panel decomposition results for different LVLM capacities. From left to right, we show outputs from 7B, 32B, and 72B variants. Larger models (e.g., 72B) tend to produce more accurate panel boundaries and richer descriptions.

3.1. Hyper-parameters

We use AdamW throughout with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.05. Optimization is decoupled into a shared backbone branch and three region heads (bbox, fine, coarse), each with its own learning rate and cosine schedule:

- shared/global branch: learning rate 1×10^{-5},
- bbox head: learning rate 5×10^{-6},
- fine head: learning rate 1×10^{-5},
- coarse head: learning rate 1×10^{-5}.

All four branches use cosine learning-rate decay with a linear warmup of $1{,}000$ updates.

3.2. Alternating Training

To preserve general-purpose captioning ability and mitigate catastrophic forgetting, we alternate Panel2Patch data with the original Biomedica figure–caption pairs [1]. Concretely, we sample 1 million figure-level image–caption pairs from the raw Biomedica dataset and mix them with our struc-

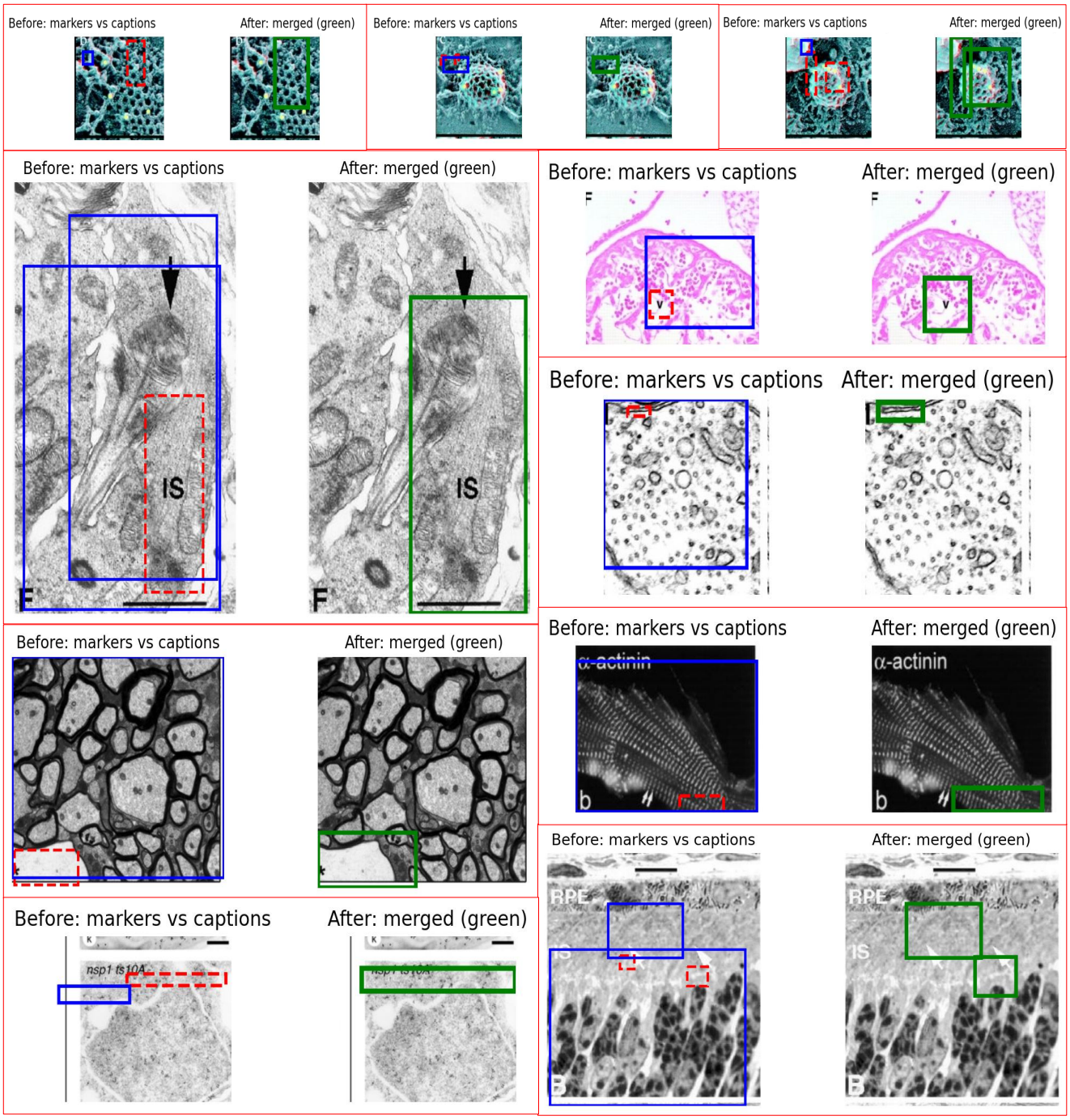

Figure 6. **Qualitative examples of region mining.** Visualizations of bounding boxes obtained from marker-guided and caption-guided mining after post-processing. Highlighted regions correspond to clinically or scientifically relevant structures emphasized in the figure.

tured panel/region data. During training, each mini-batch is drawn from this mixture, and within Panel2Patch samples we randomly interleave bbox, fine, and coarse crops. This setup keeps the model exposed to both holistic figure-level captions and fine-grained region-level supervision, improving localized grounding while maintaining performance on generic figure captioning benchmarks.

Table 1. Secondary labels and percentages of figure types (Part 1).

Label	Pct	Label	Pct
plot	0.2392039961	bar plot	0.1150651605
line plot	0.0731292006	scientific illustration	0.0560098810
2D chemical reaction	0.0410442131	microscopy	0.0392800450
diagram	0.0284265115	signal plot	0.0271832018
2D chemical structure	0.0268386763	light microscopy	0.0218300065
table	0.0216904796	immunohistochemistry	0.0211481985
flowchart	0.0179900438	immunoblot	0.0178245154
scatter plot	0.0139932794	matrix plot	0.0127300424
natural image	0.0121686572	forest plot	0.0120581146
map	0.0120131838	3D chemical structure	0.0106260247
computerized tomography	0.0106139191	signaling pathway	0.0103486139
graph	0.0095768316	lab equipment	0.0093916710
fluorescence microscopy	0.0093163344	confocal microscopy	0.0092534491
3D plot	0.0088056319	electron microscopy	0.0078804112
box plot	0.0074682040	3D protein structure	0.0069536449
heatmap plot	0.0067048960	clinical imaging	0.0065234831
laboratory specimen	0.0055550288	pie chart	0.0053961494
x-ray radiography	0.0053488378	assay	0.0051444683
specimen	0.0045556989	insects	0.0040846959
venn diagram	0.0038361605	cohort selection flowchart	0.0037718538
brain	0.0033478943	ambiguous	0.0032435385
skin lesion	0.0030236729	face	0.0028432698
ultrasound	0.0028265426	scanning electron microscopy	0.0028222899
surgical procedure	0.0026714498	histogram	0.0026000397
bacterial culture	0.0023106903	phylogenetic tree	0.0021169777
sequence plot	0.0019832298	nature	0.0019563777
endoscopy	0.0018649276	network	0.0017417046
electronic circuit	0.0017399841	epifluorescence microscopy	0.0017307059
tree	0.0016858721	angiography	0.0016805242
gel electrophoresis	0.0016659331	electrocardiography	0.0016465339

4. External Evaluation Benchmarks

Our evaluation benchmarks in the main paper are strictly non-overlapping with the pretraining data.

4.1. Cross-modal Retrieval

As shown in Tab. 2 of the main paper, we consider two retrieval-based setups: (i) figure-level retrieval and (ii) region-level retrieval. To avoid contamination, we construct both benchmarks from PubMed articles published in 2025 that are not part of Biomedica [1] and are never used in our pretraining pipeline. For each benchmark, we run the same panel splitting and bounding-box generation pipeline used for pretraining. This results in two evaluation sets that mirror our pretraining data distribution while remaining strictly disjoint in both images and text. We additionally perform manual checks to remove any residual overlaps.

4.2. Zero-shot Classification

For zero-shot classification experiments, we follow the setup of [2] to ensure a fair comparison. In particular, we adopt the same label spaces, test splits, and evaluation metrics. Our method differs only in the visual–textual encoder and training data; we do not introduce any task-specific tuning or additional supervision beyond what is permitted in that protocol.

4.3. Ablation Studies

Due to the cost of continual pretraining, we conduct ablations on a reduced subset of the pretraining corpus. Specifically, we sample an additional $100\,\mathrm{k}$ figures from our processed dataset and re-run training under modified settings (e.g., removing marker-guided regions or LVLM-generated boxes). This allows us to isolate the contribution of each component while keeping training affordable.

Table 2. Secondary labels and percentages of figure types (Part 2).

Label	Pct	Label	Pct
magnetic resonance	0.0016456367	dot plot	0.0016198410
functional magnetic resonance	0.0016157048	survival curve	0.0014810793
radial plot	0.0014789354	intraoral imaging	0.0014599673
human	0.0014181240	density plot	0.0013346044
circos plot	0.0013059153	user interface	0.0011404918
tool	0.0010539856	roc curve	0.0009932869
optical coherence tomography	0.0009341144	screenshot	0.0008343880
violin plot	0.0008291993	metabolic pathway	0.0008278167
phase contrast microscopy	0.0008206435	transmission electron microscopy	0.0007905368
aerial photography	0.0007299856	circular plot	0.0007264203
RT PCR	0.0007226259	teeth	0.0006845110
process chart	0.0006718811	checklist table	0.0005735024
3D chemical reaction	0.0005326223	intraoperative image	0.0005291735
procedural image	0.0005042245	neural network	0.0004901149
patient photo	0.0004813765	system diagram	0.0004802114
manuscript	0.0004780481	reagents	0.0004636123
humans and devices	0.0004603306	algorithm	0.0004338552
drawing	0.0004153686	immunoassay	0.0003963423
plot and chart	0.0003894797	flow diagram	0.0003860271
differential gene expression matrix	0.0003814792	medical equipment	0.0003706203
laryngoscopy	0.0003408360	2D mesh	0.0003278449
eye	0.0003246564	flowcytometry	0.0003243030
human head	0.0003202484	word cloud	0.0003137431
skull	0.0002876872	list	0.0002694608
funnel plot	0.0002647887	immunocytochemistry	0.0002596583
illustration	0.0002404416	pyramid chart	0.0001193004
mammography	0.0000844672	karyotype	0.0000812088

4.4. Cell Imaging

We further evaluate our approach on cell imaging tasks to assess its ability to capture fine-grained morphological cues.

As illustrated in Fig. 7, we visualize retrieval and captioning results on a challenging cell-imaging benchmark (e.g., fluorescence or bright-field microscopy). Our model is able to localize relevant subcellular regions (e.g., nuclei, cytoplasm, membrane structures) using the mined patches. It can also retrieve semantically similar cell images given a textual query describing morphological patterns (e.g., "cells with fragmented nuclei") and generate concise and accurate descriptions of cell states (e.g., apoptosis, mitosis, or abnormal morphology).

Qualitatively, the model trained with Panel2Patch better focuses on the discriminative parts of the cells than a baseline trained only on figure-level data, leading to improved alignment between textual descriptions and visual evidence.

References

[1] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature, 2025. 2, 5, 7

[2] Min Woo Sun, Alejandro Lozano, Javier Gamazo Tejero, Vishwesh Nath, Xiao Xiao Sun, James Burgess, Yuhui Zhang, Kun Yuan, Robert Tibshirani, Sean Huver, et al. No tokens wasted: Leveraging long context in biomedical vision-language models. *arXiv preprint arXiv:2510.03978*, 2025. 7

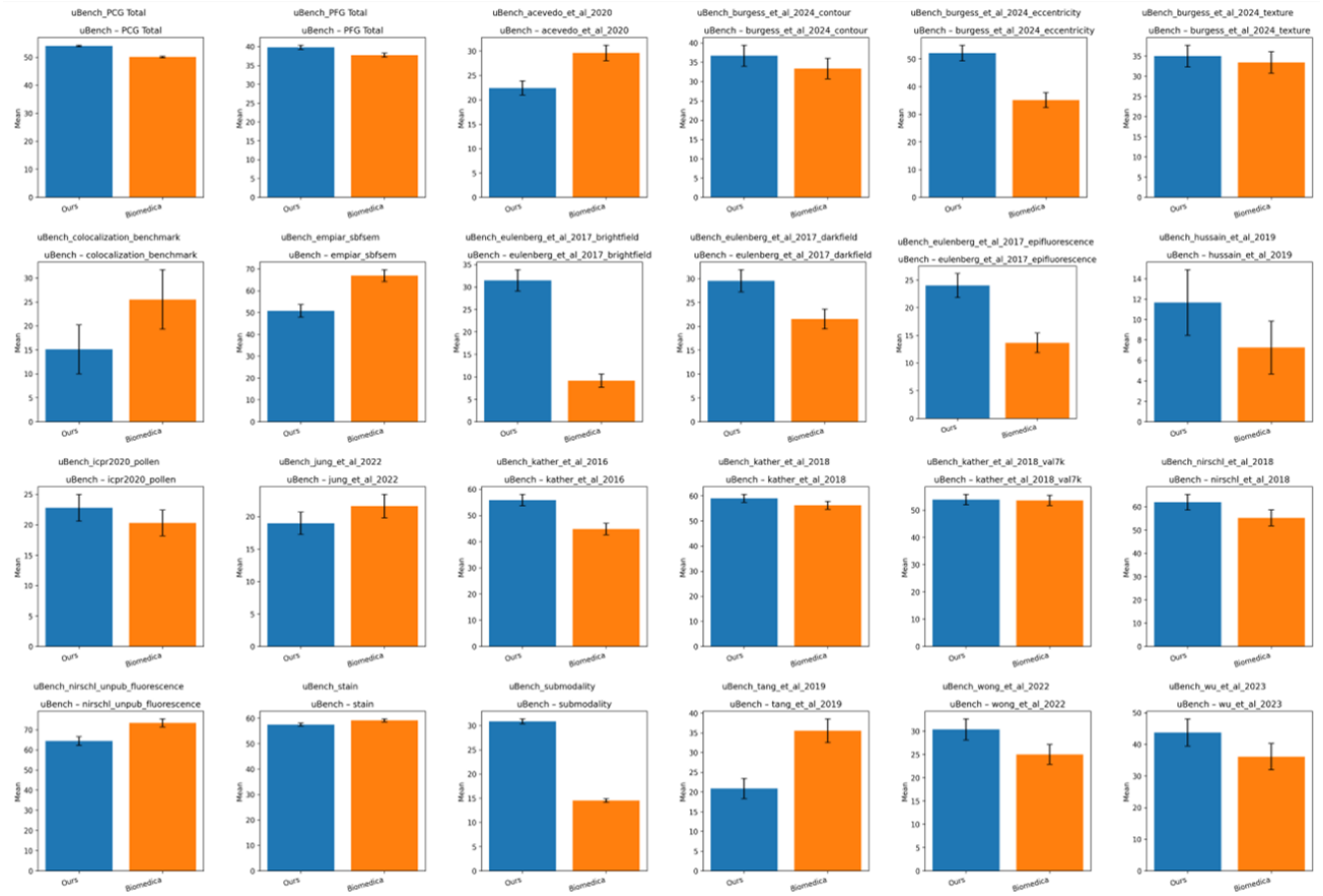

Figure 7. **Qualitative results on cell imaging.** Example retrieval and captioning results on a cell-imaging benchmark. The model trained with Panel2Patch produces more fine-grained and visually grounded descriptions compared to figure-level training, particularly for subtle morphological differences.