INFOGRAPHICS AS A VISUAL LANGUAGE: LEARNING THROUGH CONTEXT AND RECONSTRUCTION

Anonymous authors
Paper under double-blind review

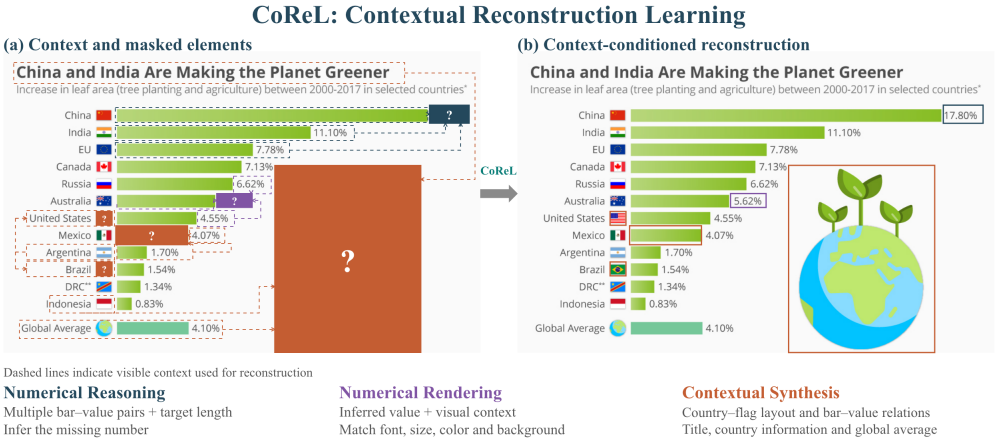


Figure 1: Learning infographic visual language through contextual reconstruction. CoReL uses infographic elements as complementary prediction targets: Numerical Reasoning recovers numerical text; Numerical Rendering expresses the inferred content within the image; Contextual Synthesis completes graphical elements. The surrounding infographic supplies context for all three.

ABSTRACT

Advances in multimodal models have improved chart understanding and image generation, yet it remains unclear how well these models capture the relationships that make an infographic meaningful. Infographics communicate through a visual language in which numerical values, text, and graphical elements complement and constrain one another. We introduce **CoReL (Contextual Reconstruction Learning)**, a data construction and pretraining recipe for learning infographic visual language through contextual reconstruction. CoReL organizes learning around three complementary tasks—**Numerical Reasoning**, **Numerical Rendering**, and **Contextual Synthesis**—that connect numerical meaning, visual expression, and contextual compatibility. We build **InfoContext**, a training corpus of 1.7M images containing 49M annotated regions, including 27M numerical-text bounding boxes and 22M graphical bounding boxes. We also introduce **InfoContext-Bench**, a human-annotated benchmark spanning the three tasks. Both resources will be made publicly available. We apply CoReL to an existing unified understanding-and-generation model through element-masked reconstruction and evaluate its performance across the three tasks. This objective requires the model to comprehensively capture the visual context, particularly the relationships among the visible visual components, the text, and the masked region. Experiments reveal substantial limitations in existing public models, whereas our model consistently achieves stronger performance across all three tasks. Beyond element reconstruction, numerical-reasoning pretraining with CoReL further improves downstream chart question answering on both BAGEL and Qwen, with consistent gains across benchmarks. These results establish contextual reconstruction as an effective framework for learning infographic visual language, with benefits extending from element completion to reasoning across chart elements.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Infographics communicate through relationships among numerical values, text, and graphical elements. A bar expresses a quantity relative to a scale, its label names that quantity, and an icon connects it to a category or topic. Together, these elements form an *infographic visual language*: understanding an element requires interpreting its context, while generating it requires preserving the relationships that give it meaning.

Existing chart question answering (QA) and image generation provide incomplete supervision for these relationships. Whole-image questions need not require connecting a value to its visual encoding. For example, Mukherjee et al. (2026) report that 81% of the human-labeled ChartQA test questions can be answered from textual annotations alone. InfoChartQA further shows that infographic-specific presentation challenges current vision-language models (VLMs) (Xie et al., 2025). On the generation side, producing a plausible-looking chart does not ensure numerical correctness or preservation of surrounding information (Tang et al., 2026; Li et al., 2026a). These observations motivate learning these relationships through individual infographic elements and their context.

We propose **CoReL** (Contextual Reconstruction Learning), which explicitly models these relationships by masking elements and reconstructing them from the remaining context (Figure 1). Removing a bar's numerical label creates an inference problem grounded in other value–mark pairs. Drawing the label also requires expressing the inferred value in a compatible visual form. Removing a graphical element reverses the direction: the model must recover geometry or pictorial content from numerical and semantic context. These complementary prediction directions define **Numerical Reasoning**, **Numerical Rendering**, and **Contextual Synthesis**. The original elements serve as targets, yielding many reconstruction examples per image without separately authored QA pairs.

The modeling contribution is formulating infographic learning as conditional reconstruction across text and image outputs. We instantiate it in BAGEL (Deng et al., 2025), an existing unified understanding-and-generation model, using its native objectives. Scaling contextual reconstruction requires infographics whose chart elements (e.g., numerical labels, quantitative marks, and human-recognizable objects (HROs)) provide mutual context, with annotations localizing elements for masking and reconstruction. InfoDet (Zhu et al., 2026) provides rich element annotations for approximately 100K infographics, while ChartGalaxy (Li et al., 2026b) offers a million-scale collection of diverse infographic designs. To enable contextual reconstruction on this larger collection, we introduce **InfoContext**, which augments ChartGalaxy's synthetic infographics with fine-grained annotations of numerical text and graphical elements. With 1.7M images and 49M annotated regions, including 27M numerical-text and 22M graphical bounding boxes, InfoContext supplies localized supervision for both textual and visual reconstruction, enabling CoReL to learn diverse element–context relationships. To evaluate recovery of these relationships under missing information, we introduce **InfoContext-Bench**, which tests numerical and graphical reconstruction under controlled removal of infographic elements. **Both resources will be publicly available**.

Experiments establish two complementary benefits. On reconstruction, CoReL outperforms the evaluated open models across the three tasks, with a growing numerical-reasoning advantage as more context is removed. Supplying correct numbers or descriptions of missing objects helps image-editing baselines, yet substantial gaps remain. On downstream chart understanding, Numerical Reasoning pretraining consistently improves both BAGEL and Qwen3.5-4B (Qwen Team, 2026b) across two public benchmarks: in-domain ChartGalaxy QA (Li et al., 2026b) and out-of-domain, unseen-task EncQA (Mukherjee et al., 2026). These gains across models and domains connect element reconstruction to broader reasoning over chart elements.

Our contributions are: (1) a modeling formulation that learns numerical, visual, and semantic relationships through three complementary reconstruction tasks; (2) InfoContext, a 1.7M-image corpus with 49M element-level annotations enabling contextual reconstruction learning at scale, and InfoContext-Bench, a human-annotated benchmark assessing numerical inference and context-consistent visual reconstruction; and (3) substantial improvements in numerical reasoning and visual reconstruction, with downstream chart-QA gains across two model families and an out-of-domain benchmark, demonstrating transferable chart understanding through reconstruction-based learning.

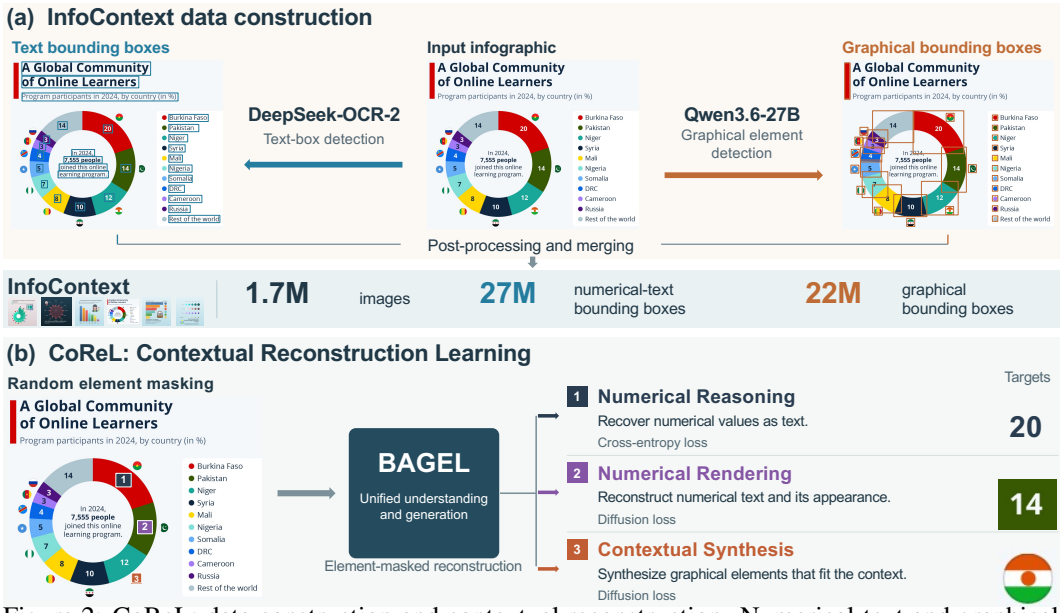


Figure 2: CoReL: data construction and contextual reconstruction. Numerical-text and graphical regions turn existing infographics into reconstruction examples. Randomly masking these elements yields text targets for Numerical Reasoning and image targets for Numerical Rendering and Contextual Synthesis. The three tasks share BAGEL's existing backbone.

2 Related Work

Chart and infographic understanding. ChartQA studies visual and logical reasoning over charts (Masry et al., 2022). ChartGalaxy expands infographic design diversity and data scale (Li et al., 2026b); InfoDet provides annotations of quantitative marks and human-recognizable objects (HROs) (Zhu et al., 2026). EncQA targets reasoning over visual encodings, while InfoChartQA examines infographic-specific difficulties (Mukherjee et al., 2026; Xie et al., 2025). These resources motivate interpreting element relationships beyond reading text. InfoContext supplies such reconstruction supervision; InfoContext-Bench measures recovery of withheld content.

Reconstruction and visual-language pretraining. Masked reconstruction learns representations from existing data (He et al., 2022). Pix2Struct uses screenshot parsing, MatCha combines chart derendering and mathematical reasoning, ChartT5 predicts masked table content, and UniChart combines extraction with chart reasoning (Lee et al., 2023; Liu et al., 2023b; Zhou et al., 2023; Masry et al., 2023). DePlot translates charts into tables for subsequent language-model reasoning (Liu et al., 2023a). CoReL applies this principle to infographic elements: a numerical region supports textual inference and visual realization, while graphical reconstruction learns dependence on numerical and semantic context. The formulation spans understanding and generation.

Unified models and information-preserving generation. BAGEL unifies text prediction and image generation through multimodal pretraining (Deng et al., 2025). Recent editors, including Qwen-Image-Edit-2511, FireRed-Image-Edit-1.1, and LongCat-Image-Edit, support instruction-based image manipulation (Qwen Team, 2025b; FireRed Team, 2026; LongCat Team, 2025). IGenBench and ChartE^3 assess information reliability in infographic generation and chart editing (Tang et al., 2026; Li et al., 2026a). Our setting requires inferring content unspecified by instructions, unlike assisted settings supplying values or target descriptions; we assess correctness and preservation.

3 Contextual Reconstruction Learning

CoReL unifies three complementary tasks: Numerical Reasoning predicts text tokens, while Numerical Rendering and Contextual Synthesis generate images.

Elements as conditional prediction targets. Let I be an infographic with N elements $\mathcal{E} = \{(b_i, k_i, y_i)\}_{i=1}^{N}$, where b_i is a bounding box, k_i is its type, and y_i is its numerical transcription when applicable. For task $t \in \{\text{reason}, \text{render}, \text{synth}\}$, let $\mathcal{M}_t \subseteq \{1, \ldots, N\}$ index the masked elements. The masked image is $\widetilde{I}_t = \text{Mask}(I, \{b_i : i \in \mathcal{M}_t\})$. The two numerical tasks select numerical-text elements, whereas Contextual Synthesis selects graphical elements. The input is the masked image, with solid boxes marking target locations, plus a task instruction and textual context. The reconstruction target is the masked content. This supervises an element's dependence on surrounding evidence rather than a global image description. Masking multiple elements further reduces available context. The shared model f_θ^t uses parameters θ for task t; instructions are implicit.

Numerical Reasoning. The model predicts masked numerical values as text:

$$\widehat{\mathbf{y}} = (\hat{y}_i)_{i \in \mathcal{M}_{\text{reason}}} = f_\theta^{\text{reason}}(\widetilde{I}_{\text{reason}}), \tag{1}$$

where $\widehat{\mathbf{y}}$ lists predicted numerical transcriptions in reading order, and $\hat{y}_i$ denotes masked element i's predicted text. For example, comparing a bar's extent with visible bars and their values can reveal its masked label. Textual sequences or aggregate constraints can also reveal some targets. Benchmark difficulty levels distinguish these cases rather than assuming uniform evidence requirements.

Numerical Rendering. Masking the same elements, this task generates a completed image:

$$\widehat{I}_{\text{render}} = f_\theta^{\text{render}}(\widetilde{I}_{\text{render}}, a), \tag{2}$$

where $\widehat{I}_{\text{render}}$ is the generated image and a is auxiliary textual input, such as the chart title. In the primary *context-only* condition, a does not disclose the target values: the model must infer them and render readable labels with appropriate placement and style. A diagnostic *value-conditioned* evaluation supplies the correct values in a, reducing inference requirements while still requiring accurate rendering and chart preservation.

Contextual Synthesis. The model reconstructs masked graphical elements in an image:

$$\widehat{I}_{\text{synth}} = f_\theta^{\text{synth}}(\widetilde{I}_{\text{synth}}, c), \tag{3}$$

where $\widehat{I}_{\text{synth}}$ is the generated image and c is available textual context. Targets include chart elements, such as bars and pie sectors, and HROs, such as icons and illustrations. A reconstructed bar should agree with its value and surrounding scale; a reconstructed HRO should fit the infographic's topic and composition. The primary evaluation supplies the existing chart title for topical context without explicitly describing the masked object.

One backbone, complementary objectives. We instantiate CoReL in BAGEL (Deng et al., 2025) without additional modules, sampling tasks at 1:1:1. The objective is:

$$\mathcal{L} = \underbrace{\lambda_{\text{CE}}\mathcal{L}_{\text{CE}}^{\text{reason}}}_{\text{Numerical Reasoning}} + \underbrace{\lambda_{\text{FM}}\mathcal{L}_{\text{FM}}^{\text{render}}}_{\text{Numerical Rendering}} + \underbrace{\lambda_{\text{FM}}\mathcal{L}_{\text{FM}}^{\text{synth}}}_{\text{Contextual Synthesis}}. \tag{4}$$

Here $\mathcal{L}_{\text{CE}}$ denotes cross-entropy on numerical text tokens, and $\mathcal{L}_{\text{FM}}$ denotes flow-matching reconstruction loss on VAE image latents (Deng et al., 2025; Lipman et al., 2023); superscripts identify task contributions. We use $\lambda_{\text{CE}} = 0.25$ and $\lambda_{\text{FM}} = 1$. Numerical Reasoning thus contributes the text-token loss, while Numerical Rendering and Contextual Synthesis contribute the two image reconstruction terms. Original pixels and numerical transcriptions supply supervision. Numerical targets encourage precise value recovery; image targets require visual expression or context-compatible graphical completion. Appendix A specifies masking and optimization.

4 InfoContext and InfoContext-Bench

4.1 InfoContext: Scalable Element-Level Supervision

We construct InfoContext from ChartGalaxy's synthetic infographics (Li et al., 2026b), with InfoDet's element vocabulary (Zhu et al., 2026) and annotations for graphical reconstruction. DeepSeek-OCR-2 (Wei et al., 2026) extracts text regions and transcriptions, and Qwen3.6-27B (Qwen Team,

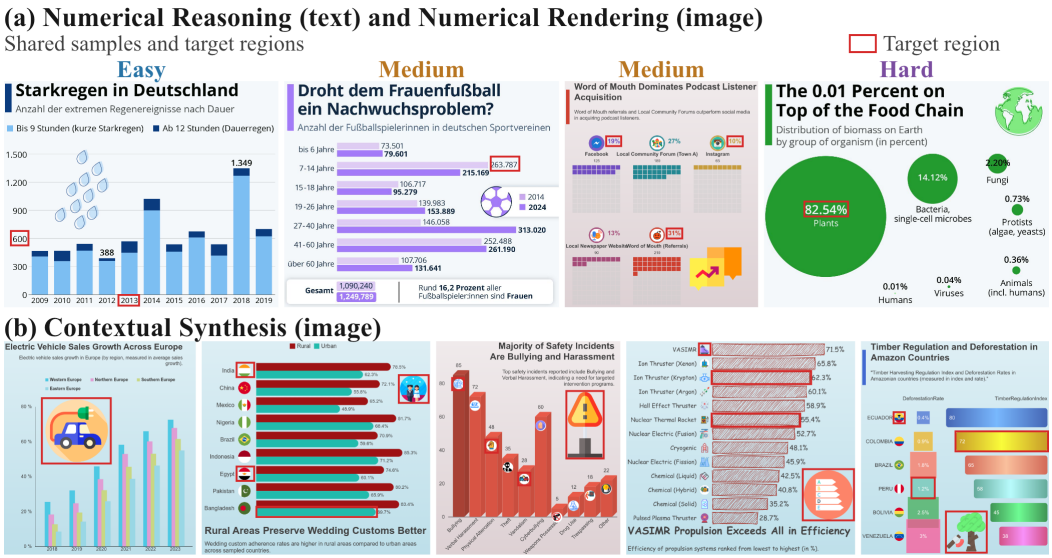


Figure 3: InfoContext-Bench. Numerical Reasoning and Numerical Rendering share numerical targets at three difficulty levels: Easy, Medium, and Hard. Their outputs differ: text versus a completed image. Contextual Synthesis covers Structural Targets (chart elements) and Semantic Targets (human-recognizable objects; HROs). Red outlines mark targets in reference examples.

Table 1: Shared numerical targets: 712 images and 5,353 annotated boxes for both tasks.

Source	Images	Easy boxes	Medium boxes	Hard boxes	Total boxes
Synthetic	391	792	1,216	24	2,032
Real	321	1,000	2,000	321	3,321
Total	712	1,792	3,216	345	5,353

2026c) detects graphical components. Filtering and merging detections yield 1.7M images with 27M numerical-text boxes and 22M graphical boxes (Figure 2). These count source images and regions, not separately authored questions. Target sampling yields multiple reconstruction examples per image. Numerical annotations support both numerical tasks; graphical annotations support Contextual Synthesis.

The training data are synthetic; real ChartGalaxy images are reserved for evaluation. All benchmark images are excluded from training. Synthetic numerical benchmark images come from ChartGalaxy's official QA evaluation split. Before graphical test sampling, DINOv2 feature clustering reduces redundancy (Oquab et al., 2024). Construction and separation details are in Appendix A.

4.2 InfoContext-Bench: Recovery from Incomplete Context

InfoContext-Bench evaluates recovery of missing numerical and graphical content from visible infographic context. Numerical Reasoning tests inference from textual evidence and quantitative visual encodings. Numerical Rendering tests in-place label restoration while preserving surrounding content, as when repairing incomplete charts. Contextual Synthesis assesses graphical completion under numerical and semantic constraints, such as recovering an occluded bar or a missing HRO.

Numerical targets. The two numerical tasks share 5,353 manually annotated numerical-text boxes across 712 images (391 synthetic and 321 real; Table 1). These shared targets match content across text and image outputs. We define three difficulty levels. Easy relies on local textual regularities; Medium uses relationships among numerical and graphical elements; and Hard requires these relationships with spatially sparser or more distant evidence. Numerical Reasoning uses single and multiple masks; Numerical Rendering uses context-only and value-conditioned inputs.

Graphical targets. Contextual Synthesis uses InfoDet's human annotations (Zhu et al., 2026) to evaluate two graphical target categories: **Structural Targets**, covering chart elements, and **Semantic Targets**, covering human-recognizable objects (HROs). This separates reconstruction of low-level

Table 2: Numerical Reasoning. Micro and difficulty scores (↑). Multi-mask columns give the maximum targets masked per image. GPT-5.5-Pro is a separate commercial reference.

Model	Single mask				Multi-mask Micro			
	Micro	Easy	Medium	Hard	2	3	4	5
Commercial reference								
GPT-5.5-Pro	97.03	98.33	97.30	87.55	95.95	95.21	94.42	94.29
Chart/document baselines								
CLI-7B	45.00	30.58	54.52	31.13	44.29	42.52	40.94	39.65
Granite-4.0-3B-Vision	46.48	31.17	56.36	33.79	44.44	43.14	40.92	39.24
General-purpose VLMs								
LLaVA-OneVision-2-8B	45.74	29.23	56.17	34.32	45.67	43.29	41.46	40.16
Kimi-VL-A3B-2506	46.46	31.44	56.47	31.20	39.05	32.48	26.71	24.19
Qwen2.5-VL-7B	46.85	30.72	56.69	38.97	48.24	44.84	42.73	41.22
MiniCPM-V-4.5	48.12	33.33	57.76	35.13	47.29	45.89	44.04	42.59
InternVL3.5-8B	50.13	36.28	58.91	40.17	48.15	44.14	42.48	41.06
Qwen3-VL-8B	50.99	32.20	62.63	40.08	51.89	48.56	46.59	44.78
Ovis2.5-9B	55.88	35.06	68.77	43.85	55.01	50.85	48.11	46.10
GLM-4.1V-9B	56.69	34.28	70.47	44.71	47.81	43.54	39.04	35.92
Qwen3.5-9B	64.50	58.07	70.28	44.07	59.41	55.84	53.14	51.13
Qwen3.6-27B	82.05	87.71	81.77	55.22	71.70	66.91	62.86	61.20
Contextual reconstruction pretraining								
CoReL (ours)	**85.19**	**92.71**	**84.10**	**56.22**	**84.31**	**83.90**	**83.81**	**83.55**

Bold: best among the evaluated open models. Exact model versions and references are in Appendix A.4.

quantitative structures from completion of high-level semantic entities. The Structural evaluation comprises 200 targets (50 each for **bars**, **areas under lines**, **lines**, and **stacked bars**), while the Semantic evaluation contains 219 images with one recognizable object per image. All primary comparisons use the masked image and its own title.

5 Experiments

We first evaluate reconstruction, then test transfer to downstream QA. The main model jointly trains all three CoReL tasks on BAGEL for 60k steps. Numerical baselines span general VLMs and chart/document models; GPT-5.5-Pro is reported separately as a commercial capability reference (OpenAI, 2026). Image comparisons use Qwen-Image-Edit-2511, FireRed-Image-Edit-1.1, and LongCat-Image-Edit. Appendix A.4 lists exact model versions and references. Direct comparisons assess our task-adapted model against released baselines; downstream experiments separately test pretraining benefits for related chart understanding.

Image reconstruction metrics. Both image reconstruction tasks use complementary visual fidelity measures. MAE and PSNR quantify pixel fidelity, while SSIM measures local structural similarity (Wang et al., 2004). LPIPS and DISTS assess perceptual differences (Zhang et al., 2018; Ding et al., 2022), and CLIP-I and DINO measure similarity between generated and reference image embeddings (Radford et al., 2021; Oquab et al., 2024). These measures assess reconstruction beyond pixel agreement. Implementation details are in Appendix B.4.

5.1 Numerical Reasoning

For target i, we parse predicted and reference texts $\hat{y}_i, y_i$ into values $\hat{v}_i, v_i$ and compute

$$s(\hat{v}_i, v_i) = 100 \max\left(0, 1 - \frac{|\hat{v}_i - v_i|}{|v_i| + \epsilon}\right), \qquad \epsilon = 10^{-8}. \tag{5}$$

Micro averages the score over target boxes. Missing or invalid predictions score zero. *All-correct accuracy* requires exact numerical recovery of every target in an input. Parsing and aggregation details are in Appendix B.

Results. We mask one to five regions, progressively reducing context. With one mask, the 7B-parameter CoReL scores 85.19 Micro, outperforming all evaluated open models, including the larger

Table 3: Numerical Rendering on the full numerical benchmark. Numerical recovery compares target-crop readout with ground-truth values; preservation metrics exclude all target boxes.

Model	Numerical recovery		Visual preservation		
	Score ↑	Exact (%) ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Context-only (infer and render)					
LongCat-Image-Edit	3.65	0.41	21.06	0.9051	0.0875
FireRed-Image-Edit-1.1	22.56	1.61	17.03	0.8208	0.2292
Qwen-Image-Edit-2511	14.63	0.86	18.06	0.8481	0.1464
CoReL (ours)	**75.51**	**55.67**	**36.42**	**0.9860**	**0.0048**
Value-conditioned (correct target values supplied)					
LongCat-Image-Edit	29.20	24.42	23.59	0.9265	0.0569
FireRed-Image-Edit-1.1	44.64	31.81	17.39	0.8547	0.1887
Qwen-Image-Edit-2511	59.83	51.13	19.91	0.8724	0.1148
CoReL (ours)	**82.78**	**78.85**	**36.97**	**0.9865**	**0.0038**

Qwen3.6-27B (82.05). With more masks, all models decline, but CoReL degrades more slowly: its lead over Qwen3.6-27B grows from 3.14 to 22.35 points with one to five masks (83.55 versus 61.20).

Single-mask scores are 92.71/84.10/56.22 on Easy/Medium/Hard targets, respectively; sparse numerical references remain challenging. For upper-bound context, the substantially larger commercial GPT-5.5-Pro reaches 97.03 and 94.29 in the single- and five-mask settings. This reference uses the API's default tools, without matching size or compute. Fixed-target comparisons appear in Appendix D (Table 12).

5.2 Numerical Rendering

We evaluate Numerical Rendering on two axes: restored-value correctness through text readout, and visual preservation of unmasked content. A Qwen3-VL-8B-Instruct reader (Qwen Team, 2025a) extracts target-crop numbers for comparison with ground truth: Score measures numerical proximity (Equation 5), and Exact is the percentage of exact numerical matches. PSNR, SSIM, and LPIPS measure unintended changes outside all target boxes.

Results. Table 3 reports both input conditions on the full numerical benchmark. In context-only completion, CoReL scores **75.51**, compared with 22.56 for the best baseline. To aid the public editors, we provide each target's ground-truth value in the value-conditioned setting, leaving only rendering and preservation. This consistently improves all three editors, but their best value-conditioned score, 59.83, remains below CoReL's context-only score. With matched value-conditioned inputs, CoReL reaches **82.78** with **78.85%** exact matches. It also achieves the best preservation scores in both conditions. These results link learned element–context relationships to numerical recovery and localized visual expression. The reader's score on original crops is 86.55; normalized scores and full diagnostics are in Appendix E.

5.3 Contextual Synthesis

We evaluate Structural Targets (chart elements) on completed whole images and Semantic Targets—human-recognizable objects (HROs) (Zhu et al., 2026)—on both whole images and target crops. Compared with simple marks like solid-color bars, HROs occupy larger regions with richer visual detail and recognizable object-level semantics. Crop metrics thus isolate reconstructed object appearance and structure from unchanged surroundings, while whole-image metrics assess completion and preservation. For Semantic Targets, FID also compares generated and reference image distributions (Heusel et al., 2017).

Results. Table 4 summarizes title-conditioned results. Figure 4 (third row) and Figures 8–9 show qualitative results. On Structural Targets—bars, areas under lines, lines, and stacked bars—CoReL leads on all seven aggregate whole-image metrics (Appendix F.1); for example, MAE falls from the best baseline's 43.46 to **20.26**. For Semantic Targets, it reduces whole-image FID from 26.31 to **21.00**

Table 4: Contextual Synthesis with title conditioning. Structural Targets comprise 200 chart elements across four classes; Semantic Targets comprise one HRO in each of 219 images. All models receive the masked image and its title.

Model	Structural Targets: whole image				Semantic Targets		
	MAE ↓	LPIPS ↓	CLIP-I ↑	DINO ↑	Full FID ↓	Crop FID ↓	Crop DISTS ↓
Qwen-Image-Edit-2511	43.46	0.4199	0.8318	0.5876	27.71	236.25	0.5662
FireRed-Image-Edit-1.1	64.73	0.6364	0.7746	0.4251	36.48	239.32	0.5876
LongCat-Image-Edit	65.68	0.6224	0.7614	0.4195	26.31	235.33	0.5382
CoReL (ours)	**20.26**	**0.2390**	**0.8624**	**0.6588**	**21.00**	**118.70**	**0.3638**

Full: completed whole image. Crop: target region. Full metrics and caption-assisted results are in Appendix F.

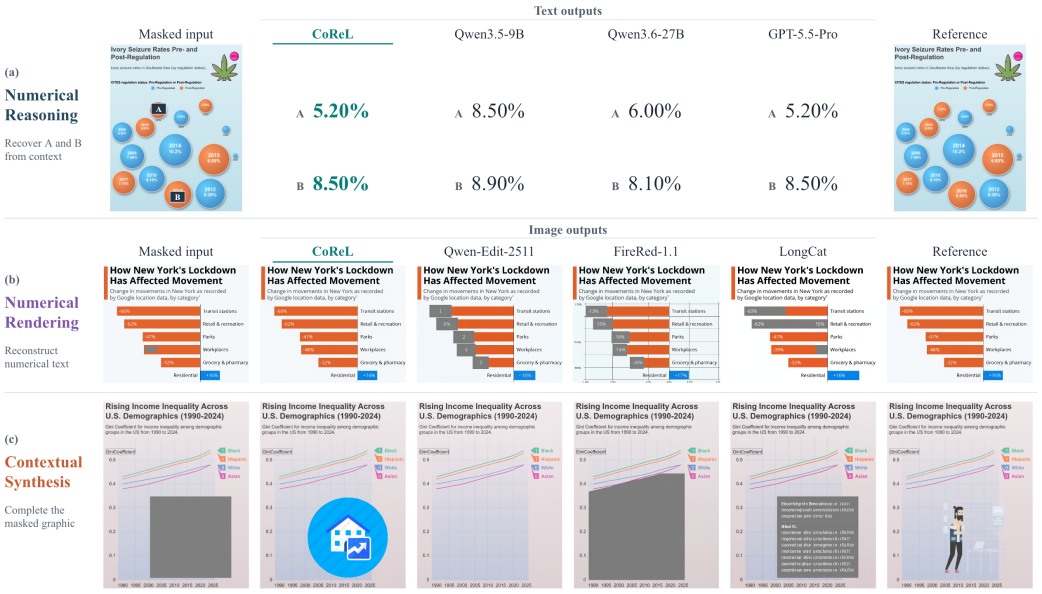


Figure 4: Qualitative results across the three tasks. Numerical Reasoning predicts missing values as text. Numerical Rendering restores labels within the image. Contextual Synthesis completes a masked graphical region. CoReL uses visible context for localized completions; public editors can leave masks unfilled or introduce unrelated content. More examples appear in Appendix C.

and crop FID from 235.33 to **118.70**. We also give baselines VLM descriptions of the unmasked reference image (image captions) or target region (object captions); CoReL uses only the title. Object captions improve the best baseline crop FID to 136.31, still above our title-conditioned result. Further analysis is in Appendix F.2 (Table 17).

5.4 Context Ablation

To test whether contextual relationships guide CoReL's reconstruction, we compare three controlled conditions with the same masked six-square target (Figure 5).

With full context (a), $+6\%$ and surrounding value–mark pairs are available: CoReL renders six orange squares, matching the reference count and form despite differing in color; Qwen-Image-Edit-2511 leaves the mask unfilled. However, erasing only $+6\%$ (b) removes the explicit quantity cue and breaks the six-unit encoding. Furthermore, erasing the $+3\%$ and $+63\%$ value–mark pairs on the left while retaining $+6\%$ (c) removes square-based examples; CoReL instead produces a circle, consistent with the remaining circular encodings on the right. This suggests CoReL uses relevant context during inference: the target value guides quantity, while surrounding value–mark pairs guide visual form.

5.5 Transfer to Chart Question Answering

Beyond reconstruction on InfoContext-Bench, we ask whether CoReL benefits downstream chart understanding. To assess transfer across backbones and domains, we evaluate BAGEL and Qwen3.5-4B (Deng et al., 2025; Qwen Team, 2026b) on two public QA benchmarks, ChartGalaxy and

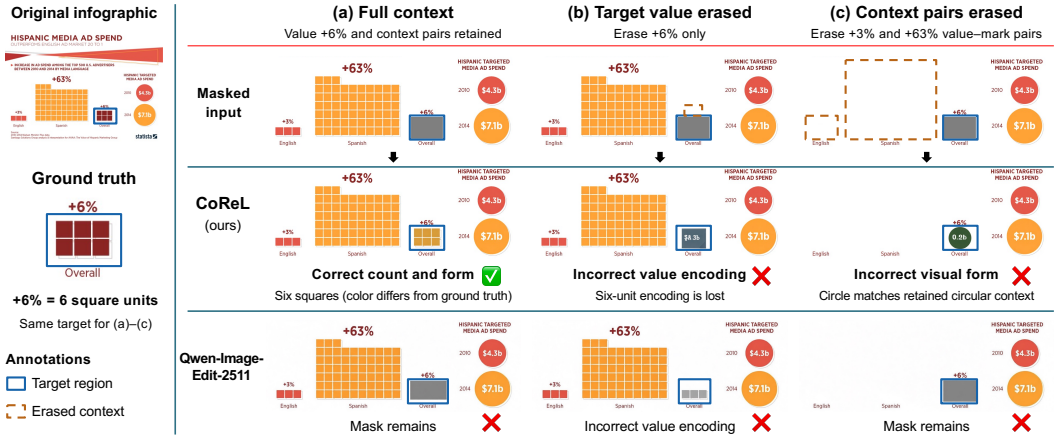


Figure 5: Context ablation with a fixed target. The ground truth is six square units. We vary the visible evidence: (a) full context; (b) target value erased; (c) neighboring value–mark pairs erased.

Table 5: Transfer to chart QA. Base: original backbone; CoReL: Numerical Reasoning pretraining followed by QA SFT. Accuracy (%, ↑); bold: best per backbone and column.

Training	ChartGalaxy						EncQA				
	Overall	Calc.	Comp.	Count	Ident.	Trend	Overall	CV	CVR	FV	FE
BAGEL											
Base	62.10	45.27	**85.46**	31.64	72.65	87.07	73.33	70.00	58.67	56.22	97.56
QA SFT	70.14	63.16	84.09	32.37	80.85	**95.69**	67.04	61.67	56.00	44.44	96.89
CoReL (ours)	**72.14**	**63.81**	84.09	**41.79**	**82.58**	91.38	**76.52**	**74.67**	**60.00**	**60.67**	**99.11**
Qwen3.5-4B											
Base	33.67	24.90	37.33	18.36	43.21	46.55	74.96	64.33	66.00	65.56	94.44
QA SFT	74.65	66.93	89.19	38.89	85.81	91.38	78.67	67.67	74.00	68.89	97.33
CoReL (ours)	**76.82**	**67.83**	**89.59**	**49.03**	**86.45**	**94.83**	**82.89**	**75.67**	**76.67**	**74.44**	**98.22**

Calc.: Calculation; Comp.: Comparison; Count: Counting; Ident.: Identification.
CV: Correlate Values; CVR: Correlate Values Relative; FV: Filter Values; FE: Find Extrema.

EncQA (Li et al., 2026b; Mukherjee et al., 2026). EncQA evaluates how accurately models interpret and compare chart elements across visual encodings such as position, length, area, and color. It also tests out-of-domain transfer: its charts differ from our training distribution, and we use no EncQA-domain data for CoReL pretraining or QA SFT.

For each backbone, we first apply CoReL Numerical Reasoning pretraining followed by QA SFT, using identical SFT data and training steps across models. Since ChartGalaxy's training QA annotations were not publicly released, we annotate its training images following the original QA-generation protocol (Li et al., 2026b). Each SFT run uses 140k images and processes 1.28M QA examples over 15k steps. To assess transfer from CoReL, we also train a counterpart on the same QA SFT data and steps without prior Numerical Reasoning pretraining. Data construction details are in Appendix G.

Relative to QA SFT alone, CoReL improves ChartGalaxy overall accuracy by 2.00/2.17 percentage points and EncQA accuracy by 9.48/4.22 points for BAGEL/Qwen, respectively (Table 5). Counting and value filtering also improve on both backbones, supporting transfer from masked reconstruction to broader chart reasoning. Gains vary across individual tasks (Appendix G); Figure 10 illustrates ChartGalaxy counting questions and the four evaluated EncQA question types.

6 Conclusion

CoReL formulates infographic learning as reconstruction of elements from their context. Its three tasks connect numerical inference, visual expression, and graphical compatibility, supported by InfoContext and InfoContext-Bench. Joint training improves reconstruction over the evaluated open baselines, while Numerical Reasoning pretraining improves chart QA on two model families under identical QA SFT data. These results support element-level reconstruction as a practical way to learn relationships that matter for both understanding and generation.

Reproducibility Statement

Section 3 defines the tasks and objective, and Section 4 describes InfoContext and InfoContext-Bench. Appendix A specifies data separation, masking, joint-training settings, and model versions. Appendix B defines evaluation, Appendices D–G provide detailed results, and Appendix I gives the evaluation instructions. InfoContext and InfoContext-Bench will be publicly available.

AI Use Statement

AI tools assisted with manuscript organization, language revision, literature discovery, figure formatting, and LaTeX preparation. OCR and multimodal models also support element annotation, sampling metadata, QA-data generation, auxiliary captions, and numerical readout; their roles are described in the paper.

Ethics Statement

This work uses existing infographic datasets and inherited or automatically constructed element labels. Automatically produced transcriptions and generated numerical content can contain errors. Visually plausible chart completion should not be treated as independently verified data.

References

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Cheng Chen, Yuyu Guo, Pengpeng Zeng, Jingkuan Song, Peng Di, Hang Yu, and Lianli Gao. From one-to-one to many-to-many: Dynamic cross-layer injection for deep vision-language fusion. In *European Conference on Computer Vision*, pp. 616–634, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

FireRed Team. FireRed-Image-Edit-1.1. Official model card, 2026. URL https://huggingface.co/FireRedTeam/FireRed-Image-Edit-1.1.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

IBM Granite Vision Team. Granite 4.0 Vision. Official model card, 2026. URL https://huggingface.co/ibm-granite/granite-4.0-3b-vision.

Sergey Kastryulin, Jamil Zakirov, Denis Prokopenko, and Dmitry V. Dylov. PyTorch Image Quality: Metrics for image quality assessment. *arXiv preprint arXiv:2208.14818*, 2022.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18893–18912, 2023.

Shuo Li, Jiajun Sun, Zhekai Wang, Xiaoran Fan, Hui Li, Dingwen Yang, Zhiheng Xi, Yijun Wang, Zifei Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. ChartE^3: A comprehensive benchmark for end-to-end chart editing. In *International Conference on Machine Learning*, 2026a.

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, Xinhuan Shu, and Shixia Liu. ChartGalaxy: A dataset for infographic chart understanding and generation. In *International Conference on Learning Representations*, 2026b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023a.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770, 2023b.

LongCat Team. LongCat-Image-Edit. Official model card, 2025. URL https://huggingface.co/meituan-longcat/LongCat-Image-Edit.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14662–14684, 2023.

Moonshot AI. Kimi-VL-A3B-Thinking-2506. Official model card, 2025. URL https://huggingface.co/moonshotai/Kimi-VL-A3B-Thinking-2506.

Kushin Mukherjee, Donghao Ren, Dominik Moritz, and Yannick Assogba. EncQA: Benchmarking vision-language models on visual encodings for charts. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):648–658, 2026.

OpenAI. Introducing GPT-5.5. Official model announcement, 2026. URL https://openai.com/index/introducing-gpt-5-5/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Qwen Team. Qwen3-VL-8B-Instruct. Official model card, 2025a. URL https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct.

Qwen Team. Qwen-Image-Edit-2511. Official model card, 2025b. URL https://huggingface.co/Qwen/Qwen-Image-Edit-2511.

Qwen Team. Qwen3.5-9B. Official model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.5-4B. Official model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Qwen Team. Qwen3.6-27B. Official model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.6-27B.

Qwen Team. Qwen3.8-27B. Official model card, 2026d. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, Luoxuan Weng, Xiuqi Huang, Minfeng Zhu, Yingchaojie Feng, Yuyu Luo, and Wei Chen. IGenBench: Benchmarking the reliability of text-to-infographic generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36919–36954, 2026.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: Image processing in Python. *PeerJ*, 2:e453, 2014.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

Tianchi Xie, Minzhi Lin, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. InfoChartQA: A benchmark for multimodal question answering on infographic charts. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. Enhanced chart understanding via visual language pre-training on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1314–1326, 2023.

Jiangning Zhu, Yuxing Zhou, Zheng Wang, Juntao Yao, Yima Gu, Yuhui Yuan, and Shixia Liu. InfoDet: A dataset for infographic element detection. In *International Conference on Learning Representations*, 2026.

A Data Construction and Experimental Setup

A.1 Constructing InfoContext

InfoContext turns existing infographic images into supervision for element reconstruction. The synthetic ChartGalaxy images provide a large and diverse source of chart layouts (Li et al., 2026b). DeepSeek-OCR-2 extracts text boxes and their transcriptions, while Qwen3.6-27B detects graphical elements (Wei et al., 2026; Qwen Team, 2026c). Filtering and merging the outputs yield numerical-text regions and graphical regions. Numerical-text regions can supply either text targets or image targets; graphical regions supply image targets. InfoDet provides the element taxonomy and human annotations used for graphical reconstruction (Zhu et al., 2026). We additionally use 45k annotated InfoDet images for the Contextual Synthesis task.

The corpus contains 1.7M source images, 27M numerical-text boxes, and 22M graphical boxes. These counts distinguish source images from regions: masking different subsets of the regions can produce multiple training examples from one image. Original pixels supply the image reconstruction targets, and numerical transcriptions supply text targets. The supervision is self-supervised with respect to the withheld image content, while element localization and transcription rely on external annotation models or inherited annotations.

A.2 Benchmark sampling and separation

The numerical benchmark includes 391 synthetic images sampled across 15 categories of the official ChartGalaxy QA evaluation split and 321 real images. To diversify real-image sampling, a VLM assigns interaction type, recoverability, and constraint type to a pool of 59k images; sampling is balanced across combinations of these attributes. The numerical targets are manually annotated and assigned to Easy, Medium, or Hard, with counts in Table 1. Both Numerical Reasoning and Numerical Rendering use the full set of 712 images and 5,353 boxes.

Graphical benchmark construction uses InfoDet's existing human annotations. We extract DINOv2 features, cluster the images, select a representative from each cluster, and sample the test set from these representatives. This reduces visual redundancy before evaluation. The Structural evaluation contains 200 targets, with 50 each for bars, areas under lines, lines, and stacked bars. In the annotation vocabulary these correspond to `bar_mark`, `area-under-line_mark`, `line_mark`, and `stacked-bar_mark`. The Semantic evaluation contains 219 images and one selected target per image. The two groups are reported separately.

All benchmark images are excluded from reconstruction training and QA SFT. ChartGalaxy real images are never included in training, and the synthetic numerical test images belong to the official evaluation split. The graphical test images selected after feature clustering are likewise held out from training.

A.3 Masking and optimization

For Numerical Reasoning and Numerical Rendering, training randomly masks one to six numerical boxes while retaining at least three numerical boxes as visible context. Contextual Synthesis randomly masks one to four graphical boxes. If their union exceeds 25% of the image area, selected boxes are dropped until the limit is satisfied, except that at least one target is retained even if that target alone exceeds the limit. For the two image-output tasks, the existing chart title is randomly included as textual input. The target image is the original, unmasked infographic.

The three tasks are sampled at a 1:1:1 ratio. Numerical Reasoning contributes the cross-entropy term in Equation 4; Numerical Rendering and Contextual Synthesis contribute the two flow-matching terms. The main BAGEL model is trained for 60k steps on 128 GPUs with AdamW (Loshchilov & Hutter, 2019). The learning rate increases linearly during 500 warm-up steps to 2×10^{-5} and remains constant thereafter. AdamW uses $\beta = (0.9, 0.95)$, $\epsilon = 10^{-15}$, and zero weight decay. Training uses token packing rather than a fixed example batch size: each GPU rank has a soft target of 32,768 tokens and a hard limit of 36,864 tokens. These settings describe the three-task joint-training experiment.

The downstream experiments use separate checkpoints trained through Numerical Reasoning followed by QA SFT. For each backbone, the QA SFT data are identical in the comparator and the CoReL route.

These experiments test transfer from numerical reconstruction independently of joint image-output training.

A.4 Models and naming

Table 6 lists the exact evaluated model variants and their references. Compact names are used where figure or table space is limited. In image comparisons, *Qwen-Edit-2511*, *FireRed-1.1*, and *LongCat* refer to Qwen-Image-Edit-2511, FireRed-Image-Edit-1.1, and LongCat-Image-Edit, respectively. *CoReL* denotes our task-trained BAGEL model.

BAGEL has two expert branches with 7B active and 14B total language-model parameters (Deng et al., 2025). The 7B designation in comparisons refers to active parameters; it is not a claim that the full system contains only 7B parameters. Kimi-VL-A3B similarly names its active language-model parameter count (Moonshot AI, 2025). Released checkpoint names are retained for all models rather than treating their suffixes as a uniform measure of total system size.

GPT-5.5-Pro is queried through the API with the default tool configuration, with tool access left at its defaults. Its result is a commercial capability reference, not a controlled comparison of model size or inference compute. We do not infer its parameter count or attribute a particular answer to a specific tool call. CLI-7B introduces cross-layer vision-language fusion, and Granite-4.0-3B-Vision targets document and chart extraction (Chen et al., 2026; IBM Granite Vision Team, 2026). The remaining numerical baselines are general-purpose VLMs. The image-output baselines are evaluated with the supplied restoration instructions, including the assisted input conditions described below.

Table 6: Model variants used in evaluation and data construction.

Model / checkpoint	Reference
BAGEL (7B active / 14B total)	Deng et al. (2025)
GPT-5.5-Pro	OpenAI (2026)
CLI-7B	Chen et al. (2026)
Granite-4.0-3B-Vision	IBM Granite Vision Team (2026)
LLaVA-OneVision-2-8B-Instruct	An et al. (2026)
Kimi-VL-A3B-Thinking-2506	Moonshot AI (2025)
Qwen2.5-VL-7B-Instruct	Bai et al. (2025)
MiniCPM-V-4.5	Yu et al. (2025)
InternVL3.5-8B	Wang et al. (2025)
Qwen3-VL-8B-Instruct	Qwen Team (2025a)
Ovis2.5-9B	Lu et al. (2025)
GLM-4.1V-9B-Thinking	GLM-V Team (2025)
Qwen3.5-9B	Qwen Team (2026a)
Qwen3.6-27B	Qwen Team (2026c)
Qwen-Image-Edit-2511	Qwen Team (2025b)
FireRed-Image-Edit-1.1	FireRed Team (2026)
LongCat-Image-Edit	LongCat Team (2025)
Qwen3.5-4B (downstream backbone)	Qwen Team (2026b)
Qwen3.8-27B (QA generation)	Qwen Team (2026d)
DeepSeek-OCR-2 (text annotations)	Wei et al. (2026)

B Evaluation Definitions

B.1 Numerical parsing and scoring

We extract the leading signed numeral from each predicted string $\hat{y}_i$ and reference string y_i, denoting the valid parsed values by $\hat{v}_i$ and v_i, respectively. Commas are disambiguated as thousands separators or decimal separators; magnitude suffixes are not used to rescale the extracted number. If either string yields no valid numeral, the score is zero. Otherwise Equation 5 computes numerical proximity on a 0–100 scale with $\epsilon = 10^{-8}$. For target set $\mathcal{T}$,

$$\text{Micro} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} s(\hat{v}_i, v_i). \tag{6}$$

Difficulty-level scores restrict this average to the corresponding targets. Exact match refers to equality of parsed numerical values, not character-level identity. For an input j with targets $\mathcal{T}_j$, all-correct accuracy is

$$\text{AllCorrect} = \frac{100}{J} \sum_{j=1}^{J} \mathbf{1}[\forall i \in \mathcal{T}_j,\ \hat{v}_i = v_i \text{ and both values are valid}]\,. \tag{7}$$

This criterion gives no partial credit. In particular, a readable but incorrect number may receive a positive proximity score while failing exact match.

B.2 Nested masks and shared targets

Numerical Reasoning uses one to five masked targets. In multi-mask evaluation, n is a maximum: inputs with fewer available targets mask those available. Consequently, increasing n can change the number and difficulty distribution of scored boxes; aggregate Micro scores alone need not isolate the loss of context.

We construct nested masks so that the two-target condition's targets are retained when constructing the three-, four-, and five-target conditions. Shared-target evaluation scores the same common targets from the two-target condition under each masked input. All-correct accuracy instead requires every target in that input to be exactly recovered. Table 12 reports both views.

B.3 Numerical image readout

The image-output model generates a completed infographic. We crop the known target boxes and pass each crop to Qwen3-VL-8B-Instruct (Qwen Team, 2025a); the reader does not receive the full infographic. Its numerical output is parsed and scored as above. We report the continuous score, exact-match percentage, and *has number*, the percentage of crops yielding any valid numeral. Has number measures readability, not correctness.

On the original target crops the reader attains a mean score of 86.55. The normalized score is $100 \times \text{Score}\,/86.55$. This empirical reader reference contextualizes the readout error; it is not a strict upper bound for every possible generated font or image. Original and generated crops can induce different reader errors, so normalization does not remove all measurement uncertainty.

B.4 Visual metrics and evaluated regions

For Numerical Rendering, PSNR, SSIM, and LPIPS compare the generated and reference images *outside* the union of the requested target boxes. They measure preservation of unmasked content. Target-crop numerical readout separately evaluates the edited content.

For Structural Targets (the four chart-element classes), all reported visual metrics are computed on the completed *whole image*. For Semantic Targets (HROs), we report both whole-image and target-crop metrics. Whole-image similarity includes unmasked content and therefore reflects overall preservation as well as reconstruction. Crop metrics focus on the generated region. Neither setting by itself establishes semantic appropriateness.

MAE is mean absolute pixel error on the 0–255 scale and uses no learned model. PSNR measures pixel fidelity in decibels. SSIM measures local structural similarity (Wang et al., 2004), using the windowed `scikit-image` implementation with the window size clamped to the available crop dimensions (van der Walt et al., 2014). LPIPS uses an AlexNet backbone (Zhang et al., 2018); DISTS measures deep structure and texture similarity using the `piq` implementation (Ding et al., 2022; Kastryulin et al., 2022). CLIP-I is cosine similarity between CLIP image embeddings (Radford et al., 2021); DINO is cosine similarity between DINOv2 embeddings (Oquab et al., 2024). Lower MAE, LPIPS, and DISTS are better; higher PSNR, SSIM, CLIP-I, and DINO are better.

For Semantic Targets, FID compares feature distributions, and CLIP-FID applies a Fréchet distributional comparison in CLIP feature space (Heusel et al., 2017; Radford et al., 2021). Lower values are better. CLIP-FID is distinct from CLIP-I, which is a paired cosine similarity. On a small set of 219 HRO targets, distributional metrics are interpreted alongside paired perceptual and pixel measures.

C Additional Qualitative Results

Figures 6–9 show additional examples for the three tasks. Reference images and model outputs are preserved, with target regions identified.

Numerical Reasoning

Difficulty ■ E: Easy ■ M: Medium ■ H: Hard 1–3 identify targets; GT = reference Green bold: exact GT match

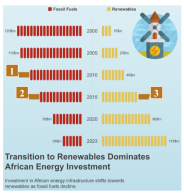
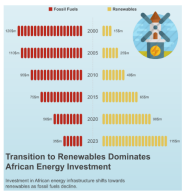
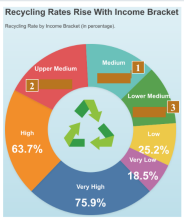
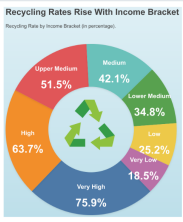
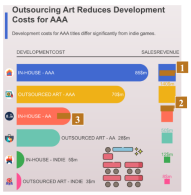
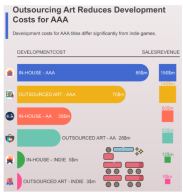
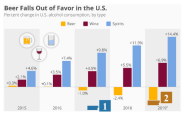
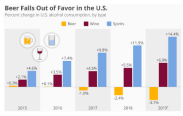
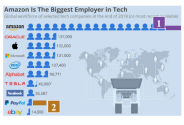
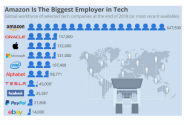
Masked input	Target	CoReL	Qwen3.5-9B	Qwen3.6-27B	GPT-5.5-Pro	GT	Reference
	1 M	**95**	85	100	90	**95**	
	2 M	**75**	70	70	50	**75**	
	3 M	**65**	**65**	**65**	100	**65**	
	1 M	45.1%	45.8%	45.3%	**42.1%**	**42.1%**	
	2 M	58.9%	38.2%	52.1%	51.8%	**51.5%**	
	3 M	**34.8%**	32.5%	38.6%	**34.8%**	**34.8%**	
	1 M	**150**	120	140	**150**	**150**	
	2 M	65	**60**	50	**60**	**60**	
	3 M	**35**	**35**	12	**35**	**35**	
	1 M	18	12	14	**16**	**16**	
	2 M	2	6	4	**3**	**3**	
	1 E	**2017**	-1.5%	-2.4%	**2017**	**2017**	
	2 M	-3.5%	-1.0%	-1.0%	-3.6%	**-3.7%**	
	1 H	**98%**	41%	**98%**	**98%**	**98%**	
	1 H	368,000	1,500,000	200,000	**647,500**	**647,500**	
	2 M	26,000	80,000	20,000	**21,800**	**21,800**	

Figure 6: Numerical Reasoning examples. Targets span Easy, Medium, and Hard. Numbers link masked regions to prediction rows; green bold values exactly match the reference. CoReL can recover multiple missing values, while precise estimation remains difficult for some encodings.

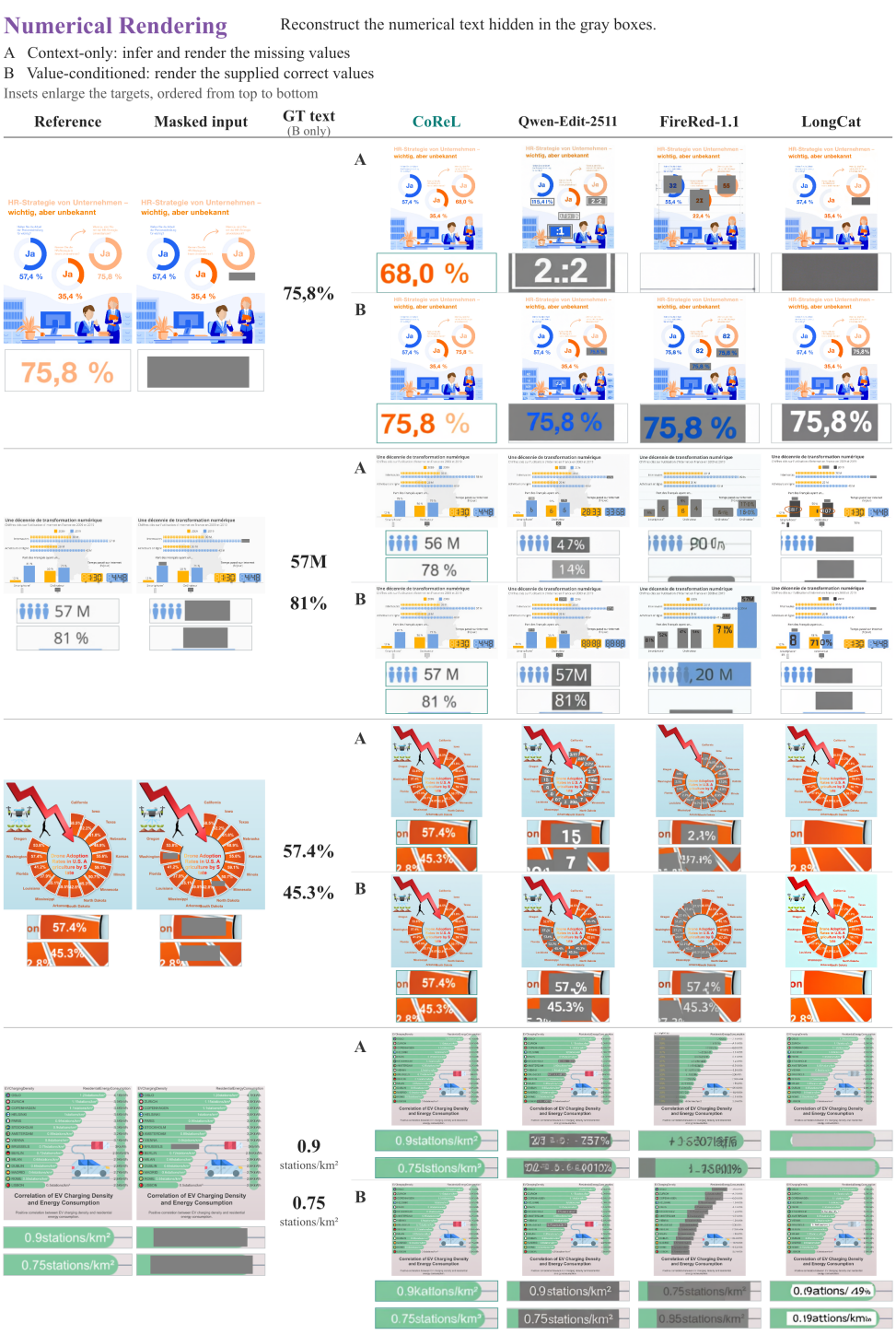


Figure 7: Numerical Rendering examples. Context-only (A) requires inferring and rendering the missing values; value-conditioned (B) supplies the correct values. Insets enlarge target regions. CoReL better preserves numerical labels and surrounding content, while baselines can misrender supplied values or alter unmasked regions.

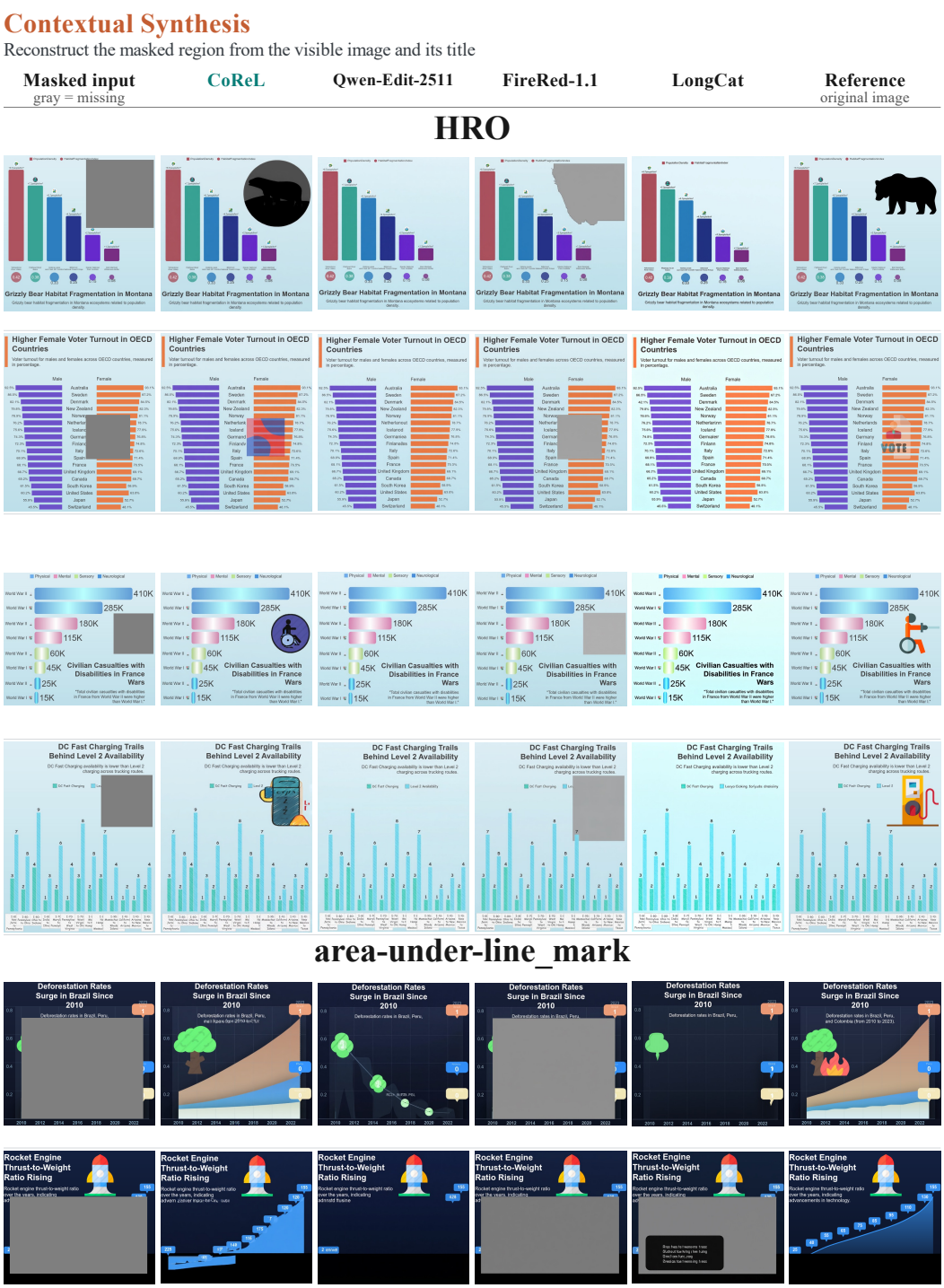


Figure 8: Contextual Synthesis examples. Models complete the masked graphical region using the visible infographic and its title. CoReL restores marks or pictorial content compatible with the context; baselines can leave the region incomplete or introduce structural changes. A suitable pictorial completion need not reproduce the reference exactly.

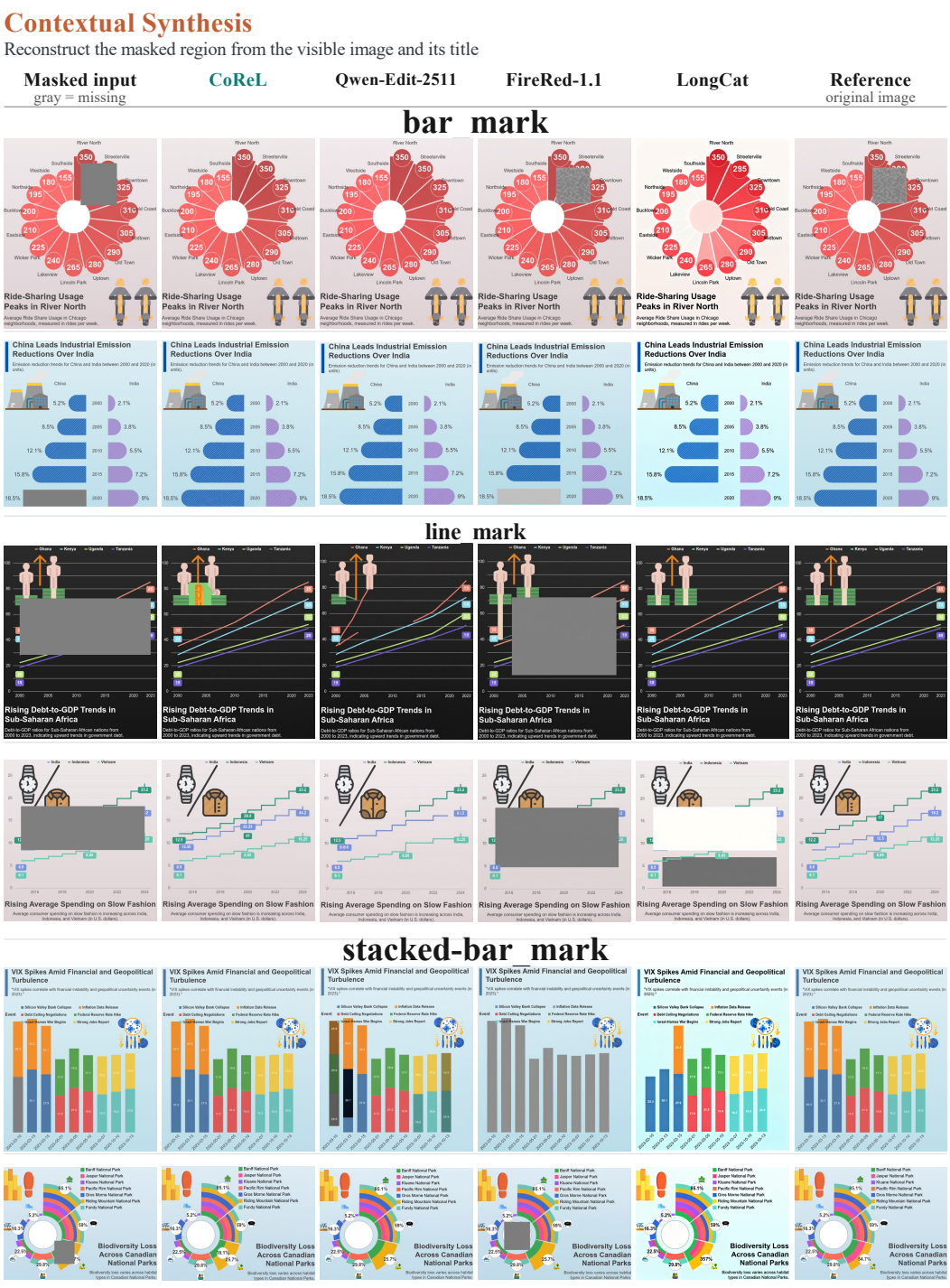


Figure 9: Contextual Synthesis examples (continued). Structural Targets span bars, lines, and stacked bars. All models complete the masked region using the visible infographic and its title.

D Detailed Numerical Reasoning Results

Tables 7–11 report Micro and difficulty-level scores at each mask limit. Our results use the BAGEL checkpoint after 60k steps of joint pretraining. The commercial reference is separated from the evaluated open models. CoReL achieves the best open-model score in every reported difficulty and mask condition.

Table 7: Numerical Reasoning with one masked target per input. Bold denotes the best open-model score; the commercial reference is separated.

Model	Micro	Easy	Medium	Hard
GPT-5.5-Pro	97.03	98.33	97.30	87.55
CLI-7B	45.00	30.58	54.52	31.13
Granite-4.0-3B-Vision	46.48	31.17	56.36	33.79
LLaVA-OneVision-2-8B	45.74	29.23	56.17	34.32
Kimi-VL-A3B-2506	46.46	31.44	56.47	31.20
Qwen2.5-VL-7B	46.85	30.72	56.69	38.97
MiniCPM-V-4.5	48.12	33.33	57.76	35.13
InternVL3.5-8B	50.13	36.28	58.91	40.17
Qwen3-VL-8B	50.99	32.20	62.63	40.08
Ovis2.5-9B	55.88	35.06	68.77	43.85
GLM-4.1V-9B	56.69	34.28	70.47	44.71
Qwen3.5-9B	64.50	58.07	70.28	44.07
Qwen3.6-27B	82.05	87.71	81.77	55.22
CoReL (ours)	**85.19**	**92.71**	**84.10**	**56.22**

Table 8: Numerical Reasoning with up to 2 masked targets per input. Bold denotes the best open-model score; the commercial reference is separated.

Model	Micro	Easy	Medium	Hard
GPT-5.5-Pro	95.95	96.48	95.98	88.09
CLI-7B	44.29	34.99	49.84	31.51
Granite-4.0-3B-Vision	44.44	33.70	50.94	28.86
LLaVA-OneVision-2-8B	45.67	35.41	51.54	34.06
Kimi-VL-A3B-2506	39.05	31.73	43.71	26.12
Qwen2.5-VL-7B	48.24	37.21	54.43	36.85
MiniCPM-V-4.5	47.29	39.78	52.30	31.97
InternVL3.5-8B	48.15	40.46	53.10	34.13
Qwen3-VL-8B	51.89	39.86	59.24	33.73
Ovis2.5-9B	55.01	43.41	61.97	38.73
GLM-4.1V-9B	47.81	40.17	52.58	35.25
Qwen3.5-9B	59.41	63.79	60.22	32.57
Qwen3.6-27B	71.70	78.09	72.05	40.55
CoReL (ours)	**84.31**	**90.11**	**84.42**	**58.03**

Table 9: Numerical Reasoning with up to 3 masked targets per input. Bold denotes the best open-model score; the commercial reference is separated.

Model	Micro	Easy	Medium	Hard
GPT-5.5-Pro	95.21	96.17	95.34	88.49
CLI-7B	42.52	32.48	48.30	31.69
Granite-4.0-3B-Vision	43.14	32.57	48.82	35.72
LLaVA-OneVision-2-8B	43.29	34.14	48.56	33.24
Kimi-VL-A3B-2506	32.48	29.05	35.42	19.39
Qwen2.5-VL-7B	44.84	34.69	50.56	35.11
MiniCPM-V-4.5	45.89	38.49	50.90	30.55
InternVL3.5-8B	44.14	35.24	49.51	32.08
Qwen3-VL-8B	48.56	37.13	55.31	34.48
Ovis2.5-9B	50.85	39.40	57.77	35.21
GLM-4.1V-9B	43.54	36.58	47.93	32.30
Qwen3.5-9B	55.84	58.11	57.16	32.65
Qwen3.6-27B	66.91	72.70	66.72	42.50
CoReL (ours)	**83.90**	**89.94**	**84.08**	**54.58**

Table 10: Numerical Reasoning with up to 4 masked targets per input. Bold denotes the best open-model score; the commercial reference is separated.

Model	Micro	Easy	Medium	Hard
GPT-5.5-Pro	94.42	95.47	94.54	85.73
CLI-7B	40.94	30.80	46.96	32.07
Granite-4.0-3B-Vision	40.92	29.46	47.51	33.02
LLaVA-OneVision-2-8B	41.46	31.29	47.46	32.92
Kimi-VL-A3B-2506	26.71	22.43	29.51	20.43
Qwen2.5-VL-7B	42.73	31.37	49.29	34.58
MiniCPM-V-4.5	44.04	35.20	49.90	30.48
InternVL3.5-8B	42.48	32.87	48.26	33.41
Qwen3-VL-8B	46.59	34.17	54.19	33.68
Ovis2.5-9B	48.11	35.45	55.69	36.53
GLM-4.1V-9B	39.04	32.35	43.00	33.23
Qwen3.5-9B	53.14	53.42	54.88	35.01
Qwen3.6-27B	62.86	67.30	62.68	43.09
CoReL (ours)	**83.81**	**90.11**	**83.72**	**54.07**

Table 11: Numerical Reasoning with up to 5 masked targets per input. Bold denotes the best open-model score; the commercial reference is separated.

Model	Micro	Easy	Medium	Hard
GPT-5.5-Pro	94.29	95.63	94.26	85.27
CLI-7B	39.65	28.41	46.52	30.79
Granite-4.0-3B-Vision	39.24	27.06	46.60	30.53
LLaVA-OneVision-2-8B	40.16	28.17	47.44	31.22
Kimi-VL-A3B-2506	24.19	20.66	27.06	14.51
Qwen2.5-VL-7B	41.22	28.51	48.96	31.56
MiniCPM-V-4.5	42.59	32.22	49.66	27.31
InternVL3.5-8B	41.06	29.81	47.96	31.99
Qwen3-VL-8B	44.78	31.23	53.44	30.49
Ovis2.5-9B	46.10	32.90	54.35	33.97
GLM-4.1V-9B	35.92	27.86	40.78	30.33
Qwen3.5-9B	51.13	49.44	53.63	35.44
Qwen3.6-27B	61.20	65.42	61.30	38.39
CoReL (ours)	**83.55**	**89.81**	**83.38**	**52.63**

Table 12: Nested-mask evaluation. Shared-target scores hold the scored targets fixed while removing more context. All-correct accuracy requires exact numerical recovery of every target in an input. Bold: best open-model result.

Model	Shared-target score ↑				All-correct accuracy (%) ↑			
	2	3	4	5	2	3	4	5
GPT-5.5-Pro	95.95	95.20	94.79	94.75	64.24	58.01	54.77	52.06
CLI-7B	44.29	43.94	43.38	43.01	0.52	0.26	0.26	0.26
Granite-4.0-3B-Vision	44.44	44.12	42.78	41.95	0.78	0.78	0.78	0.69
LLaVA-OneVision-2-8B	45.67	44.38	42.96	42.78	1.12	0.86	0.86	0.86
Kimi-VL-A3B-2506	39.05	34.89	31.43	29.58	1.82	0.95	0.95	1.21
Qwen2.5-VL-7B	48.24	45.97	45.28	45.15	0.95	0.61	0.61	0.61
MiniCPM-V-4.5	47.29	46.87	46.18	45.81	2.07	1.38	1.30	1.21
InternVL3.5-8B	48.15	45.57	44.79	44.18	1.99	1.38	1.21	1.38
Qwen3-VL-8B	51.89	50.20	49.26	48.56	2.16	1.64	1.64	1.64
Ovis2.5-9B	55.01	52.32	51.50	50.77	3.37	2.25	1.82	1.99
GLM-4.1V-9B	47.81	45.26	42.40	41.45	4.75	2.77	2.85	2.68
Qwen3.5-9B	59.41	58.18	56.53	56.07	8.82	6.22	5.10	5.19
Qwen3.6-27B	71.70	68.41	66.15	65.38	18.67	13.48	11.58	11.32
CoReL (ours)	**84.31**	**83.39**	**83.15**	**82.48**	**39.84**	**34.31**	**32.24**	**29.30**

Holding the scored targets fixed. Changing the mask limit can change the difficulty distribution of the evaluated targets. We therefore keep the scored targets from the two-mask condition fixed while masking additional regions in the three- to five-mask conditions (Table 12). From two to five masks, CoReL's shared-target score changes from 84.31 to 82.48, while Qwen3.6-27B drops from 71.70 to 65.38. Five-mask all-correct accuracy is also higher, 29.30% versus 11.32%. Thus the advantage persists when target identity is controlled and extends to exact joint recovery.

Difficulty and context. Single-mask improvements over Qwen3.6-27B are 5.00 points on Easy, 2.33 on Medium, and 1.00 on Hard. Hard targets remain challenging despite the overall advantage. The gap widens when multiple labels are removed: at five masks, CoReL scores 83.55 overall versus 61.20 for Qwen3.6-27B. Shared-target scores in Table 12 confirm that this pattern is not solely due to changes in which targets are scored.

Approximate versus exact recovery. CoReL's five-mask Micro score of 83.55 coexists with 29.30% all-correct accuracy. The former gives partial credit for numerically close predictions; the latter requires exact recovery of all requested values. The two views distinguish useful numerical estimation from complete exact reconstruction.

E Detailed Numerical Rendering Results

Tables 13 and 14 report the full numerical benchmark under both input conditions. Context-only inputs require inference and rendering; value-conditioned inputs explicitly bind the correct number to each target box. Each condition is applied to all four models.

Table 13: Numerical Rendering: target-crop readout on 712 images and 5,353 boxes. The reader reference score is 86.55.

Model	Score ↑	Exact (%) ↑	Has number (%) ↑	Normalized ↑
Context-only (infer and render)				
LongCat-Image-Edit	3.65	0.41	13.58	4.22
FireRed-Image-Edit-1.1	22.56	1.61	73.60	26.07
Qwen-Image-Edit-2511	14.63	0.86	74.52	16.90
CoReL (ours)	**75.51**	**55.67**	**97.48**	**87.25**
Value-conditioned (correct target values supplied)				
LongCat-Image-Edit	29.20	24.42	41.75	33.74
FireRed-Image-Edit-1.1	44.64	31.81	73.29	51.58
Qwen-Image-Edit-2511	59.83	51.13	87.75	69.13
CoReL (ours)	**82.78**	**78.85**	**97.85**	**95.65**

Table 14: Numerical Rendering: preservation outside the target boxes. Higher PSNR and SSIM and lower LPIPS are better.

Model	PSNR ↑	SSIM ↑	LPIPS ↓
Context-only (infer and render)			
LongCat-Image-Edit	21.06	0.9051	0.0875
FireRed-Image-Edit-1.1	17.03	0.8208	0.2292
Qwen-Image-Edit-2511	18.06	0.8481	0.1464
CoReL (ours)	**36.42**	**0.9860**	**0.0048**
Value-conditioned (correct target values supplied)			
LongCat-Image-Edit	23.59	0.9265	0.0569
FireRed-Image-Edit-1.1	17.39	0.8547	0.1887
Qwen-Image-Edit-2511	19.91	0.8724	0.1148
CoReL (ours)	**36.97**	**0.9865**	**0.0038**

With correct values supplied, the numerical scores of LongCat, FireRed, and Qwen rise from 3.65/22.56/14.63 to 29.20/44.64/59.83. CoReL rises from 75.51 to 82.78. The improvement isolates a useful change in input information, though it is not a strict decomposition of latent reasoning and generation errors.

Readability is distinct from correctness: Qwen's value-conditioned has-number rate is 87.75%, while its exact-match rate is 51.13%. CoReL reaches 97.85% and 78.85%, respectively. In both conditions it also achieves higher PSNR and SSIM and lower LPIPS outside the target boxes, indicating more faithful preservation under this protocol. Pixel scores remain sensitive to output resizing or resampling, so they are interpreted with the perceptual metric and qualitative examples.

F Detailed Contextual Synthesis Results

F.1 Structural Targets

Table 15 reports all seven metrics for the 200 Structural Targets. Every model receives the masked image and the image's existing title; all metrics are whole-image measures. CoReL leads across pixel fidelity, perceptual distance, and image-embedding similarity. Because the metric region includes the rest of the image, these results measure completion together with preservation.

F.2 Semantic Targets and text conditioning

The Semantic evaluation contains one HRO target from each of 219 images. The primary *title* condition uses the existing title together with the masked image. In the *image-caption* condition, a VLM describes the complete reference image. In the *object-caption* condition, it describes the held-out target region. These captions can reveal target identity and appearance, providing stronger

Table 15: Contextual Synthesis of Structural Targets across four chart-element classes. Fifty target boxes per class (200 total) are reconstructed with title conditioning. All metrics are computed on the completed whole image.

Model	MAE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	CLIP-I ↑	DINO ↑
Qwen-Image-Edit-2511	43.46	15.53	0.7078	0.4199	0.2848	0.8318	0.5876
FireRed-Image-Edit-1.1	64.73	11.57	0.6413	0.6364	0.4087	0.7746	0.4251
LongCat-Image-Edit	65.68	11.33	0.6182	0.6224	0.3949	0.7614	0.4195
CoReL (ours)	**20.26**	**20.79**	**0.7947**	**0.2390**	**0.1863**	**0.8624**	**0.6588**

assistance than a title. They remain image-conditioned editing tasks, but reduce the need to infer the missing content from context.

CoReL is evaluated only with the title. The baselines are evaluated with all three text inputs. Tables 16 and 17 therefore show our title-conditioned result once, rather than repeating it as if it came from three different runs.

Table 16: Contextual Synthesis of Semantic Targets (HROs): whole-image metrics on 219 targets. Our model uses only the title; captions provide additional target information to the baselines. Bold denotes the best result under matched title conditioning.

Model	Text input	FID ↓	CLIP-FID ↓	PSNR ↑	SSIM ↑	DISTS ↓
Qwen-Image-Edit-2511	Title	27.71	8.85	16.44	0.8563	0.1622
FireRed-Image-Edit-1.1	Title	36.48	11.80	12.39	0.7359	0.2125
LongCat-Image-Edit	Title	26.31	8.25	17.91	0.8434	0.1699
CoReL (ours)	Title	**21.00**	**4.89**	**20.12**	**0.9262**	**0.1310**
Qwen-Image-Edit-2511	Image caption	33.42	8.67	16.00	0.8300	0.1897
FireRed-Image-Edit-1.1	Image caption	43.82	11.08	12.17	0.7069	0.2484
LongCat-Image-Edit	Image caption	27.50	7.83	17.89	0.8615	0.1733
Qwen-Image-Edit-2511	Object caption	30.89	8.25	16.90	0.8643	0.1729
FireRed-Image-Edit-1.1	Object caption	39.11	10.45	13.09	0.7530	0.2206
LongCat-Image-Edit	Object caption	36.63	8.66	15.54	0.8260	0.2102

Table 17: Contextual Synthesis of Semantic Targets (HROs): target-crop metrics on 219 targets. Our model uses only the title; captions provide additional target information to the baselines. Bold denotes the best result under matched title conditioning.

Model	Text input	FID ↓	CLIP-FID ↓	PSNR ↑	SSIM ↑	DISTS ↓
Qwen-Image-Edit-2511	Title	236.25	48.92	7.57	0.3114	0.5662
FireRed-Image-Edit-1.1	Title	239.32	52.08	4.27	0.1063	0.5876
LongCat-Image-Edit	Title	235.33	48.86	**10.48**	**0.5420**	0.5382
CoReL (ours)	Title	**118.70**	**19.86**	9.30	0.3469	**0.3638**
Qwen-Image-Edit-2511	Image caption	188.52	36.35	8.63	0.3855	0.5042
FireRed-Image-Edit-1.1	Image caption	207.75	41.35	5.84	0.1871	0.5432
LongCat-Image-Edit	Image caption	229.50	44.81	9.74	0.4912	0.4947
Qwen-Image-Edit-2511	Object caption	140.74	20.97	8.30	0.3087	0.3742
FireRed-Image-Edit-1.1	Object caption	154.31	23.32	6.59	0.2091	0.4164
LongCat-Image-Edit	Object caption	136.31	21.07	7.71	0.3240	0.3761

Object captions lower all three baselines' crop FID relative to title conditioning. Their best object-caption crop FID is 136.31, compared with CoReL's 118.70 using only the title. This demonstrates a remaining gap even after the target is described. Assistance does not improve every metric: whole-image FID and crop pixel similarity can follow different rankings.

Not every metric on Semantic Targets favors CoReL: LongCat has higher crop PSNR and SSIM under title conditioning, whereas CoReL has lower FID, CLIP-FID, and DISTS (Table 17). Since multiple

HRO completions can fit a context, reference similarity must be read together with perceptual and distributional measures.

G Downstream Data and Task-Level Analysis

G.1 QA data and controlled comparison

The training QA annotations described in ChartGalaxy were not publicly released. We therefore generate our own QA SFT annotations for ChartGalaxy training images using Qwen3.8-27B (Qwen Team, 2026d) and the original QA-generation prompts (Li et al., 2026b). DINOv2 features are used to reduce visual redundancy before QA generation (Oquab et al., 2024). The questions cover Data Identification, Data Comparison, Fact Checking, and Data Extraction with Condition. The last category includes trend/pattern, calculation, and hypothetical questions, yielding six question types in total. The actual SFT runs use 140k images; each run trains for 15k steps, processing 1.28M QA examples.

For both BAGEL and Qwen3.5-4B, we compare the original backbone, QA SFT alone, and Numerical Reasoning pretraining followed by QA SFT. Within each backbone, the two trained routes use exactly the same QA SFT data and the same 15k-step SFT duration. Numerical Reasoning is the only reconstruction task used in these transfer runs; the Qwen backbone does not generate images. Downstream evaluation focuses on QA categories from ChartGalaxy and EncQA related to Numerical Reasoning pretraining (Table 5). EncQA evaluates transfer to a different chart domain: neither CoReL pretraining nor QA SFT uses data from that domain, and we perform no EncQA-specific tuning.

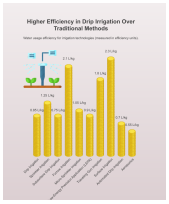
G.2 What transfers across tasks and backbones

Across both backbones, gains occur on tasks that select elements or relate values across a chart. Counting improves by 9.42/10.14 points on BAGEL/Qwen, and EncQA Filter Values improves by 16.23/5.55 points. Correlate Values and Correlate Values Relative also improve on both backbones (Table 5), supporting transfer beyond masked-label prediction.

Improvements are not uniform: BAGEL's comparison score remains 84.09, and trend analysis decreases from 95.69 to 91.38. On EncQA, QA SFT lowers BAGEL from 73.33 to 67.04; CoReL raises it to 76.52. Its 9.48-point gain over QA SFT therefore includes recovery from that degradation, with a 3.19-point gain over the original backbone. Qwen improves over both baselines.

ChartGalaxy: Counting

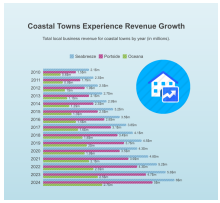
(a) C1: Count vs. threshold



Q: In how many Technology is the Water Usage Efficiency greater than 1.13 units?

Answer: 4

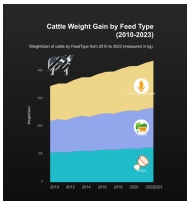
(b) C2: Count vs. average



Q: In how many Year is the Total of Seabreeze greater than the average Total of Seabreeze across all Year?

Answer: 7

(c) C3: Visual element count

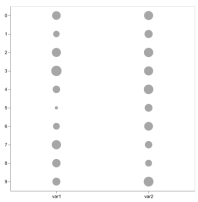


Q: How many icons and images are present in this infographic chart?

Answer: 4

EncQA: Out-of-domain chart understanding

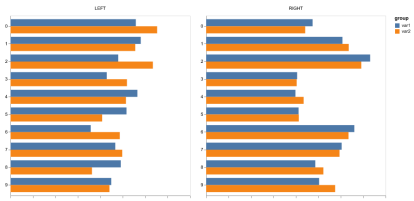
(d) Correlate values



Q: Are the areas of the circles for var1 and var2 correlated?

Answer: no

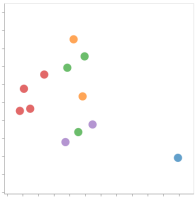
(e) Correlate values (relative)



Q: In which chart are the lengths of the bars for var1 and var2 more correlated?

Answer: RIGHT

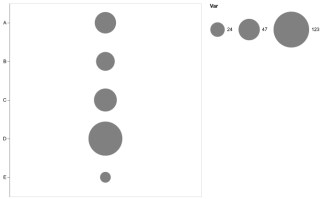
(f) Find extrema



Q: which color has the least observations?

Answer: Blue

(g) Filter values (size)



Q: Which circle(s) have a value of Var less than 26?

Answer: ['E']

(h) Filter values (color)



Q: Which circle(s) have a value of Var less than 34?

Answer: ['C']

Figure 10: Downstream chart QA examples. ChartGalaxy (a–c) illustrates counting relative to a threshold, counting relative to an average, and counting visual elements. EncQA (d–h) covers Correlate Values, Correlate Values Relative, Find Extrema, and Filter Values; the last category is illustrated with both size and color encodings. Each example includes the chart, question, and answer.

H LIMITATIONS

Attribution and comparison. Direct reconstruction results compare a task-adapted model with released baselines. The transfer experiment controls the QA SFT data, but adds a pretraining stage and therefore does not equalize total training compute. It establishes the benefit of Numerical Reasoning pretraining under the matched-data protocol; it does not isolate the incremental QA benefit of the image-output tasks or prove synergy among all three objectives. The commercial reference also differs in model and inference resources.

Available evidence and ambiguity. Some numerical targets admit textual shortcuts or redundant constraints; the difficulty levels expose this variation, but scores do not identify internal reasoning. Graphical targets can admit multiple appropriate completions. Future evaluation could directly assess contextual compatibility alongside reference similarity.

I EVALUATION PROMPT TEMPLATES

The templates below specify the restoration instruction. Variables are filled per example; they are not fixed content shown in every evaluation. Numerical Reasoning uses red masks, while the image-output tasks use gray masks. Target values and coordinates are supplied only in the corresponding assisted conditions.

I.1 NUMERICAL REASONING

Here `{n}` is the number of masked boxes and `{sep}` is the output separator, `|`.

> You are looking at a chart image. {n} numeric/text value(s) have EACH been hidden behind a solid red rectangle. Using only the surrounding context of the chart (axis labels, other data values, legend, trend), predict the exact value that was originally written inside EACH red rectangle.
>
> There are {n} red rectangles. Report their values in reading order: top-to-bottom, then left-to-right.
>
> Rules:
>
> - Reply with ONLY the {n} predicted value(s), in reading order, separated by '{sep}'. Example for 3 boxes: 12% | 3.4 | -5
> - Keep symbols like %, $, +, -, k, M, commas, decimals exactly as they would appear on the chart.
> - Do NOT include explanations, labels, units not in the chart, or any extra words.
> - Output EXACTLY {n} value(s) separated by '{sep}', in the given order. If uncertain about any, still give your single best guess for it.

I.2 NUMERICAL RENDERING

Context-only.

> This image is a chart in which {n} region(s) have been covered by solid {color} boxes. Each covered region originally contained a number—a data value, axis tick, percentage, or similar numeric label. Do not change anything outside the {color} boxes. Using the full context of the chart (its axes, scale, gridlines, the other visible numbers, and the relative size or position of the bars, lines, points or slices), infer the single most plausible number that belongs in each {color} box and paint that number back into the box in a font, size and color that match the surrounding labels, so the chart reads as a clean, complete image with no {color} boxes left.

Value-conditioned.

> This image is a chart in which {n} region(s) have been covered by solid {color} boxes. Each covered region originally contained a number—a data value, axis tick, percentage, or similar numeric label. {numbers} Do not change anything outside the {color} boxes. Paint exactly that given number back into each {color} box, matching the font, size and color of the surrounding labels, so the chart reads as a clean, complete image with no {color} boxes left.

Here `{numbers}` lists the correct values in reading order, each paired with a coarse region name and the box center as percentages of image width and height.

I.3 CONTEXTUAL SYNTHESIS

The masked image is paired with its title in the primary condition. In the diagnostic conditions for Semantic Targets, the accompanying text is an image caption or target-object caption. The restoration instruction is:

> This image has {n} region(s) covered by a solid {color} box. Reconstruct the original visual content that belongs in each {color} box—using the surrounding context of the image (its layout, colors, style, and the neighboring content)—and paint it back so the picture reads as a clean, complete, natural image with no {color} boxes left. Do not change anything outside the {color} boxes.