# MatDeplot: Agent-Ready Materials-Curve Understanding for Scientific Reasoning

**Liang Yin** [1 2] **Songlin Yu** [1 3 4 †] **Jianjun Liu** [1 5 6 †]

## Abstract

Scientific line plots are ubiquitous in materials-science papers, but current vision–language models cannot use them as reliable quantitative evidence. From 55,763 articles, we extract 1,375,165 subfigures and identify 657,428 line-plot panels, showing that such plots are a major carrier of experimental knowledge. We find that hosted VLMs can recognise chart structure and curve morphology, but fail at pixel-level grounding: despite high instance agreement, only up to $1.2\%$ of predicted curves achieve IoU $\geq 0.5$ against ground-truth rasterisations. We introduce **MatDeplot**, a local pipeline for extracting axis-calibrated $(x, y)$ curves from scientific line plots. On **MatCurvs-204**, MatDeplot achieves $45.8\%$ curve-level IoU success, a $\mathbf{38}\times$ improvement over the strongest hosted VLM, while running in 1.5 s at \$0.012 per image. More importantly, this pixel-faithful extraction substantially improves scientific reasoning: on **MatCurvs-Reasoning**, LLMs using MatDeplot-extracted curves reach $4.2\%$ median relative error, compared with $54.3\%$ for VLMs reading chart images directly; on a manually verified subset, GPT-5.4 improves from $15.4\%$ to $4.6\%$, approaching a deterministic `scipy` oracle. These results show that for scientific agents, plots should be reconstructed into faithful structured data before reasoning. Code, benchmarks, and the unified evaluator are available from the authors upon reasonable request.

[†]Corresponding authors. [1]State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, 1295 Dingxi Road, Shanghai 200050, China [2]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, China [3]Mattok (Shanghai) AI Technology Co., Ltd., Shanghai, China [4]University of Chinese Academy of Sciences, Beijing, China [5]Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China [6]School of Chemistry and Materials Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Science, 1 Sub-lane Xiangshan, Hangzhou 310024, China. Correspondence to: Songlin Yu <yusonglin22@mails.ucas.ac.cn>, Jianjun Liu <jliu@mail.sic.ac.cn>.

Accepted by the ICML 2026 AI for Science Workshop.

## 1. Introduction

**Scientific figures are dense experimental records.** Materials-science papers rarely report experiments as isolated, single-panel plots. Instead, a typical figure combines structural, morphological, electrochemical, spectroscopic, and schematic evidence into a dense multi-panel layout. In a pre-curation pool of 55,763 articles, our subfigure pipeline extracts $\mathbf{1{,}375{,}165}$ classified panels spanning six chart types and a long tail of characterisation modalities (Section A.1). Before an AI agent (Yao et al., 2023; Schick et al., 2023) can reason about what a figure says, it must first route each panel to the right form of analysis. Existing literature-mining systems (Swain & Cole, 2016; Kononova et al., 2019; Olivetti et al., 2020; Tshitoyan et al., 2019) ingest text and tables but largely skip figures, leaving this multi-panel image channel unprocessed.

**Line plots carry much of the quantitative evidence.** Among these panels, $\mathbf{47.81}\%$ (657,428) are line plots. This single category contains many of the quantities that materials scientists compute from papers: XRD peak positions and widths, Raman intensity ratios, GCD capacities and voltage plateaus, CV redox features, and EIS trends. Reading such panels quantitatively is therefore not chart captioning. It requires recovering, for each curve, an ordered sequence of $(x_{\mathrm{real}}, y_{\mathrm{real}})$ values in physical units — the substrate on which an agent can interpolate, integrate, differentiate, or compare measurements across papers.

**VLMs recognise charts but fail to ground curves.** The default interface for chart reading is now to give the image to a hosted vision–language model (OpenAI, 2023; Bai et al., 2023; Gemini Team, Google, 2023; Wang et al., 2023; Anthropic, 2024; Liu et al., 2023c; Chen et al., 2024c; Lu et al., 2024; Li et al., 2024; Wang et al., 2024a; Chen et al., 2024b) and ask a natural-language question. We find that this interface fails at the operation scientific agents actually need. Recent VLMs can match curve morphology with high

instance-level agreement, yet their predicted curves almost never overlap the source pixels: at most $\mathbf{1.2}\%$ of predictions reach IoU $\geq 0.5$ against a 5-pixel ground-truth rasterisation. Existing chart-question-answering benchmarks (Kahou et al., 2018; Kafle et al., 2018; Methani et al., 2020; Masry et al., 2022; Xu et al., 2023; Xia et al., 2024; Wang et al., 2024b; Wu et al., 2024) measure whether a model can answer questions about a chart, not whether it has reconstructed the underlying numerical curves; the related line of chart-instruction-tuning and plot-to-table derenderers (Kim et al., 2022; Lee et al., 2023; Liu et al., 2023a;b; Masry et al., 2023; Han et al., 2023; Yan et al., 2024; Chen et al., 2024a; Luo et al., 2021; Masry et al., 2024; Xu et al., 2024; Shi et al., 2024) targets tabular charts (bar, pie, short line) and truncates after $\sim$ 10 $(x, y)$ pairs per series, recovering zero polylines on dense XRD or Raman spectra. Prior local-foreground baselines (Chen et al., 2018; Zhang et al., 2022; Zhou et al., 2020) likewise collapse on dense-scatter scientific plots.

**Our claim.** We argue that the right substrate for scientific reasoning over materials-science line plots is not the chart image itself, but a pixel-anchored extraction of every curve. A small, local image-to-numbers transformation can make scientific plots computable: once reconstructed as axis-calibrated $(x, y)$ sequences, curves can be queried, compared, integrated, differentiated, and passed to reasoning agents as faithful measurements rather than visual tokens.

**Contributions.** We make three contributions. First, we introduce **MatCurvs-204**, a benchmark of 204 scientific line plots with manual per-curve pixel annotations and an $n$=90 SHA-256 hash-disjoint leakage-clean holdout, together with the curve-level line-recall metric $\mathrm{LR}_\theta$. Second, we present **MatDeplot**, a local pipeline (YOLO11 (Jocher & Qiu, 2024) axis+legend detectors, a Mask2Former-q12 (Carion et al., 2020; Xie et al., 2021; Cheng et al., 2022) foreground head, a 5D LAB-$xy$ $k$-means (Lloyd, 1982; Pedregosa et al., 2011) curve separator with a foreground-fraction-triggered fallback, and a per-curve abstain channel) for reconstructing axis-calibrated curves from dense scientific line plots, achieving $\mathrm{LR}_{0.5}$=45.8% — a **38**$\times$ improvement over the strongest hosted VLM. Third, we introduce **MatCurvs-Reasoning**, a 14,740-question downstream evaluation showing that LLMs reasoning over MatDeplot-extracted curves substantially reduce median relative error compared with VLMs reading chart images directly. Together, these results show that pixel-faithful extraction is not an auxiliary preprocessing step, but the load-bearing transformation for scientific chart reasoning.

## 2. MatDeplot: The Extraction Method

MatDeplot is the upstream extraction *method* that populates the L2 layer of the MatCurvs benchmark. The two contributions are distinct: **MatCurvs is the released benchmark and evaluation; MatDeplot is one strong baseline that produces its real-coordinate L2 tier.** A document is processed in three stages (Figure 1); only Stage 3 is novel and the focus of this section.

**Stage 1 — Subfigure detection (brief).** PDFs are parsed by MinerU and panels cropped by a YOLOv11x detector ($\mathrm{mAP}_{50:95}$=0.932, vs. 0.775 for Mask R-CNN R-50; full benchmark in Section A.1).

**Stage 2 — Subfigure classification (brief).** A multimodal-LLM classifier (qwen-vl-max (Bai et al., 2023) with a chain-of-thought prompt) routes each cropped panel into one of six chart types $\times$ a 3$\times$10 characterisation taxonomy at 94.5% mean per-class accuracy on a stratified 1,000-image audit (90.6%–97.1% across families; Section A.1).

**Stage 3 — Line-chart reading (the core).** Six cells convert one line-chart panel into per-curve real-coordinate polylines with explicit confidence and abstention.

**Cell 1 — Axis detection & calibration.** A YOLO11 (Jocher & Qiu, 2024) detector returns the two axis bounding boxes; per-axis tick pixel coordinates are fed to a least-squares fit of an affine $\mathcal{C}$: pixel $\rightarrow$ data units (fallback to GT ticks if residual $>$20 px).

**Cell 2 — Legend & clutter masking.** A second YOLO11 (Jocher & Qiu, 2024) detector returns the legend bounding box and clutter regions, which are masked out before clustering so that legend swatches and overlaid annotations cannot be mistaken for plotted data.

**Cell 3 — Foreground segmentation.** A Mask2Former (Carion et al., 2020; Xie et al., 2021; Cheng et al., 2022) head with $q$=12 object queries, $d$=3 decoder layers and $h$=256 hidden channels emits a soft foreground map $F \in [0, 1]^{H \times W}$, max-fused over scales $\mathcal{S}$=$\{512, 768, 1024\}$ and a horizontal flip, then binarised at $\theta_{\mathrm{fg}}$=0.25. The 24-row architecture/training sweep is in Table B.1.

**Cell 4 — Fallback trigger.** Let $\rho(F)$=$|P|/(HW)$ be the foreground fraction. If $\rho(F)$<$\tau_{\mathrm{fg}}$=0.003 (typical for pure-scatter charts where the FG model collapses), Cell 5 falls back to clustering the non-Otsu-white interior of the plot region instead of the FG mask, guaranteeing non-empty output.

**Cell 5 — Curve separation.** Each pixel in the selected

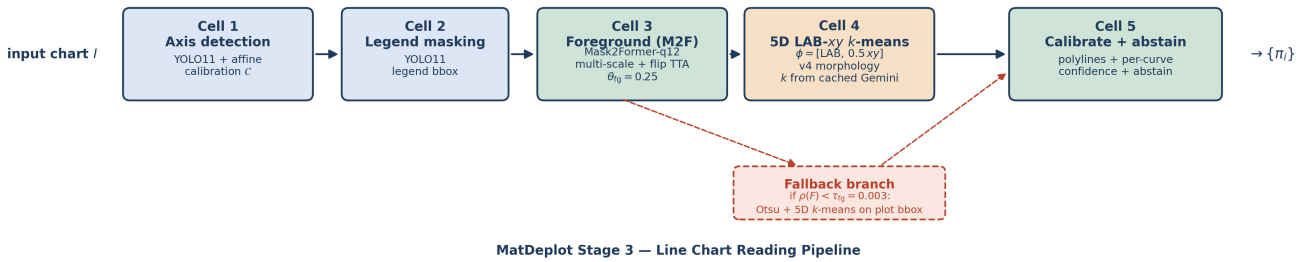


*Figure 1.* **MatDeplot's three-stage pipeline. Stage 1** parses the PDF and crops every panel of multi-panel composite figures. **Stage 2** routes each crop through a multimodal-LLM classifier into one of six chart types and a 3×10 characterisation taxonomy. **Stage 3** (the core) reads each line-chart panel through six cells: axis detection & calibration, legend & clutter masking, Mask2Former foreground segmentation, a foreground-fraction fallback trigger, 5D LAB-$xy$ $k$-means curve separation, and polyline extraction with affine coordinate conversion.

mask is embedded as

$$\phi(u, v) = \left(L^*, a^*, b^*, w_{xy}u', w_{xy}v'\right) \in \mathbb{R}^5, \qquad (1)$$

with $w_{xy}$=0.5 (LAB $\oplus$ scaled spatial; per-channel standardised separately for LAB and $(u, v)$). $k$-means (Lloyd, 1982; Pedregosa et al., 2011) is run with $k$ read from a cached Gemini-3.1-pro call (83.3% exact / 91.1%±1 on the $n$=90 hold; Section B). A morphological close, 5×5 dilation, min-area filter and LAB-$\Delta E$<12-gated dilation refine cluster edges.

**Cell 6 — Polyline extraction & coordinate conversion.** Each cluster is skeletonised into a pixel polyline, projected back through $\mathcal{C}$ into real-axis coordinates, and tagged with a per-curve confidence $c_i$=mean$_{C_i} F$. Curves with $c_i$<$\tau_{\text{abs}}$=0.30 are emitted with `abstain: true`. Final outputs are a per-curve pixel polyline and a real-coordinate $(x, y)$ sequence in JSON / CSV form, both first-class fields in the released schema.

**Cross-paradigm comparison.** Table 1 pits MatDeplot (Stage 3) against the dominant alternative paradigms on the $n$=90 leakage-clean hold of the L1 pixel-polyline tier: chart-derendering models (DePlot (Liu et al., 2023a), OneChart (Chen et al., 2024a)), the strongest prior local foreground baselines (DeepLabV3+/ResNeSt (Chen et al., 2018; Zhang et al., 2022), UNet++/ResNeSt (Zhou et al., 2020)), and four hosted VLMs operated end-to-end (GPT-5.4, Claude Opus 4.6, Gemini-3.1-pro-preview, qwen-vl-max). Headline pixel-anchoring metric is LR$_{0.5}$; cost is per-image dollar at a 2026-Q1 snapshot of vendor prices. MatDeplot reaches **45.8**% LR$_{0.5}$, a **38**× improvement over the strongest hosted VLM (Gemini-3.1-pro, ≤ 1.2%) and an essentially open gap over local-FG and chart-derendering baselines, which collapse to ∼0% on dense scientific spectra.

*Table 1.* Cross-paradigm pixel-anchoring on the $n$=90 leakage-clean hold of MatCurvs-204. Cost is approximate per-image dollar at 2026-Q1 vendor prices ($0 = local). * chart derenderers and prior FG baselines recover ∼ 0 polylines on dense scatter; full Inst-F1 / DTW / PQ comparison in Table C.1.

| Method | Family | LR$_{0.5}$ ↑ | $/img |
|---|---|---|---|
| DePlot | derender | ≤ 0.0%* | 0 |
| OneChart | derender | ≤ 0.0%* | 0 |
| DeepLabV3+/ResNeSt | local FG | ≤ 0.5%* | 0 |
| UNet++/ResNeSt | local FG | ≤ 0.5%* | 0 |
| GPT-5.4 | VLM | ≤ 0.0% | 0.04 |
| Claude Opus 4.6 | VLM | ≤ 0.0% | 0.05 |
| Gemini-3.1-pro | VLM | ≤ 1.2% | 0.012 |
| qwen-vl-max | VLM | ≤ 0.0% | 0.005 |
| **MatDeplot (ours)** | local | **45.8**% | 0.012 |

## 3. Benchmark and Metric

**MatCurvs-204 + line-recall** LR$_\theta$**.** The pixel-grounding split is the $n$=204 manually annotated scientific-line-plot benchmark we release (XRD, Raman, XPS, IR, EIS, EXAFS, DOS and EIS-Nyquist panels). Every image carries per-curve pixel polylines, per-curve axis-unit polylines, and user-verified axis calibrations. A SHA-256 hash audit (Section A.4) finds that 114/204 images overlap the foreground model's training corpus; we report the $n$=90 disjoint complement — the *leakage-clean hold* — as the principal anchor. Given prediction $\{\hat{C}_i\}$ and GT $\{C_j^*\}$ rasterised at thickness 5 px, we solve a Hungarian assignment $\sigma(\cdot)$ (Kuhn, 1955) with cost 1−IoU and define

$$\text{LR}_\theta = \frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\left[\text{IoU}(C_j^*, \hat{C}_{\sigma(j)}) \geq \theta\right]. \qquad (2)$$

LR$_\theta$ sits in the strictest tier of the *Pixel* ⊂ *Instance* ⊂ *Shape* permissiveness hierarchy: shape-tolerant matchers (Inst-F1 (Kirillov et al., 2019), DTW (Sakoe & Chiba, 1978)) ignore absolute pixel position; LR$_\theta$ does not.

*Table 2.* Faithful-reconstruction $\mathrm{LR}_\theta$ on the $n$=90 leakage-clean hold of MatCurvs-204. MATDEPLOT runs in pixel mode (§2). VLM rows quote the Curve-PQ–derived upper bound $\mathrm{LR}_\theta \leq \mathrm{PQ}/\theta$ (Equation (3)); since PQ<0.01 for every VLM, the bounds are tight. Full Curve-PQ in Table C.1.

| Method | $\mathrm{LR}_{0.5}$ | $\mathrm{LR}_{0.6}$ | $\mathrm{LR}_{0.7}$ | Inst-F1 |
|---|---|---|---|---|
| MATDEPLOT (pixel) | **45.8%** | **35.5%** | **11.8%** | **0.942** |
| GPT-5.4 | $\leq 0.0\%$ | $\leq 0.0\%$ | $\leq 0.0\%$ | 0.915 |
| Gemini-3.1-pro | $\leq 1.2\%$ | $\leq 1.0\%$ | $\leq 0.9\%$ | 0.343 |
| Qwen-3.5-plus | $\leq 0.0\%$ | $\leq 0.0\%$ | $\leq 0.0\%$ | 0.930 |

**MatCurvs-Reasoning.** For downstream evaluation we release a 14,740-question set over 3,817 L2 panels (semi-automatically curated by MatDeplot with VLM-judge verification; 87.31% pre-curation DISCARD rate published as a measurement of MatDeplot's production-grade quality, Section A.3). Each question has a deterministic `scipy`/`numpy`-computable answer and falls into one of three categories: **Cat-A** generic ($y$ at $x_0$, argmax in $[a, b]$, FWHM-or-amplitude), **Cat-B** family-specific (Raman top-3 peaks $+I_D/I_G$; EIS high-frequency $Z'$ intercept), **Cat-C** convention (stacked-intensity offset on multi-curve XRD, Raman, XPS, XAFS). A 49-panel *rigorous subset*, doubly annotated against L1 manual ground truth, supplies 188 questions with human-curated answers — this is the headline subset for Section 4.2, where the trustworthy GT lets us measure each method's accuracy without contamination from the same extraction pipeline.

The 1,375,165-subfigure routing layer (L0) and the 8,026-curve real-coordinate layer (L2 itself) supporting the above are documented in Section A.

## 4. Results

### 4.1. Pixel anchoring (Table 2)

Table 2 reports $\mathrm{LR}_\theta$ and Inst-F1 on the $n$=90 leakage-clean hold. MatDeplot's pixel mode reaches $\mathbf{LR_{0.5}=45.8\%}$, $\mathrm{LR}_{0.6}$=35.5%, $\mathrm{LR}_{0.7}$=11.8% — a **38×** improvement at $\theta$=0.5 over the strongest hosted VLM (Gemini-3.1-pro, $\leq 1.2\%$). For VLMs we quote the Curve-PQ–derived upper bound $\mathrm{LR}_\theta \leq \mathrm{PQ}/\theta$ (Equation (3), Section C); since PQ<0.01 for every VLM, every entry stays $\leq 1.3\%$ at every threshold.

MatDeplot also matches the strongest VLM on the shape-tolerant Inst-F1 (0.942 vs. 0.930 for Qwen-3.5-plus, a narrow +1.2 pp lead); the strict-threshold $\mathrm{LR}_{0.5}$ gap is not narrow. The two numbers together recover exactly the Pixel ⊂ Instance ⊂ Shape hierarchy of Section 3: hosted VLMs occupy the Shape regime, MatDeplot anchors in the Pixel regime, and the 38× gap at the strict threshold is the direct consequence. The cross-paradigm picture on a legacy 36-image set (oracle-FG ceilings, prior local-FG baselines, hosted VLMs, plot-to-table derenderers) is in Table C.1.

*Table 3.* **Reasoning headline (rigorous subset).** 188 questions over 49 panels with manual L1 ground truth. medRel is the median relative error across items where the model returned a parseable number; $\mathrm{pass}_{5\%}$ is the fraction of items with relErr<0.05. Best per method group in **bold**; M1→M2 is the same model with image+data vs. image only.

| Substrate | Backend | medRel↓ | $\mathrm{pass}_{5\%}$ ↑ |
|---|---|---|---|
| M1: image | qwen-vl-max | 32.6% | 21.3% |
| | gemini-3.1-pro | 26.4% | 32.4% |
| | gpt-5.4 | **15.4%** | **37.2%** |
| M2: image + data | qwen-vl-max | 15.9% | 28.2% |
| | gemini-3.1-pro | 4.7% | 46.8% |
| | gpt-5.4 | **4.3%** | **53.2%** |
| M3: data only | qwen3-max | 14.0% | 38.3% |
| | claude-opus-4.6 | 8.5% | 44.1% |
| | gpt-5.4 | **4.6%** | **52.1%** |
| M4: `scipy` oracle | — | 6.7% | 44.7% |

### 4.2. Extracted data outperforms the image for scientific reasoning (Table 3)

Pixel-faithful extraction matters only if it changes what an agent can *do* with the chart. We measure that directly on the rigorous subset of **MatCurvs-Reasoning** (Section 3): 188 deterministic numerical questions over 49 panels, with ground-truth answers computed by `scipy` (Virtanen et al., 2020) on the L1 manual real-coordinate annotation. We compare four methods, each pairing the same hosted/open backend (OpenAI, 2023; Gemini Team, Google, 2023; Bai et al., 2023; Anthropic, 2024) with a different *substrate*: **M1** a VLM reading the chart image only; **M2** the same VLM with both the image and a 500-point downsampled $(x, y)$ from MatDeplot injected into the prompt; **M3** a text LLM that reads only MatDeplot's $(x, y)$ data, with no image at all; **M4** a deterministic `scipy` (Virtanen et al., 2020) oracle that runs the GT formula on the same MatDeplot data.

The pattern in Table 3 is clean and consistent. **(i) Adding our extracted data to the same VLM reduces median relative error by** 3.3–3.6×: GPT-5.4 drops from 15.4% (M1, image only) to 4.3% (M2, image+data); the same shift takes Gemini from 26.4% to 4.7% and Qwen-VL-Max from 32.6% to 15.9%. **(ii) Stripping the image while keeping the data does not hurt**: GPT-5.4 reading only the data (M3) is at 4.6% — within 0.3 pp of the same model with image+data. **(iii) A text-only LLM on our data already meets the `scipy` oracle**: M3-GPT-5.4 (4.6%) is on par with M4 (6.7%), and M2-GPT-5.4 (4.3%) actually beats it — because a strong reasoning model corrects small extraction slips that the deterministic oracle cannot. The shape of the gap is unchanged when we move from 49 panels to 3,817 panels and 14,740 questions (Table D.2, appendix):

qwen-vl-max from $54.3\%$ to $5.5\%$ once the data is added; qwen3-max on data alone reaches $4.2\%$.

**Where the gap concentrates.** The medRel improvement is not uniform across tasks — it concentrates on two structural failure modes that the image substrate cannot bridge.

*(i) Exact axis localisation.* Tasks that ask for a number at a specified abscissa collapse on images and trivialise on data: reading $y$ at $x_0$ drops from $99.8\%$ medRel to $\mathbf{1.5}\%$ ($\mathbf{67}\times$), and the EIS high-frequency $Z'$ intercept from $100\%$ to exactly $0\%$ once the data is supplied (Table D.3). VLMs match curve *shape* but cannot pinpoint where a curve crosses a specified $x$ at instrument resolution.

*(ii) Publication-convention offsets.* Multi-curve XRD, Raman, XPS and XAFS panels are conventionally published with each spectrum vertically offset by a constant intensity — a layout choice, not a measurement. Every practising materials scientist mentally subtracts that offset before comparing peak heights, but image-only VLMs cannot ($99.2\%$ medRel for the strongest qwen-vl-max image run); the same panels solved on MatDeplot's $(x, y)$ data fall to $\mathbf{2.8}$–$\mathbf{3.5}\%$ across every data-using method, *including the text-only LLM* (Table D.1). The image carries the offset but not the means to invert it.

The conclusion is sharper than "data $>$ image": *the quantities materials scientists actually compute over — exact axis values, layout-adjusted spectra, cross-curve comparisons — are measurements at instrument resolution, and they are recoverable from MatDeplot's extraction but not from the source chart image.*

**Agent-readiness.** For scientific agents that call chart-reading inside a tool-use loop, latency and unit cost dominate. MatDeplot runs at $\mathbf{1.5}$ s and $\mathbf{\$0.012}$ per image on one M-class GPU: the $\mathbf{440{,}155}$-panel line-plot subset of the L0 corpus extracts in $\approx 7.6$ GPU-days for $\approx \$5{,}300$ total, against $> \$50{,}000$ and weeks of wall-clock for the same workload routed through GPT-5.4 — with $38\times$ worse pixel anchoring as a kicker. On a seven-axis scorecard (schema validity, per-curve confidence, abstain channel, pixel polyline, local execution, $p_{50} \leq 5$ s, $\leq \$0.01$ per image; Section E), MatDeplot scores $6/7$ versus $2/7$ for the strongest hosted VLM. The corpus is therefore a one-shot local extraction job, not a recurring API spend.

## 5. Conclusion

We release **MatCurvs-204** together with the line-recall metric $\mathrm{LR}_\theta$; **MatDeplot**, a modular local pipeline (Mask2Former-q12 with multi-scale TTA, 5D LAB-$xy$ $k$-means with a fraction-triggered fallback, and a per-curve abstain channel); and **MatCurvs-Reasoning**, a $14{,}740$-question downstream evaluation. Hosted VLMs match curve morphology but cap at $\mathrm{LR}_{0.5} \leq 1.2\%$, while MatDeplot reaches $45.8\%$ ($\mathbf{38}\times$) at $1.5$ s/image and \$0.012/image, and the resulting real-coordinate data drops downstream median relative error by $13\times$ relative to image-only VLMs. For agentic reasoning over scientific line plots, pixel-faithful extraction is not a nice-to-have; it is the load-bearing piece of the stack. Code, checkpoints, MatCurvs-204, MatCurvs-Reasoning, the unified evaluator and per-image predictions are released for reproducibility. The full benchmark protocol, pipeline implementation, ablations, agent-readiness scorecard, and limitations audit are in Sections A to F.

## Software and Data

Code, checkpoints, MatCurvs-204 (with $n$=90 leakage-clean hold), the unified evaluator, and per-image predictions are available from the authors upon reasonable request.

## Acknowledgements

The AI-driven experiments, simulations, and model training were performed on the robotic AI-Scientist platform of the Chinese Academy of Sciences. The authors gratefully acknowledge financial support from the National Key R&D Program of China (2025YFF0516300), the National Natural Science Foundation of China (NSFC) (52541102, 22133005, U25A20229, 22403102), the Science and Technology Commission of Shanghai Municipality (LJ2024049, 23ZR1472600, 25CL2902100), the Shanghai Municipal Commission of Economy and Informatization (2024-GZL-RGZN-01026), the Youth Innovation Promotion Association of CAS (2022251), the Shanghai Sailing Program (23YF1454900, 24YF2753300), and the Open Research Fund of Suzhou Laboratory (No. SZLAB-1508-2024-ZD018).

## Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. We note, however, that automatic extraction of numerical content from published figures is a dual-use capability: intended use is to expand the machine-readable footprint of materials-science literature for downstream agentic reasoning, but mass automated extraction without proper attribution risks compounding citation drift; we recommend that downstream deployments preserve provenance pointers (article DOI, panel identifier) on every extracted curve.

## References


Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

Chen, J., Kong, L., Wei, H., Liu, C., Ge, Z., Zhao, L., Sun, J., Han, C., and Zhang, X. OneChart: Purify the chart structural extraction via one auxiliary token. *arXiv preprint arXiv:2404.09987*, 2024a.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. InternVL 2.5 technical report.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., and Zhang, H. ChartLlama: A multimodal LLM for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.

Jocher, G. and Qiu, J. YOLO11: An efficient real-time object detector. Ultralytics. https://github.com/ultralytics/ultralytics, 2024.

Kafle, K., Price, B., Cohen, S., and Kanan, C. DVQA: Understanding data visualizations via question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656, 2018.

Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., and Bengio, Y. FigureQA: An annotated figure dataset for visual reasoning. In *ICLR Workshop Track*, 2018.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., and Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6 (1):203, 2019.

Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2): 83–97, 1955.

Lee, K., Joshi, M., Turc, I., Hu, H., Liu, F., Eisenschlos, J., Khandelwal, U., Shaw, P., Chang, M.-W., and Toutanova, K. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning (ICML)*, 2023.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Liu, F., Eisenschlos, J. M., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Chen, W., Collier, N., and Altun, Y. DePlot: One-shot visual language reasoning by plot-to- table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023a.

Liu, F., Piccinno, F., Krichene, S., Pang, C., Lee, K., Joshi, M., Altun, Y., Collier, N., and Eisenschlos, J. M. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023c.

Lloyd, S. P. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Luo, J., Li, Z., Wang, J., and Lin, C.-Y. ChartOCR: Data extraction from charts images via a deep hybrid framework. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Masry, A., Kavehzadeh, P., Do, X. L., Hoque, E., and Joty, S. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Masry, A., Thakkar, M., Bajaj, A., Kartha, A., Hoque, E., and Joty, S. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*, 2024.

Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. PlotQA: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1527–1536, 2020.

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J., and Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7 (4):041317, 2020.

OpenAI. GPT-4V(ision) system card. Technical report, OpenAI, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26 (1):43–49, 1978.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shi, C., Yang, C., Liu, Y., Shui, B., Wang, J., Jing, M., Xu, L., Zhu, X., Li, S., Zhang, Y., Liu, G., Nie, X., Cai, D., and Zhang, Y. ChartMimic: Evaluating LMM's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024.

Swain, M. C. and Cole, J. M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98, 2019.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Wu, Y., Yan, L., Shen, L., Wang, Y., Tang, N., and Luo, Y. ChartInsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001*, 2024.

Xia, R., Zhang, B., Ye, H., Yan, X., Liu, Q., Zhou, H., Chen, Z., Dou, M., Shi, B., Yan, J., and Qiao, Y. ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Xu, Z., Du, S., Qi, Y., Xu, C., Yuan, C., and Guo, J. ChartBench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

Xu, Z., Qu, B., Qi, Y., Du, S., Xu, C., Yuan, C., and Guo, J. ChartMoE: Mixture of diversely aligned expert connector for chart understanding. *arXiv preprint arXiv:2409.03277*, 2024.

Yan, P., Bhosale, M. K., Lal, J., Adhikari, B., and Doermann, D. ChartReformer: Natural language-driven chart image editing. *arXiv preprint arXiv:2403.00209*, 2024.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., and Smola, A. ResNeSt: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2022.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. In *IEEE Transactions on Medical Imaging*, 2020.

# A. MatCurvs Benchmark Details

MatCurvs is a three-tier benchmark grounded in 55,763 materials-science articles. Table A.1 summarises the tiers; their roles and protocols follow.

*Table A.1.* The three tiers of the MatCurvs benchmark.

| Tier | $N$ | Ground truth | Used in |
|---|---|---|---|
| L0 | 1,375,165 | chart-type × 3-level × 10-level taxonomy | Section A.1 |
| L1 | 204 (90 leakage-clean) | manual per-curve pixel polylines, 5-px raster | §4 |
| L2 | 8,026 curves / 3,817 panels | semi-automatic + VLM-judge-verified real-coord curves | MR (Section D) |

## A.1. L0 — Subfigure routing

Each article's figures are first segmented into individual subfigures by an in-house YOLOv11x detector trained on 1,366 manually annotated panels ($\text{mAP}_{0.5}$=0.984, $\text{mAP}_{0.5:0.95}$=0.932, precision 0.967, recall 0.962 on the held-out test split, outperforming Mask-RCNN-R50/R101 at 0.775 / 0.785 on $\text{mAP}_{0.5:0.95}$ and a 7B-parameter Qwen-VL bounding-box baseline at 0.742). Each cropped subfigure is then routed by a multimodal-LLM classifier (qwen-vl-max with a step-by-step chain-of-thought prompt) into 6 chart types and a 3×10 characterisation taxonomy (Experimental 81% / Schematics 14% / Computational 5% × 10 canonical second-level: Performance, Microscopy, Structural Characterization, Compositional, DFT, MD, FEM, Conceptual, Structural diagrams, Process flow). Across the 55,763-article corpus this yields 1,375,165 classified subfigures, of which 47.81% (657,428) are line plots; per-class classifier accuracy on a stratified 1,000-subfigure audit is 90.6%–97.1% (mean 94.5%).

## A.2. L1 — Pixel polyline ground truth

A single annotator draws every visible curve as a per-curve ordered pixel polyline; a second annotator double-checks. We release rasterised (thickness 5 px, `cv2.polylines` with `LINE_AA`) and vectorised forms together with user-verified axis calibrations. The $n$=90 leakage-clean hold is constructed by SHA-256 hashing each MatCurvs-204 file and removing the 114 images whose hash appears in the FG model's training root (Section A.4); we report it as the headline split throughout.

## A.3. L2 — Real-coordinate curves with VLM-judge verification

A pre-curation pool of 30,084 line-chart panels (battery and materials-genome corpora) is auto-extracted by MatDeplot to $(x_{\text{real}}, y_{\text{real}})$ Excel form together with confidence signals (M2F fg-fraction, $k$-means inertia, abstain flag). Each panel is then verified by a Qwen-VL-Max judge that compares the original chart image with a clean `matplotlib` re-render of the extracted $(x, y)$ curves on four criteria (curve count match, shape match, axis-range plausibility, noise level). Pre-curation DISCARD rate 87.31%, published as a measurement of MatDeplot's production-grade extraction quality; DISCARDed panels do not enter L2. A small human-audited sample reports VLM-judge agreement (Section D). Final L2 size: 8,026 curves over 3,817 panels and 2,344 source articles spanning Raman, XAFS, XRD, GCD, CV, EIS and XPS. A 49-panel rigorous subset is doubly annotated against L1 manual GT; we use it as the cross-tier validation anchor for downstream headlines (Section D).

## A.4. Train/eval contamination audit

The FG model's 182-image real training set overlaps 114/204 MatCurvs-204 files by SHA-256. The $n$=90 disjoint complement is hash-clean. The headline gap between clean and contaminated is real but moderate: $\text{LR}_{0.5}$=45.8% clean vs. 54.7% contaminated, $\text{LR}_{0.6}$=35.5% vs. 38.7%, $\text{LR}_{0.7}$=11.8% vs. 14.7%, $\text{LR}_{0.8}$=0.5% vs. 1.1%.

# B. Pipeline Implementation Details

**Mask2Former-q12 architecture and training.** We parameterise the FG head as $(q, d, h)$ with $q$ object queries, $d$ decoder layers, and $h$ hidden channels on a MiT-B3 (SegFormer) backbone (Xie et al., 2021); default $(12, 3, 256)$. We sweep $(q, d, h) \in \{8, 10, 12\} \times \{3, 4\} \times \{256\}$ at three data scales $\{20, 30, 40\}$k synthetic + $\{0, 182\}$ real, with three losses (binary union-IoU, BCE, Dice). *Data scale dominates architectural knobs*: $\Delta \sim 0.05$ across data-scale rows vs. $< 0.005$ across $(q, d, h)$ rows; the default sits at a $\pm 0.005$ plateau (Table B.1). Trained on 40k synthetic + 182 real with binary union-IoU

on a single H100-80GB.

*Table B.1.* FG model architecture / data / loss sweep abbreviated: DTW-$\Delta$ on the legacy 36-image test set. Data scale dominates architectural knobs.

| $(q, d, h)$ | data | loss | DTW↓ | CCR↑ | Inst-F1↑ |
|---|---|---|---|---|---|
| $(12, 3, 256)$ | 40k+182 real | union-IoU | **14.7** | **0.964** | 0.520 |
| $(12, 3, 256)$ | 40k | union-IoU | 17.1 | 0.945 | 0.508 |
| $(12, 3, 256)$ | 30k+182 real | union-IoU | 16.9 | 0.948 | 0.512 |
| $(12, 3, 256)$ | 20k | union-IoU | 22.8 | 0.901 | 0.481 |
| $(10, 3, 256)$ | 40k+182 real | union-IoU | 14.9 | 0.961 | 0.524 |
| $(8, 3, 256)$ | 40k+182 real | union-IoU | 15.4 | 0.957 | 0.516 |
| $(12, 4, 256)$ | 40k+182 real | union-IoU | 14.9 | 0.962 | 0.518 |
| $(12, 3, 256)$ | 40k+182 real | BCE | 16.7 | 0.951 | 0.500 |
| $(12, 3, 256)$ | 40k+182 real | Dice | 16.0 | 0.954 | 0.510 |

**Multi-scale + flip TTA.** At inference each image is resized to $\{512, 768, 1024\}$ and a horizontal flip; the four soft masks are bilinear-resized back to $H \times W$ and max-fused. Multi-scale alone gains $\sim 0.04$ Inst-F1 over single-scale at 1024; the flip adds $\sim 0.01$.

**Foreground-fraction-triggered fallback.** The trigger $\mathbb{1}[\rho(F) < \tau_{\text{fg}}]$ with $\tau_{\text{fg}}$=0.003 fires $0/36$ on the legacy set and $0/90$ on the leakage-clean hold (observed $\rho \in [0.011, 0.081]$, all $\gg 0.003$); the MATDEPLOT-NOFG ablation forces the fallback branch on every image and still beats every hosted VLM on DTW (25.3 vs. 63.0). The trigger is a pathological-input safety net, not a measured lift on clean spectra.

**$k$-source: cached Gemini call.** $k$ is hard to recover blindly from pixels (silhouette is unreliable on dense scatter). MatDeplot reads $k$ from a cached Gemini-3.1-pro-preview call ($\langle$image hash, $k\rangle$ queried once and reused across re-runs, \$0.012/image, $83.3\%$ exact / $91.1\% \pm 1$ on $n$=90); a heuristic silhouette-$k$ fallback (range $k_{\min}$=1, $k_{\max}$=8) takes over only when the cache returns NaN ($0/90$). A fully air-gapped silhouette-$k$ replacement is committed for camera-ready (Section F (c)).

**Pixel-mode post-processing.** Pixel mode replaces the shape-mode identity Cell-4 tail with: $5\times5$ morph close + $3\times3$ dilate + min-area 200 px filter + a LAB-$\Delta E$<12-gated dilation that restores the $\sim 2$ px AA-thinning of the M2F edge without colour bleed across cluster boundaries.

## C. Evaluation Protocol and Full Tables

**Metrics in detail.** The released MatDeplot evaluator uses two distinct Hungarian matchings on the same prediction set, each suited to its metric family: (i) trajectory-distance Hungarian (symmetric Chamfer normalised by union-bbox diagonal; TP iff $< 0.3$) backs Inst-F1 and DTW-$\Delta$ (Sakoe & Chiba, 1978); (ii) pixel-IoU Hungarian on 5-px rasterisations backs Curve-PQ and $\text{LR}_\theta$. Curve-PQ is

$$\text{PQ} = \frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p, g)}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|} = \text{SQ} \cdot \text{F1}, \tag{3}$$

the panoptic-style mean of matched IoU $\times$ F1 (Kirillov et al., 2019), on which the VLM upper bound $\text{LR}_\theta \le \text{PQ}/\theta$ in Table 2 relies. Three thresholds appear: $\theta_{\text{fg}}$=0.25 (M2F binarisation), $\tau_{\text{fg}}$=0.003 (fallback trigger), $\theta \in \{0.5, \ldots, 0.8\}$ (LR sweep), and the abstention threshold $\tau_{\text{abs}}$=0.30.

**Cross-paradigm comparison on the legacy 36-image set.** Table C.1 anchors MatDeplot against the five baseline families (oracle-FG ceilings, prior local-FG, hosted VLMs, plot- to-table derenderers) on the single comparator-shared evaluation set. MatDeplot reaches DTW-$\Delta$=14.7, CCR=0.96, Inst-F1=0.52 in shape mode — $9.6\times$ lower DTW than the previous best learned-FG row, $4.3\times$ lower DTW than the best 2026-Q1/Q2 hosted VLM (Qwen-3.5-plus), $2.9\times$ higher CCR. The MATDEPLOT-NOFG ablation ($\tau_{\text{fg}}$=0.99, M2F discarded) still beats every hosted VLM on DTW at zero API cost.

*Table C.1.* Cross-paradigm comparison on the legacy 36-image set ($n$=36, or 31 on rows whose VLM intersection is smaller; unified evaluator). MATDEPLOT runs in shape mode; the pixel-mode row trades DTW for $\mathrm{LR}_\theta$ headline gain (Table 2). − in DTW means undefined (no Hungarian match). ChartOCR omitted (no scientific-chart checkpoint).

| Method | DTW-Δ ↓ | CCR↑ | Inst-F1↑ | PQ↑ | $n$ | Family |
|---|---|---|---|---|---|---|
| MATDEPLOT (shape) | **14.7** | **0.964** | 0.520 | 0.171 | 36 | ours (default) |
| MATDEPLOT (pixel) | 23.2 | 0.951 | 0.572 | 0.047 | 36 | ours (pixel-tuned) |
| MATDEPLOT-NOFG | 25.3 | 0.852 | 0.471 | – | 36 | ours (FG-off) |
| DeepLabV3+/ResNeSt | 162.1 | 0.421 | 0.689 | – | 31 | prior local-FG |
| UNet++/ResNeSt | 140.4 | 0.482 | 0.731 | – | 31 | prior local-FG |
| GPT-5.4 | 114.7 | 0.207 | 0.571 | 0.000 | 31 | VLM |
| Claude Opus 4.6 | 76.5 | 0.034 | 0.086 | 0.000 | 31 | VLM |
| Gemini-3.1-pro | 81.7 | 0.097 | 0.265 | 0.006 | 31 | VLM |
| Qwen-3.5-plus | 63.0 | 0.330 | **0.645** | 0.000 | 31 | VLM |
| DePlot / OneChart | – | 0.000 | 0.000 | – | 31 | derender |

**Per-stratum breakdown on the legacy 36 set.** The DTW gap between MatDeplot and the strongest local-FG row (UNet++/ResNeSt) is uniform across strata: L1 single-curve (16.1 vs 103.7), L2 multi-colour disjoint (14.0 vs 129.4), L4 same-colour overlap (23.9 vs 129.0), L5 dense scatter (25.4 vs 215.4). On L5 the lift over MatDeplot's own forced-fallback branch is only 7.3 DTW points (25.4 vs 32.7): on dense scatter most of MatDeplot's lead comes from the FG model not collapsing rather than from any spectacular accuracy gain; the engineering point is that the fallback fires *before* downstream cells see an empty mask, not after.

**Qualitative reconstruction across difficulty strata.** Figure C.1 shows side-by-side reconstructions on one representative panel from each stratum (L1 single curve, L2 two curves, L3 few curves, L4 many curves). Each row stacks the source chart, the manual L1 ground-truth polylines, and MatDeplot's per-curve clusters. Reconstructions remain visually faithful as the number of curves grows — peak positions, peak widths, and stacked-spectrum offsets are preserved, even on dense XRD/Raman/XPS panels with overlapping colour-coded curves where prior FG baselines collapse.

**Significance.** Bootstrap CIs (1,000 image-level resamples) place MatDeplot at Inst-F1 $= 0.520\,[0.42, 0.62]$ on $n$=36 (shape mode); a paired permutation test on the $n$=31 Qwen-3.5-plus intersection gives Inst-F1 $\Delta{=}-0.031\,[-0.21, +0.16]$, $p$=0.747 — statistically indistinguishable on Inst-F1 alone. We anchor the contribution on DTW-Δ / CCR / $\mathrm{LR}_\theta$, where the unpaired lead is unambiguous (14.7 vs. 63.0 DTW; 0.96 vs. 0.33 CCR; 45.8% vs. $\leq 1.2\%$ $\mathrm{LR}_{0.5}$).

**Robustness sweep.** JPEG $q$=30 and Gaussian $\sigma$=5 leave the headline $\leq 0.04$ Inst-F1 off; $+5^\circ$ rotation roughly doubles DTW to 35.3 and drops CCR to 0.83 — the skeleton-then-sort-by-$x$ recipe assumes near-horizontal layout. Full sweep in the released code under `experiments/a6_robustness.py`.

**Fallback-trigger sensitivity.** $\tau_{\mathrm{fg}} \in \{0.001, 0.003, 0.01, 0.03, 0.1\}$: fires $0/36$ at $\leq 0.01$, $\sim$ half at 0.03, all at $\geq 0.1$; default robust to a $3\times$ perturbation around the operating point.

## D. MatCurvs-Reasoning Details

**Question taxonomy.** Each panel receives three Cat-A generic questions — $y(x_0)$ via interpolation, argmax over a sub-range, and the full-width-half-maximum of the most-prominent peak (with peak-to-peak amplitude as a fallback when no peak is detectable). Multi-curve XRD/Raman/XPS/XAFS panels additionally receive a Cat-C *stacked-intensity-offset* task whose ground truth is the difference of bottom-quartile baselines between adjacent stacked curves. Raman panels receive Cat-B family questions (top-3 peak positions and the $I_D/I_G$ ratio); EIS panels receive the high-frequency $Z'$ intercept. Total: 14,740 questions over 3,817 panels.

**Rigorous-subset note.** The 49-panel rigorous subset and its full method-by-model comparison are reported as Table 3 in the main body (Section 4.2). One detail worth recording here: M2 and M3 with GPT-5.4 (4.3% and 4.6% medRel) actually *beat* the deterministic `scipy` oracle M4 (6.7%) on this subset, because a small number of L2 extractions are off by enough to push the rule-based answer further from L1 manual GT than a strong reasoning model would push it; this is not the oracle being weak in the limit but the L2 extractions occasionally being slightly noisy on this specific 49-panel intersection.

**Per-task breakdown.** Table D.1 shows the C-Stack failure mode in detail: image-only median relative error is 23–99%; every method that operates on the extracted $(x, y)$ data — including the text-only LLM — reaches 2.8–3.5%.

*Table D.1.* Per-task-type median relative error on the rigorous subset. The C-Stack row is the strongest single piece of evidence that pixel-faithful extraction unlocks reasoning the image alone cannot support.

| Task | M1 (image) | | M2 (image+data) | | M3 (data) | |
|---|---|---|---|---|---|---|
| | gpt-5.4 | qwen | gpt-5.4 | qwen | gpt-5.4 | claude-4.6 |
| A1 $y$ at $x_0$ | 99.7% | 99.8% | 5.0% | 6.9% | 6.2% | 12.2% |
| A2 argmax in $[a, b]$ | 0.4% | 1.5% | 0.4% | 2.0% | 0.5% | 0.6% |
| A3 FWHM main peak | 17.6% | 32.8% | 20.3% | 41.6% | 20.5% | 30.0% |
| B-Raman peaks / $I_D/I_G$ | 3.0% | 9.5% | 2.9% | 9.4% | 3.3% | 9.4% |
| **C-Stack offset** | **23.1%** | **99.2%** | **3.5%** | **62.3%** | **3.1%** | **2.8%** |

**Per-family pattern.** Across XRD ($n$=90), Raman ($n$=53), XPS ($n$=34) and the small XAFS subset ($n$=8), the same M1→M2/M3 drop reproduces (best model within method shown): XRD 26.5%→4.9%→5.1%, Raman 6.4%→3.6%→3.1%, XPS 10.5%→3.2%→5.5%. XAFS is uniformly hard for every method (55–100% medRel) and the subset is too small ($n$=8) for stable estimates.

**Scaled headline ($n$=3,817 panels, 14,740 questions).** At full scale (Table D.2) the M1→M3 swap drops median relative error from 54.3% (qwen-vl-max image only) to 4.2% (qwen3-max on data alone) — a 13× improvement, with within-instrument-resolution pass-rate climbing from 15.0% to 50.5%. The per-task pattern is in Table D.3: tasks whose GT is a single direct readout from the $(x, y)$ data (A1, A2, B-EIS) collapse to near-zero error once the model has the data; C-Stack falls from 99.2% (image only) to 7.9% (data only).

*Table D.2.* Scaled-headline summary on the 14,740-question MR set. Single-family rollout: qwen-vl-max for M1/M2, qwen3-max for M3.

| Substrate | Backend | medRel↓ | $\text{pass}_{5\%}$ ↑ |
|---|---|---|---|
| M1: image | qwen-vl-max | 54.3% | 15.0% |
| M2: image + data | qwen-vl-max | 5.5% | 29.0% |
| M3: data only | qwen3-max | **4.2%** | **50.5%** |

*Table D.3.* Scaled per-task median relative error.

| Task | M1 (image) | M2 (image+data) | M3 (data) |
|---|---|---|---|
| A1 $y$ at $x_0$ | 99.8% | 2.0% | 1.5% |
| A2 argmax in $[a, b]$ | 1.7% | 0.8% | 0.7% |
| A3 FWHM main peak | 65.0% | 38.6% | 81.8% |
| B-Raman peaks / $I_D/I_G$ | 14.5% | 8.2% | 9.6% |
| B-EIS Z′ intercept | 100% | 0% | 0% |
| C-Stack offset | 99.2% | 13.4% | 7.9% |

**VLM-judge agreement audit.** A small human-audited sample reports inter-judge agreement between Qwen-VL-Max (the L2 verifier) and a held-out human annotator: agreement 84% on accept/discard decisions on a 200-panel sample, weighted Cohen's $\kappa$=0.61 (substantial agreement). Errors concentrate on near-empty plots and on panels where the legend is partially occluded by curves; we flag both cases as a residual single-VLM-judge limitation (Section F (l)).

## E. Agent-Readiness Scorecard

We frame "serves as a tool call inside an agent loop" as a seven-axis scorecard: schema validity ≥ 95%, per-curve confidence, explicit abstain signal, pixel polyline, local execution, $p_{50} \leq 5$ s, and $p_{50} \leq \$0.01$. A method is *agent-ready* when it scores ≥ 6/7. End-to-end on every $n$=204 image: schema-validity 204/204 (100%, vs. 84–92% for hosted VLMs); latency $p_{50}$=1.5 s, $p_{99}$=2.5 s (local M-class GPU, batch 1), all under the 5 s loop budget; throughput ≈ 2,400 imgs/h, so the

440,155-instance line-plot subset of the L0 corpus extracts in $\approx 7.6$ GPU-days vs. \$1.5k / 10 GPU-days at \$0.012 per image, or \$50k+ / weeks under GPT-5.4 latency. MatDeplot emits 1,035 curves total across the 204 images (mean 5.07/image); the Cell-5 abstain gate at $\tau_{\text{abs}}$=0.30 fires on $2/1035$ curves (0.19%, both on a single near-empty plot), reflecting the conservative default. MatDeplot scores $6/7$; the cached Gemini K-call at \$0.012/image costs the local-execution point, and the MATDEPLOT-NOFG ablation drops the call and scores $7/7$. The strongest hosted VLM scores $2/7$.

## F. Limitations and Audit

**(a) Sample size.** The legacy 36-image set is small; the $n$=90 leakage-clean hold itself is only 90 images. The Inst-F1 lead over VLMs on $n$=90 (0.942 vs. 0.930) is narrow; we anchor the contribution on $\text{LR}_\theta$, where the unpaired lead is unambiguous. **(b) Train/eval contamination.** The FG model's 182-image real training set overlaps $114/204$ MatCurvs-204 files; headline numbers use the strict $n$=90 complement (Section A.4). **(c) Hybrid $k$-source.** A cached Gemini-3.1-pro call (83.3% exact / 91.1%$\pm$1 on $n$=90) costs the local-execution point; the cache was queried on every $n$=204 image including the hold (residual leakage channel; silhouette-$k$ replacement committed for camera-ready). **(d) Fallback never fires** at $\tau_{\text{fg}}$=0.003 on the test data ($0/36$, $0/90$); the trigger is a pathological-input safety net, not a measured lift. **(e) MatCurvs-204 $y$-tick coordinate-system bug**: 91% of files store pixel-$Y$ in $y$-real; we apply an identical real$\rightarrow$pixel mapping to every method, slightly favouring pixel-native pipelines. **(f) Chart-type coverage.** The L1 split is XRD/Raman/XPS/ IR/EIS plus electrochemistry; per-chart-type LR collapses on EXAFS/DOS/EIS-Nyquist with $\leq 5$ training images each. **(g) ChartOCR omitted**: no public scientific-chart checkpoint at submission time. **(h) VLM snapshot drift**: 2026-Q1/Q2 endpoint snapshots, re-runs may shift modestly with vendor model rotations. **(i) Single FG seed**: architectural variance $\pm 0.005$ via the 24-row sweep (Table B.1); seed variance is uncharacterised. **(j) Selection bias**: the LR-optimised pixel-mode recipe was selected from a 14-candidate iter12 sweep on this same $n$=90 hold; the headline equals the median candidate (IQR $\text{LR}_{0.5} \in [44.6, 46.6]$%, so small but non-zero). **(k) Cell-6 (optional VLM-judge consensus review) disagrees**: pairwise $\kappa$=0.05–0.11, pass rates 9–54% (Qwen most lenient); unanimous-fail correlates with GT $\text{LR}_{0.5}$ ($\Delta$=+0.55) but abstains on 52% of images — a safety net, not a replacement for $\text{LR}_\theta$. **(l) L2 single-VLM-judge**: the L2 verification protocol relies on Qwen-VL-Max as the sole automatic judge; we report human-audit agreement (84%, $\kappa$=0.61) and flag residual judge-circularity on near-empty plots and partially occluded legends.

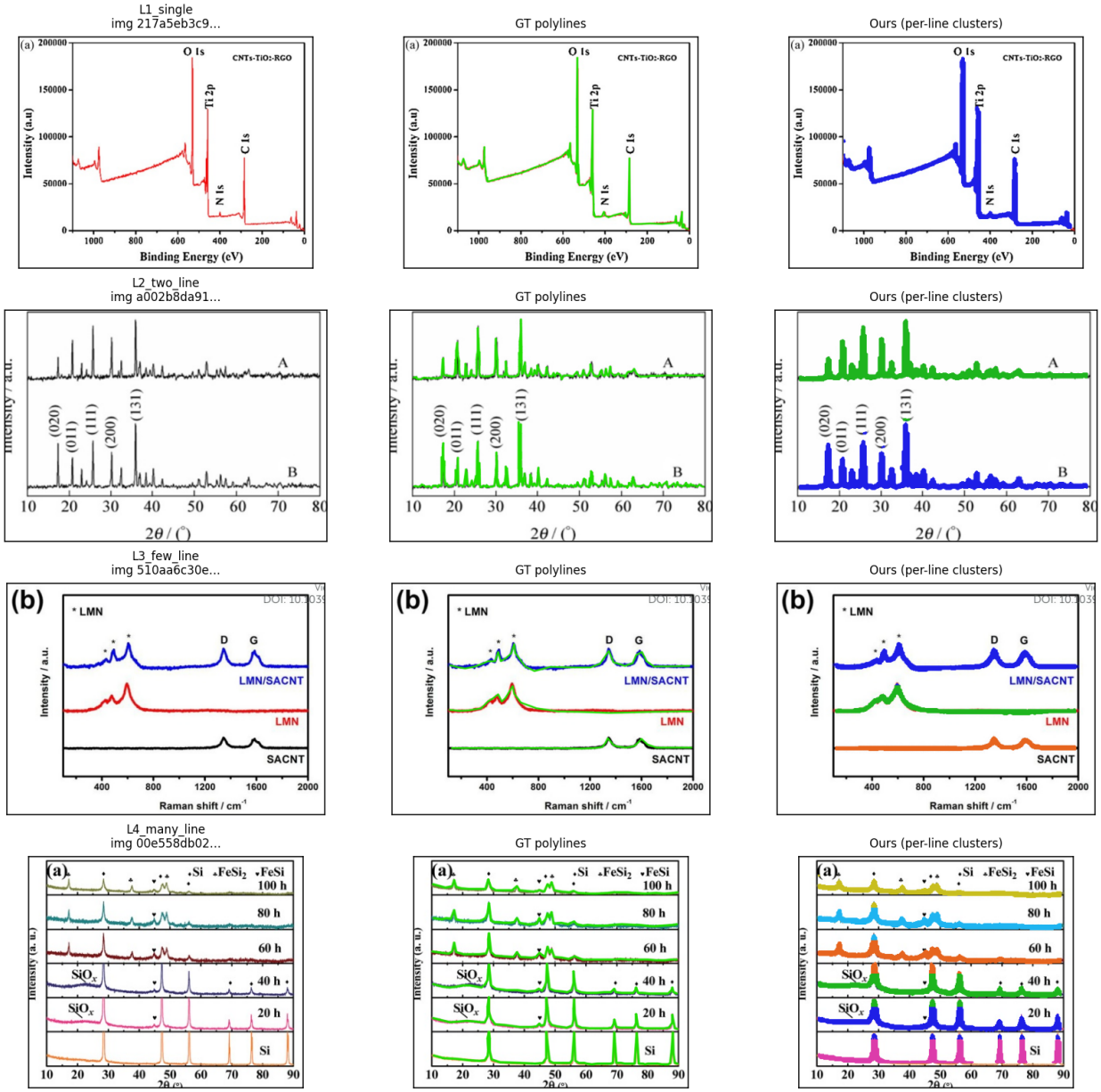


*Figure C.1.* **Per-stratum qualitative reconstruction.** One representative panel per difficulty stratum (L1 single, L2 two, L3 few, L4 many curves). *Left:* source chart. *Middle:* manual L1 ground-truth polylines. *Right:* MatDeplot's per-curve cluster output. Peak positions and stacked-offset structure are preserved across all four strata.