VIZANCHOR: DECODING MANIPULATION INTENT FROM TAMPERING VISUALIZATIONS VIA DUAL-ANCHOR REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data visualizations are widely used for communicating information, but they are also vulnerable to intentional manipulations that induce misleading interpretations. Existing methods focus on locating tampered regions or recovering hidden information, without explaining how the visualization has been manipulated or why the resulting changes may mislead viewers. We propose **VizAnchor**, a framework for visualization manipulation understanding through dual-anchor evidence construction and VLM-based reasoning. In the first stage, VizAnchor constructs a semantic anchor to recover authentic chart information and a spatial anchor to localize tampered regions. In the second stage, three task-specific VLM reasoning steps progressively ground manipulation attributes, reconstruct original and tampered chart narratives, and infer misleading intent by integrating visual evidence with intermediate reasoning results. We further construct a dataset for tampering localization and a dataset for misleading intent inferring. Evaluation shows that VizAnchor accurately localizes manipulations and produces explanations of their manipulation, misleaders, and misleading intents.

1 INTRODUCTION

Visualizations are widely used to communicate quantitative evidence in scientific articles, financial reports, public dashboards, and news media. Their persuasive power, however, also makes them vulnerable to subtle manipulation. Small edits to data marks, axes, legends, labels, colormaps, or annotations can alter the conclusions drawn by viewers while preserving a visually plausible appearance. Unlike generic image tampering, visualization manipulation is inherently semantic, as its impact depends on how visual edits change data encodings, comparisons, and chart narratives.

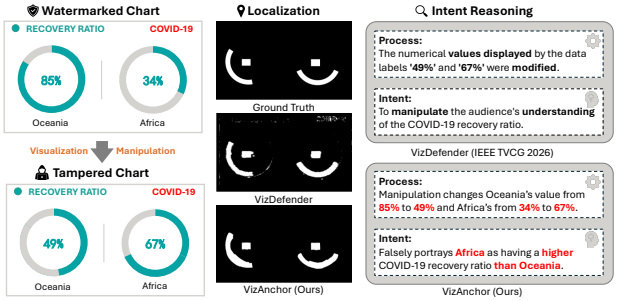

Figure 1: An example of manipulation intent decoding with VizAnchor. The manipulation reverses the original recovery ratio comparison between Oceania and Africa. Guided by semantic and spatial anchors, VLM-based reasoning identifies the specific modifications, reconstructs both chart narratives, and infers the misleading intent.

Existing trustworthy-visualization methods mainly focus on provenance verification and data recovery. Methods such as VisCode, InvVis, and VisGuard embed metadata, source data, or recoverable information into visualization images, enabling users to trace a chart back to its original data or generation context (Zhang et al., 2021; Ye et al., 2024; 2026). However, they are not designed to precisely localize malicious edits or explain their semantic consequences. VizDefender further combines tamper localization with multimodal reasoning (Song et al., 2026), but its analysis relies on the tampered chart and detected suspicious regions. Although such evidence can reveal potential modifications, it provides limited information about how the chart deviates from its original message. Without an untampered visual reference, VLMs may struggle to identify the exact changes and distinguish the intended misleading interpretation from the original narrative.

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.

To address these limitations, as shown in Figure 1, we introduce **VizAnchor**, a two-stage framework for fine-grained visualization manipulation understanding. By comparing the original and tampered charts, VLMs can identify concrete modifications, reconstruct changes in the communicated message, and infer evidence-supported misleading intents. In Stage 1, VizAnchor embeds recoverable provenance metadata and a crop-synchronization cue into an original chart, producing a watermarked chart. Given a potentially cropped and manipulated chart, VizAnchor first decodes its position map and restores it to the canonical canvas, producing aligned tampered chart. Metadata is then recovered from the aligned chart and used to resolve a reference link, through which the original chart and its untampered watermarked counterpart are retrieved. The former serves as the Semantic Anchor, while the latter provides the reference for constructing the Spatial Anchor.

In Stage 2, the dual anchors are organized into a four-panel visual prompt containing Semantic Anchor, tampered chart, a Tamper Map, and a Localized Comparison. Misleader Grounding (MG) predicts the tamper type, tampered components, and manipulation process from this prompt. Chart Narrative Reconstruction (CNR) separately analyzes Semantic Anchor and tampered chart to reconstruct their respective chart narratives. Finally, Intent Inferring (II) integrates the visual prompt and outputs of the preceding steps to infer the underlying misleading intent.

Our contributions are summarized as follows:

- We propose VizAnchor, a framework for visualization manipulation analysis integrating provenance recovery, tampering localization, and misleading-intent decoding.
- To address the lack of authentic references and precise spatial evidence in visualization manipulation analysis, we introduce a dual-anchor evidence construction mechanism that combines a retrieved Semantic Anchor with a pixel-level Spatial Anchor.
- We develop a task-structured VLM reasoning approach that progressively grounds manipulation attributes, reconstructs original and tampered chart narratives, and infers misleading intent from structured visual evidence.
- We construct two datasets for localization training and manipulation-understanding evaluation. Extensive experiments demonstrate reliable watermark recovery, crop and local-edit localization, and reasoning about manipulation processes and misleading intents.

2 Related Work

Misleading Visualization. Prior work has examined how distorted scales, encodings, visual proportions, and rhetorical designs can bias chart interpretation (Pandey et al., 2015; Lo et al., 2022; Lisnic et al., 2023; 2024). Recent benchmarks further evaluate whether MLLMs can detect or explain misleading cues in standalone charts (Chen et al., 2025; Mahbub et al., 2025). Most of these works focus on misleaders introduced during the initial chart-creation process, which may arise from either deliberate design choices or unintentional design flaws.

We regard tampered visualizations as a more specific subset of misleading visualizations. In this setting, an initially valid chart is intentionally modified after creation at the image level to alter its original message. Therefore, the task requires not only recognizing that the resulting chart is misleading, but also identifying what is changed relative to the authentic chart and inferring the intent behind the modification. Existing misleading-chart benchmarks, such as Misviz (Tonglet et al., 2026) and Misleading ChartQA (Chen et al., 2025), provide complementary evaluation of general misleading-chart understanding, but do not directly target post-creation tampering intent. VizAnchor addresses this gap by reasoning from authentic and tampered visualization evidence to localize the edit, explain the manipulation process, and infer its misleading intent.

Visualization Protection. Existing visualization protection methods generally follow two directions. One line embeds chart metadata through robust watermarking to support provenance verification and source-data recovery (Fu et al., 2021; Zhang et al., 2021; Ye et al., 2024). Another line embeds location-aware or semi-fragile signals to detect and localize post-creation tampering (Ye et al., 2026; Song et al., 2026). However, provenance recovery and tamper localization alone do not fully explain how an edit changes the chart message or what misleading intent it serves. VizAnchor bridges these objectives by combining dual-anchor evidence with structured VLM reasoning,

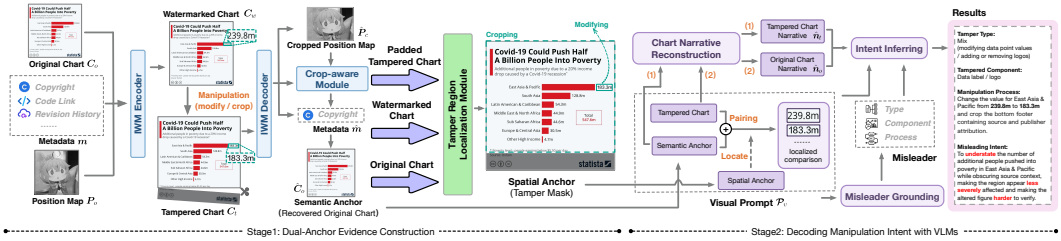

Figure 2: Overview of the proposed VizAnchor framework.

enabling provenance retrieval, tamper localization, and fine-grained manipulation-intent analysis within a unified framework.

Image Tampering Detection. Image tampering detection is commonly divided into passive and active settings. Passive tampering detectors localize edits from forensic traces such as noise residuals, boundary artifacts, and cross-channel inconsistencies (Zhou et al., 2018; Wu et al., 2019; Chen et al., 2021; Liu et al., 2022; Guillaro et al., 2023). These methods are powerful for generic image manipulation, but they are not tailored to visualization semantics, where small edits to axes, legends, labels, or marks may cause large interpretive shifts. Active methods instead embed signals before dissemination (Zhu et al., 2018; Tancik et al., 2020; Zhang et al., 2019; Jia et al., 2021; Kou et al., 2025), with EditGuard and OmniGuard extending this paradigm to proactive tamper localization (Zhang et al., 2024; 2025). While these methods focus on watermark robustness or tampering localization, they do not explicitly address how manipulations alter the semantic narratives conveyed by data visualizations. In contrast, VizAnchor focuses on visualization-specific manipulation understanding. Rather than treating watermark recovery and localization as the final objectives, we construct complementary semantic and spatial anchors to support VLM-based reasoning. By explicitly comparing the original and tampered chart narratives and grounding their differences in localized visual evidence, VizAnchor identifies concrete manipulations and infers misleading intents.

3 METHOD

VizAnchor moves beyond tamper localization to understand how a chart is manipulated and why it becomes misleading. As shown in Figure 2, it consists of two stages: Dual-Anchor Evidence Construction and Decoding Manipulation Intent with VLMs.

3.1 SEMANTIC ANCHOR CONSTRUCTION

The Semantic Anchor is the recovered original chart $\hat{C}_o$, which provides the semantic reference for comparison with the aligned tampered chart $\widetilde{C}_t$. To bind this reference to the distributed image, VizAnchor performs crop-robust metadata hiding and produces the watermarked chart C_w.

Metadata Embedding and Recovery. As illustrated in Figure 3, IWM is an INN-based reversible network built upon affine coupling transformations (Dinh et al., 2017). Following the attention flow-based message embedding design of RMSteg (Ye et al., 2025), IWM sequentially embeds metadata and position information into an original chart. Its forward and inverse paths are referred to as the IWM Encoder and IWM Decoder, respectively.

Given an original chart C_o and a K-bit metadata payload m, the payload is reshaped into a binary map and spatially tiled to introduce redundancy. The chart and tiled metadata map are separately tokenized using ViT-based encoders (Dosovitskiy et al., 2021). Unlike the chart features, the metadata features are additionally processed by an invertible token shuffle, which disperses adjacent metadata tokens across different spatial positions and improves robustness against localized corruption. The two feature streams are then concatenated and mixed through ActNorm and an invertible 1×1 convolution before entering a sequence of Transflow Blocks. After the final block, the chart feature stream is projected back to image space to produce the intermediate watermarked chart C'_w.

Let x_t^i and x_o^i denote the shuffled tiled-metadata features and original-chart features entering the i-th Transflow Block, respectively. Its forward transformation is

$$\begin{aligned} x_o^{i+1} &= x_o^i + \phi_{\mathrm{T}}(x_t^i), \\ x_t^{i+1} &= x_t^i \odot \exp\big(\rho_{\mathrm{T}}(x_o^{i+1})\big) + \eta_{\mathrm{T}}(x_o^{i+1}), \end{aligned} \tag{1}$$

where ϕ_{T}, ρ_{T}, and η_{T} are learnable transformations. The corresponding inverse transformation is

$$\begin{aligned} \hat{x}_t^i &= \big[\hat{x}_t^{i+1} - \eta_{\mathrm{T}}(\hat{x}_o^{i+1})\big] \oslash \exp\big(\rho_{\mathrm{T}}(\hat{x}_o^{i+1})\big), \\ \hat{x}_o^i &= \hat{x}_o^{i+1} - \phi_{\mathrm{T}}(\hat{x}_t^i), \end{aligned} \tag{2}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively.

Once the aligned tampered chart $\widetilde{C}_t$ is obtained by the position-recovery stage described below, it is directly fed into the metadata-decoding branch. The inverse Transflow Blocks recover the shuffled metadata features, after which the inverse token shuffle restores their original spatial ordering. The reordered features are decoded into the tiled metadata map, and the redundant predictions are aggregated to obtain the recovered payload $\hat{m}$.

Position Embedding and Recovery. The position-embedding stage embeds the full position map P_o into the intermediate watermarked chart C'_w. Their features are concatenated and processed by a sequence of Position Flow Blocks, producing the final watermarked chart C_w. Let z_w^j and z_p^j denote the features of C'_w and P_o entering the j-th Position Flow Block.

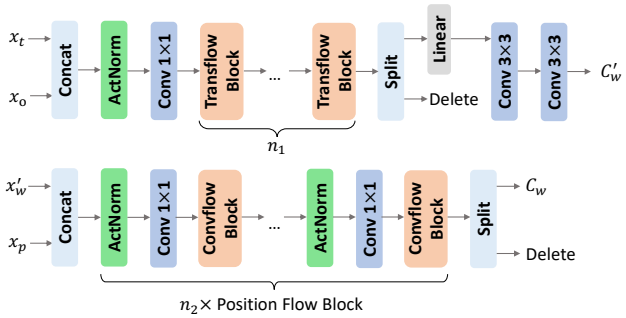

Figure 3: Architecture of the Invertible Watermarking Module. The IWM Encoder embeds metadata into the chart through Transflow Blocks to produce the watermarked chart C'_w, and embeds the position map through Position Flow Blocks to obtain C_w. The IWM Decoder reverses these stages for position and metadata recovery.

Within each Position Flow Block, the Convflow Block follows the forward and inverse affine-coupling formulations of the Transflow Block in Eqs. (1) and (2). Its learnable transformations, ϕ_{P}, ρ_{P}, and η_{P}, are parameterized independently from those used for metadata embedding. This separation enables metadata and positional information to be modeled by distinct transformations.

During recovery, given a potentially cropped and manipulated chart C_t, a Nested U-Net first enhances its features, and the Crop-Aware Module produces the aligned tampered chart $\widetilde{C}_t$. The aligned chart is then passed to the metadata-decoding branch to recover $\hat{m}$.

Crop-Aware Module. Given a potentially cropped and manipulated chart C_t, the IWM first recovers a cropped position map $\hat{P}_c$. The Crop-Aware Module matches $\hat{P}_c$ against the full position map P_o to estimate the crop location and scale. The estimated transformation is then used to restore C_t to the canonical canvas, producing the aligned tampered chart $\widetilde{C}_t$. Meanwhile, the missing region outside the estimated crop boundary is converted into a crop mask $\hat{M}_{\mathrm{crop}}$. The aligned chart is subsequently passed back to the inverse path of IWM for metadata recovery, while $\hat{M}_{\mathrm{crop}}$ is retained as part of the spatial evidence.

Semantic-Anchor Retrieval. Before dissemination, an 81-bit identifier m is assigned to each original chart and embedded into the image. The identifier encodes a compact reference link under a predefined URL prefix. After crop-aware alignment, the recovered metadata $\hat{m}$ is decoded to resolve the reference link, through which the original chart $\hat{C}_o$ and untampered watermarked chart $\hat{C}_w$ are retrieved.

3.2 Spatial Anchor Construction

The Spatial Anchor integrates two complementary localization cues: the crop mask $\hat{M}_{\text{crop}}$ produced by the Crop-Aware Module and the local-edit mask $\hat{M}_{\text{edit}}$ predicted by the Localization Module. Both masks are represented on the canonical canvas and combined by pixel-wise union:

$$\hat{M} = \hat{M}_{\text{crop}} \vee \hat{M}_{\text{edit}}, \tag{3}$$

where $\vee$ denotes the pixel-wise logical OR operation. The resulting mask $\hat{M}$ serves as the Spatial Anchor, capturing both geometric cropping and fine-grained content modifications.

Tamper Localization. The recovered position map $\hat{P}_c$ is matched against the full position map P_o to estimate the crop region and align the observed chart to the canonical canvas. The estimated cropped-away region is represented as the binary crop mask $\hat{M}_{\text{crop}}$. After crop-aware alignment and reference retrieval, the Localization Module compares the untampered watermarked chart $\hat{C}_w$ with the aligned tampered chart $\widetilde{C}_t$. Its input is a seven-channel tensor consisting of the two RGB charts and their binary difference map:

$$\begin{aligned} D_{\text{bin}} &= \mathbb{I}\left(\max_{c \in \{R,G,B\}} \left| \hat{C}_w^{(c)} - \widetilde{C}_t^{(c)} \right| > \tau_{\text{diff}} \right), \\ X &= \text{Concat}\left(\hat{C}_w, \widetilde{C}_t, D_{\text{bin}} \right). \end{aligned} \tag{4}$$

The Localization Module adopts a U-Net-based architecture to jointly model the paired RGB content and changed-pixel cues (Ronneberger et al., 2015). The local-edit mask is obtained as

$$\hat{M}_{\text{edit}} = \mathbb{I}\left(F_\theta(X) > \tau_{\text{mask}} \right), \tag{5}$$

where $F_\theta(X)$ denotes the predicted local-edit probability map.

3.3 Training Objectives

The IWM and Crop-Aware Module are jointly trained on the VisGuard dataset, while the Localization Module is trained separately on VAD-LocTrain, the localization-training subset of VizAnchor Dataset (VAD), which contains 1,500 automatically generated chart pairs with pixel-level local-edit masks. The joint objective for IWM and crop-aware recovery is

$$\begin{aligned} \mathcal{L}_{\text{IWM}} = \lambda_{\text{steg}} \mathcal{L}_{\text{steg}} + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}} + \lambda_{\text{meta}} \mathcal{L}_{\text{meta}} \\ + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}} + \lambda_{\text{pf}} \mathcal{L}_{\text{pf}}, \end{aligned} \tag{6}$$

where $\mathcal{L}_{\text{steg}}$ is the L1 loss between the original and watermarked charts, $\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}$, $\mathcal{L}_{\text{meta}}$ is the BCEWithLogits loss for metadata recovery, $\mathcal{L}_{\text{lpips}}$ is the VGG-based LPIPS perceptual loss, and $\mathcal{L}_{\text{pf}}$ is the L1 loss between the predicted and target position maps. The Localization Module is optimized using

$$\mathcal{L}_{\text{edit}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{7}$$

The loss-balancing coefficients and implementation details are reported in the appendix.

3.4 Decoding Manipulation Intent with VLMs

After constructing the Semantic and Spatial Anchors, VizAnchor decodes manipulation intent with VLMs through structured reasoning. It first converts the authentic chart $\hat{C}_o$, the tampered chart C_t, and the predicted tamper mask $\hat{M}$ into a four-panel visual prompt. It then uses three task-specific VLM reasoning steps to progressively ground visual edits, reconstruct chart narratives, and infer the misleading intent.

Visual Prompt Construction. Given $\hat{C}_o$, C_t, and $\hat{M}$, we construct a four-panel visual prompt $\mathcal{P}_v = \langle \hat{C}_o, C_t, \hat{C}_m, \hat{C}_l \rangle$. The first two panels show the authentic and tampered charts. The third panel, denoted as the Tamper Map $\hat{C}_m$, overlays the Spatial Anchor on $\hat{C}_o$ by drawing a thin contour around the tamper region. This highlights localized evidence while preserving the semantic context of the authentic chart. The fourth panel, denoted as the Localized Comparison $\hat{C}_l$, crops and enlarges the regions from $\hat{C}_o$ and C_t according to $\hat{M}$ and places them side by side. This prompt provides both global chart context and fine-grained local comparison.

Misleader Grounding. MG grounds visual differences into structured manipulation semantics. Given $\mathcal{P}_v$, it predicts the tamper type, identifies the tampered components, and generates a manipulation-process description:

$$f_{\mathrm{MG}}(\mathcal{P}_v) \rightarrow \left\{\hat{y}_{\mathrm{type}}, \hat{\mathcal{C}}_{\mathrm{comp}}, \hat{y}_{\mathrm{proc}}\right\}.$$

Following VizDefender (Song et al., 2026), we adopt its taxonomy of nine chart-specific tamper types and seven tampered-component categories. The predicted type and components specify what misleader is used and which visual element is affected, while the process description explains how the tampered chart deviates from the authentic one.

Chart Narrative Reconstruction. CNR analyzes how the chart message changes after manipulation. It receives $\hat{C}_o$ and C_t and reconstructs two chart narratives:

$$f_{\mathrm{CNR}}\left(\hat{C}_o, C_t\right) \rightarrow \{\hat{n}_o, \hat{n}_t\},$$

where $\hat{n}_o$ summarizes the authentic trends, comparisons, rankings, or conclusions, and $\hat{n}_t$ summarizes the message conveyed by the tampered chart. This step captures the semantic shift caused by the manipulation, complementing the localized evidence extracted by MG.

Intent Inferring. II performs the final reasoning step. It receives the visual prompt $\mathcal{P}_v$, the grounded manipulation outputs from MG, and the reconstructed narratives from CNR:

$$f_{\mathrm{II}}\left(\mathcal{P}_v, \hat{y}_{\mathrm{type}}, \hat{\mathcal{C}}_{\mathrm{comp}}, \hat{y}_{\mathrm{proc}}, \hat{n}_o, \hat{n}_t\right) \rightarrow \hat{y}_{\mathrm{intent}}.$$

The output $\hat{y}_{\mathrm{intent}}$ represents the inferred misleading intent, describing how the manipulation is intended to alter viewers' interpretation of the chart. By integrating localized visual evidence, structured manipulation semantics, and original–tampered narrative contrast, II infers the underlying misleading intent based on observable modifications and their semantic consequences.

4 Experiments

Our experiments evaluate the two stages of VizAnchor: dual-anchor evidence construction and VLM-based manipulation understanding. We further conduct ablation studies on visual evidence and task decomposition. Additional qualitative results are provided in the appendix.

4.1 Experimental Details

Datasets. We conduct experiments on VisGuard (VGD) (Ye et al., 2026), VizDefender (VDD) (Song et al., 2026), and our VizAnchor Dataset (VAD). VAD comprises VAD-LocTrain, containing 1,500 automatically generated chart pairs with pixel-level masks, and VAD-Eval, containing 120 manually created pairs for evaluating localization and manipulation understanding. VAD-LocTrain employs three synthetic edit families, while VAD-Eval additionally includes crop-based manipulations. Detailed construction protocols, sampling strategies, and quality-control procedures are provided in Appendix **??**.

Baselines. We compare VizAnchor with image-hiding, tamper-localization and manipulation-understanding methods, including HiNet, ISN, StampOne, StegaStamp, WAM, EditGuard, OmniGuard, VizDefender, and VisGuard (Jing et al., 2021; Lu et al., 2021; Shadmand et al., 2024; Tancik et al., 2020; Sander et al., 2025; Zhang et al., 2024; 2025; Song et al., 2026; Ye et al., 2026).

Metrics. We evaluate visual fidelity via PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), reference retrieval via Correct Retrieval Rate (CRR), defined as the proportion of test images for which the decoded metadata successfully retrieves the correct original and untampered watermarked charts, metadata recovery via bit accuracy (BitAcc), and tamper localization via IoU, Precision, Recall, and F1. Tamper-type classification is evaluated via Accuracy and Macro-F1, while component recognition uses Exact Match Accuracy and Macro-F1. For manipulation-process and

Table 1: Watermarked-chart fidelity and reference retrieval under modified-pixel corruption on VGD. Parentheses indicate payload length for methods and BitAcc (%) for results. Bold and underlined values denote the best and second-best results, respectively.

Method	Image Fidelity			Modified-Pixel CRR (BitAcc) (%)			
	PSNR ↑	SSIM ↑	LPIPS ↓	0	15	30	45
HiNet (81)	29.95	0.8306	0.3692	85.83 (98.99)	69.90 (97.87)	49.28 (94.68)	18.91 (87.29)
ISN (81)	34.38	0.9478	0.2340	81.69 (99.08)	68.03 (98.20)	53.32 (96.73)	37.45 (94.49)
StampOne (81)	29.44	0.9107	0.1913	2.94 (91.90)	1.24 (91.07)	0.80 (89.95)	0.43 (88.17)
StegaStamp (81)	10.32	0.5733	0.6852	78.78 (98.45)	70.03 (98.07)	58.94 (97.56)	43.17 (96.61)
WAM (32)	31.55	0.9440	0.1513	9.12 (92.36)	6.55 (90.87)	3.71 (88.01)	1.64 (83.50)
EditGuard (64)	33.00	0.8752	0.2416	31.98 (83.11)	21.45 (83.09)	10.69 (81.26)	2.84 (78.07)
OmniGuard (100)	<u>41.64</u>	**0.9839**	**0.0443**	96.46 (99.68)	12.50 (95.85)	0.10 (84.65)	0.00 (72.55)
VizDefender	34.46	0.8840	0.2207	–	–	–	–
VisGuard (81)	41.13	0.9734	<u>0.0610</u>	<u>98.24</u> (99.93)	<u>99.03</u> (99.90)	<u>97.75</u> (99.83)	<u>96.17</u> (98.83)
VizAnchor (81)	**43.28**	<u>0.9809</u>	0.0664	**99.95** (99.97)	**99.91** (99.96)	**99.88** (99.93)	**98.84** (99.87)

Table 2: Reference retrieval and crop-localization robustness under different crop ratios on VGD.

Metric	Method	Cropped Area Ratio (%)				
		10	20	30	40	50
CRR (BitAcc) (%)	WAM	2.04 (87.47)	1.94 (87.36)	1.54 (87.14)	1.17 (86.62)	0.74 (85.49)
	VisGuard	<u>93.05</u> (97.77)	<u>96.19</u> (98.65)	<u>96.83</u> (98.97)	<u>96.59</u> (98.73)	<u>93.82</u> (97.63)
	VizAnchor	**98.86** (99.55)	**99.03** (99.61)	**99.03** (99.59)	**98.76** (99.55)	**98.26** (99.36)
IoU	VisGuard	0.9789	0.9815	0.9824	0.9814	0.9684
	VizAnchor	**0.9866**	**0.9875**	**0.9860**	**0.9837**	**0.9790**

misleading-intent evaluation, we report Cos-FA, computed using text-embedding-3-large. However, semantic similarity alone may not adequately capture factual correctness, specificity, or consistency with visual evidence. We therefore introduce AI-FA, an LLM-as-a-judge metric using Gemini-3.1-Pro-Preview (Liu et al., 2023; Zheng et al., 2023). AI-FA scores are normalized from a 0–5 scale to $[0, 1]$. Detailed evaluation protocols and scoring criteria are provided in the appendix.

4.2 Semantic Anchor Construction

Watermarked-Chart Fidelity. While VizAnchor achieves the highest PSNR, its SSIM and LPIPS remain competitive rather than consistently outperforming all baselines. Our objective is not to optimize individual image-quality metrics in isolation, but to preserve visual fidelity while embedding the information required for semantic reference recovery and spatial alignment. These capabilities enable the subsequent construction of complementary semantic and spatial evidence for fine-grained manipulation understanding.

Reference Retrieval Robustness. We evaluate reference retrieval under local modifications and cropping using Correct Retrieval Rate (CRR), with BitAcc reported for comparison. As shown in Tables 1 and 2, VizAnchor consistently achieves the highest CRR under both conditions. It maintains a CRR of 98.84% under 45% pixel modification and 98.26% under 50% cropping, demonstrating robust reference retrieval despite substantial content loss.

4.3 Spatial Anchor Construction

Crop Localization. We evaluate crop localization on VGD and the VAD-LocEval crop subset. Since existing localization methods rarely address crop-based manipulations, we compare with VisGuard. As shown in Tables 2 and 3, VizAnchor consistently improves localization across both evaluations. These results are relevant because cropping can selectively remove critical context and mislead viewers, highlighting the importance of crop-aware localization.

Table 3: Tamper-localization results on VAD-LocEval and VDD-LocEval.

Manipulation	Method	IoU ↑	Prec. ↑	Recall ↑	F1 ↑
Crop	VisGuard	0.9710	**0.9982**	0.9726	0.9851
	VizAnchor	**0.9789**	0.9981	**0.9808**	**0.9893**
Local Edit	EditGuard	0.5062	0.5969	0.7399	0.6289
	VizDefender	0.5324	0.6178	0.7584	0.6556
	VisGuard	0.6067	0.7668	0.7622	0.7152
	VizAnchor	**0.7418**	**0.8011**	**0.9105**	**0.8375**

Local-Edit Localization. We evaluate the Localization Module on VDD-LocEval and the local-edit subset of VAD-LocEval. OmniGuard is not included as a localization baseline because its public release does not provide the tamper localization weights, while only the watermark recovery component is available. Table 3 shows that VizAnchor substantially outperforms all baselines, improving overall IoU to 0.7418 and F1 to 0.8375.

4.4 Decoding Manipulation Intent with VLMs

We evaluate VLM-based manipulation understanding on VDD-ReasonEval and VAD-ReasonEval across four tasks. All three reasoning steps share the same gemini-3.5-flash backbone.

Tamper Type Classification. We formulate tamper type prediction as closed-set single-label classification over the predefined taxonomy. As shown in Figure 4, VizAnchor achieves an Accuracy of 0.91 and a Macro-F1 of 0.90, outperforming VizDefender by 0.11 and 0.09, respectively. Access to the original chart enables explicit comparison of visual encodings and data semantics, reducing the ambiguity of inferring manipulation types from the tampered chart alone.

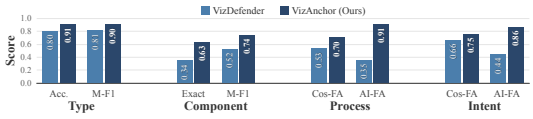

Figure 4: Sample-weighted tamper-type, component recognition, manipulation-process and misleading-intent results across VAD-ReasonEval and VDD-ReasonEval. Exact denotes exact-match accuracy, and M-F1 denotes Macro-F1.

Tampered Component Recognition. We formulate tampered-component recognition as multi-label classification because one manipulation may affect multiple chart elements. Figure 4 shows that VizAnchor improves Exact Match Accuracy from 0.34 to 0.63 and Macro-F1 from 0.52 to 0.74. These substantial gains indicate that the Semantic Anchor provides the original visual context, while the Spatial Anchor highlights specific modified regions, providing complementary evidence for identifying affected chart components.

Manipulation Process and Misleading Intent. We evaluate the manipulation processes and misleading intents generated by our complete framework. As shown in Figure 4, VizAnchor outperforms VizDefender on both tasks. Unlike VizDefender, which relies solely on the tampered visualization, VizAnchor leverages dual-anchor evidence to incorporate the original chart as a semantic reference and explicitly localize manipulated regions. This enables more precise identification of manipulation processes and more informed inference of misleading intents through original–tampered narrative comparison.

4.5 Ablation Study

Visual Evidence. We conduct ablation experiments to investigate the contributions of different visual evidence. As shown in Figure 5, introducing the Semantic Anchor substantially improves tamper-type Macro-F1 from 0.473 to 0.778, highlighting the importance of original–tampered semantic comparison. The Spatial Anchor complements this reference by localizing manipulated regions and grounding subsequent reasoning in specific visual evidence, supporting interpretability and evidence traceability. Localized Comparison (LC) facilitates fine-grained comparison of the

modified regions. Together, these results illustrate the complementary roles of semantic and spatial evidence in manipulation understanding. Detailed settings and results are provided in the appendix.

Task Decomposition. We evaluate the contribution of each reasoning step to the final intent inference. The complete framework achieves the best performance with an Intent Cos. FA of 0.753 and an AI FA of 0.856. Removing MG or CNR degrades performance (dropping to 0.734/0.806 and 0.743/0.832 for Cos./AI FA, respectively). Direct Intent Inference yields the lowest scores (0.721 and 0.826), demonstrating the importance of misleader grounding and narrative reconstruction for accurate intent reasoning. Detailed per-dataset metrics are provided in the appendix.

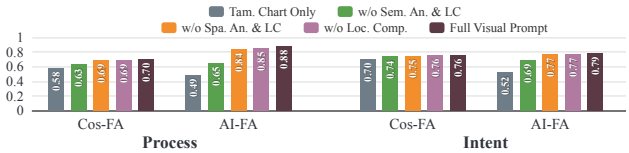

Figure 5: Overall visual-evidence ablation results across VAD-ReasonEval and VDD-ReasonEval.

Contribution Analysis. These ablations clarify the sources of VizAnchor's performance gains. Metadata recovery primarily enables access to the authentic reference rather than directly contributing to VLM reasoning. The Semantic Anchor provides the main improvement in tamper-type recognition through original–tampered semantic comparison, whereas the Spatial Anchor provides complementary region-level evidence for fine-grained manipulation grounding. Task-structured VLM reasoning further improves intent inference by integrating grounded modifications and narrative differences. Thus, the improvements over VizDefender reflect both access to additional visual evidence and structured reasoning, rather than the reasoning structure alone.

5 User Study

To evaluate human-perceived explanation quality, we conducted a blind user study with 30 participants, including visualization experts and individuals from diverse professional backgrounds. Participants rated 10 anonymized manipulation examples on a five-point Likert scale ($\mu_{\text{age}} = 24.07$ years, $\sigma = 4.28$; 16 males, 13 females, and 1 who preferred not to disclose their gender). We compared VizAnchor against VizDefender and Dual-Anchor with a vanilla VLM. As shown in Figure 6, VizAnchor received higher average ratings across both dimensions, scoring 1.437 for manipulation process and 1.370 for misleading intent, compared with VizDefender (-0.547 and -0.010) and the vanilla VLM (1.197 and 1.027). These results demonstrate the potential of combining dual-anchor evidence with task-structured VLM reasoning to improve human-perceived explanation quality.

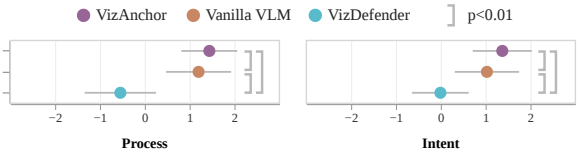

Figure 6: The results of user study. Each circle represents the mean score, with error bars indicating the standard deviation. The statistical significance is assessed using the Wilcoxon signed-rank test.

6 Conclusion

We present **VizAnchor**, a dual-anchor framework that combines authentic semantic references with fine-grained spatial evidence to support visualization manipulation understanding. Through task-structured VLM reasoning, VizAnchor grounds manipulation attributes, reconstructs chart narratives, and infers misleading intents. Experiments demonstrate reliable reference recovery, accurate tamper localization, and effective evidence-grounded reasoning. However, VizAnchor requires proactively watermarked charts and remains vulnerable to holistic chart regeneration that destroys embedded signals. Future work will integrate proactive protection with passive forensic techniques to support unprotected charts and verify global semantic consistency.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of the manuscript, including language editing and LaTeX consistency checking. They were also used for retrieval and discovery purposes, such as identifying potentially relevant related work. All retrieved sources were independently verified by the authors, and all technical content, experimental settings, analyses, citations, and reported results were reviewed by the authors, who take full responsibility for the manuscript.

ETHICS STATEMENT

Our user study was approved by the relevant Institutional Review Board (IRB) prior to data collection. All 30 participants provided informed consent before participation. Participants evaluated anonymized outputs from VizAnchor and baseline methods in a blind study, with method identities concealed throughout the evaluation. Responses were analyzed at the participant level using appropriate statistical tests. Further details on the study protocol, participant demographics, and statistical analyses are provided in the supplementary material.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility in the main paper and supplementary material. The main paper specifies the proposed architecture, training objectives, datasets, baselines, evaluation metrics, and experimental settings. The supplementary material further documents dataset construction and splits, random seeds, model architectures and hyperparameters, training configurations, VLM inference settings, complete prompts and output schemas, automated evaluation procedures and scoring rubrics, ablation settings, and user-study design and statistical analysis. The computing hardware used for model training is also reported.

REFERENCES

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14185–14193, October 2021.

Zixin Chen, Sicheng Song, KaShun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13756–13789, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.695. URL https://aclanthology.org/2025.emnlp-main.695/.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jiayun Fu, Bin Zhu, Weiwei Cui, Song Ge, Yun Wang, Haidong Zhang, He Huang, Yuanyuan Tang, Dongmei Zhang, and Xiaojing Ma. Chartem: Reviving chart images with data embedding. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):337–346, 2021. doi: 10.1109/TVCG.2020.3030351.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606–20615, June 2023.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 41–49, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475324. URL https://doi.org/10.1145/3474085.3475324.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4733–4742, October 2021.

Feifei Kou, Yuhan Yao, Siyuan Yao, Jiahao Wang, Lei Shi, Yawen Li, and Xuejing Kang. Iwrn: A robust blind watermarking method for artwork image copyright protection against noise attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 370–378, 2025.

Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580910. URL https://doi.org/10.1145/3544548.3580910.

Maxim Lisnic, Alexander Lex, and Marina Kogan. "yeah, this graph doesn't show that": Analysis of online engagement with misleading data visualizations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642448. URL https://doi.org/10.1145/3613904.3642448.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. doi: 10.1109/TCSVT.2022.3189545.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. Misinformed by visualization: What do we learn from misinformative visualizations? *Computer Graphics Forum*, 41(3):515–525, 2022. doi: https://doi.org/10.1111/cgf.14559. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14559.

Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L. Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10816–10825, June 2021.

Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. The perils of chart deception: How misleading visualizations affect vision-language models. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pp. 6–10, 2025. doi: 10.1109/VIS60296.2025.00006.

Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 1469–1478, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702608. URL https://doi.org/10.1145/2702123.2702608.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, 2025.

Farhad Shadmand, Iurii Medvedev, Luiz Schirmer, João Marcos, and Nuno Gonçalves. Stampone: Addressing frequency balance in printer-proof steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4367–4376, June 2024.

Sicheng Song, Yanjie Zhang, Zixin Chen, Huamin Qu, Changbo Wang, and Chenhui Li. Vizdefender: Unmasking visualization tampering through proactive localization and intent inference. *IEEE Transactions on Visualization and Computer Graphics*, 32(6):4720–4730, 2026. doi: 10.1109/TVCG.2026.3694448.

Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117–2126, 2020.

Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. Is this chart lying to me? automating the detection of misleading visualizations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8823–8844, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.398. URL https://aclanthology.org/2026.acl-long.398/.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Huayuan Ye, Chenhui Li, Yang Li, and Changbo Wang. Invvis: Large-scale data embedding for invertible visualization. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1139–1149, 2024. doi: 10.1109/TVCG.2023.3326597.

Huayuan Ye, Shenzhuo Zhang, Shiqi Jiang, Jing Liao, Shuhang Gu, Dejun Zheng, Changbo Wang, and Chenhui Li. Robust message embedding via attention flow-based steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12840–12849, June 2025.

Huayuan Ye, Juntong Chen, Shenzhuo Zhang, Yipeng Zhang, Changbo Wang, and Chenhui Li. Visguard: Securing visualization dissemination through tamper-resistant data retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):1295–1305, 2026. doi: 10.1109/TVCG.2025.3634818.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *CoRR*, abs/1909.01285, 2019. URL http://arxiv.org/abs/1909.01285.

Peiying Zhang, Chenhui Li, and Changbo Wang. Viscode: Embedding information in visualization images using encoder-decoder network. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):326–336, 2021. doi: 10.1109/TVCG.2020.3030343.

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 (CVPR)*, June 2018.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11964–11974, June 2024.

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3008–3018, June 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.