TRACE: Token Reference Attention for Auditable Deepfake Region Localization

Yi-Fan Wang, **You-Cheng Chao**, **Chia-Ming Lee**, **Chih-Yu Jian** and **Chih-Chung Hsu**
National Yang Ming Chiao Tung University
{yifanwang652, 19951225abc, zuw408421476, ru0354m3, cc.jess.hsu}@gmail.com

Abstract

Modern image manipulations can be photorealistic within the edited region. A real-or-fake label alone is therefore insufficient for forensic analysis, which also needs to identify where the evidence appears. We present **TRACE**, a supervised deepfake localizer built around an explicit *same-image comparison*. For each visual token, source-gated attention first constructs a reference from the same image, and an MLP encodes the token, its reference, and their difference into a pre-decoder evidence field. A lightweight decoder then predicts a mask from this field, and its connected components define boxes. On an internally held-out DDL-X test split, TRACE achieves 0.7249 BBox IoU under the project metric, and a stricter one-to-one evaluation on validation yields 0.7024 IoU. The exposed pre-decoder field reaches 0.6806 token AUPRC and 0.9348 AUROC on fake validation images, compared with 0.5118 and 0.9167 for an image-blind control built from distance to the image center. When real images are scored against boxes borrowed from fake images, this control retains 0.9171 while the field falls to 0.6313. Together, these results demonstrate strong in-domain localization and show that, before dense decoding, TRACE exposes an inspectable spatial field whose alignment exceeds that of the image-blind control and changes with image content.

1 Introduction

Generative editing has pushed image forgery well beyond conspicuous copy–paste artifacts, as splicing, inpainting, object replacement, and face manipulation can now appear photorealistic within the edited region. Image-level detection remains important [Li *et al.*, 2019; Frank *et al.*, 2020; Yan *et al.*, 2023], but a binary label does not tell an analyst which region supports the decision. That gap motivates benchmarks with spatial annotations, including DDL and MFFI [Miao *et al.*, 2025a; Miao *et al.*, 2025b], and the DDL-X challenge, which evaluates classification, localization, and explanation [IJCAI–ECAI 2026, 2026; DDL-X Challenge Organizers, 2026]. This paper addresses the visual-localization component by predicting suspicious-region masks and boxes while exposing an intermediate spatial signal.

Localization is difficult because a forged region need not look unusual on its own. An inpainted object can carry plausible texture and a replaced face can look convincing up close, yet either may still disagree with the lighting, material, frequency content, boundary behavior, or semantics of the surrounding image. The forgery is therefore often revealed by relationships across the image. Earlier forensic systems have exploited within-image consistency and source–target correspondence [Huh *et al.*, 2018; Wu *et al.*, 2018; Zhao *et al.*, 2021]. We ask how to make this comparison an explicit part of a learned dense localizer.

A frozen visual encoder provides rich spatial features, although raw token statistics alone do not serve as reliable forgery scores because high magnitude may reflect ordinary texture, contrast, or semantic salience. Dense decoders can also learn accurate masks without exposing the pre-decoder signal that supports them. These observations motivate a model that retains the decoder's spatial capacity and makes the preceding comparison inspectable.

We introduce **TRACE** (Token Reference Attention for Auditable Deepfake Region Localization). Given two spatial token fields from a DINOv2 encoder [Oquab *et al.*, 2023], TRACE uses source-gated attention to build a same-image reference for each query, encodes the query, reference, and their difference into an evidence feature, and decodes the resulting field into a dense mask. Figure 1 previews these stages on one example. The base encoder stays frozen during TRACE training, while lightweight LoRA adapters [Hu *et al.*, 2022] and the TRACE head are optimized jointly. Supervision comes from box-derived masks and image labels alone. The evidence field has no auxiliary token target and is shaped by gradients that the dense mask loss propagates through the decoder.

The reference is constructed from the input image by a learned mechanism and is not anchored to a certified clean region. Accordingly, TRACE's auditability comes from exposing and measuring the pre-decoder field, while causal attribution of the final prediction remains outside the scope of this paper.

On an internally held-out DDL-X test split, TRACE reaches 0.7249 BBox IoU under the project metric, and a stricter one-to-one evaluation on validation yields 0.7024

Figure 1: TRACE intermediate representations. For an illustrative example, panels show (a) input with ground-truth boxes, (b) raw encoder-token norm, (c) pre-decoder evidence, and (d) decoded predictions. Green and red boxes denote ground truth and predictions, respectively. Panels (b)–(c) share the displayed color scale.

IoU. The pre-decoder field reaches 0.9348 token AUROC on fake validation images and falls to 0.6313 when real images are scored against boxes borrowed from fake images. The output-level and intermediate-signal results together show that TRACE provides strong in-domain localization while exposing a spatial field that can be examined before dense decoding. Additional controls separate alignment that changes with image content from dataset-level spatial bias and clarify the behavior of the source gate and learned reference.

Contributions.

- **Formulation.** We cast deepfake region localization as a reference-conditioned same-image comparison, evaluating each token against image-derived context rather than assigning it an independent anomaly score.
- **Method.** TRACE constructs source-gated references and exposes a query–reference-conditioned evidence field before dense mask decoding.
- **Evaluation.** We report held-out localization under the project box metric, a stricter one-to-one evaluation on validation, and controlled analyses of the exposed pre-decoder field and the learned same-image reference.

2 Related Work

2.1 Forgery Localization

Image forensics has increasingly moved from binary recognition toward spatial evidence. Early methods learn dense RGB/noise traces [Zhou *et al.*, 2018; Wu *et al.*, 2019]. Later systems model compression history [Kwon *et al.*, 2021], boundary and semantic cues [Chen *et al.*, 2021], long-range interactions [Hao *et al.*, 2021], or object-level reasoning [Wang *et al.*, 2022]. Face and synthesis detectors also use blending boundaries, frequency artifacts, and self-blended training data [Li *et al.*, 2020; Frank *et al.*, 2020; Shiohara and Yamasaki, 2022]. Recent methods decode frozen vision-language features or learn weakly supervised regions for diffusion-generated images [Smeu *et al.*, 2025; Tantaru *et al.*, 2024], and we adopt DeCLIP and Dolos as peer methods. TRACE makes the same-image comparison explicit and exposes the resulting field before the final dense decoder.

2.2 Same-Image Consistency and References

Relational forensic methods compare acquisition cues across regions [Huh *et al.*, 2018; Zhao *et al.*, 2021], model copy–move source/target correspondence [Wu *et al.*, 2018], or learn cross-region agreement [Zhai *et al.*, 2023]. Anomaly-localization methods compare a query with camera fingerprints [Cozzolino and Verdoliva, 2020], population or memory-bank references [Cohen and Hoshen, 2020; Defard *et al.*, 2021; Roth *et al.*, 2022], reconstructions [Zavrtanik *et al.*, 2021; Guo *et al.*, 2025], or teacher features [Bergmann *et al.*, 2020; Deng and Li, 2022]. TRACE builds a reference dynamically from the input image for every token, without an external normal bank or fixed clean template. This design removes the need for external memory, while allowing manipulated content to enter the reference. Section 4.5 measures this behavior directly.

2.3 Frozen Backbones and Adaptation

Large pretrained vision transformers [Dosovitskiy *et al.*, 2021] have become standard foundations for recognition and dense prediction. CLIP, DINOv2, DPT, and Segment Anything demonstrate transferable representations, self-supervised spatial features, dense prediction priors, and promptable segmentation [Radford *et al.*, 2021; Oquab *et al.*, 2023; Ranftl *et al.*, 2021; Kirillov *et al.*, 2023]. LoRA, visual prompt tuning, and AdaptFormer reduce task-specific updates [Hu *et al.*, 2022; Jia *et al.*, 2022; Chen *et al.*, 2022]. Frozen-backbone and adapter methods reuse these features for dense prediction [Chen *et al.*, 2023; Xu *et al.*, 2023; Zhang *et al.*, 2024]. TRACE follows this parameter-efficient setting with a frozen base encoder, LoRA updates, and a task-specific head. The head retains the query and reference features alongside their difference, so our claims apply to the complete relational representation.

3 Method

3.1 Overview and Motivation

TRACE represents the same-image comparison explicitly (Figure 2). For each token, it constructs a reference from the same image and encodes the query, reference, and their difference into a learned evidence feature. The resulting feature field is then decoded into a mask. The feature also retains the

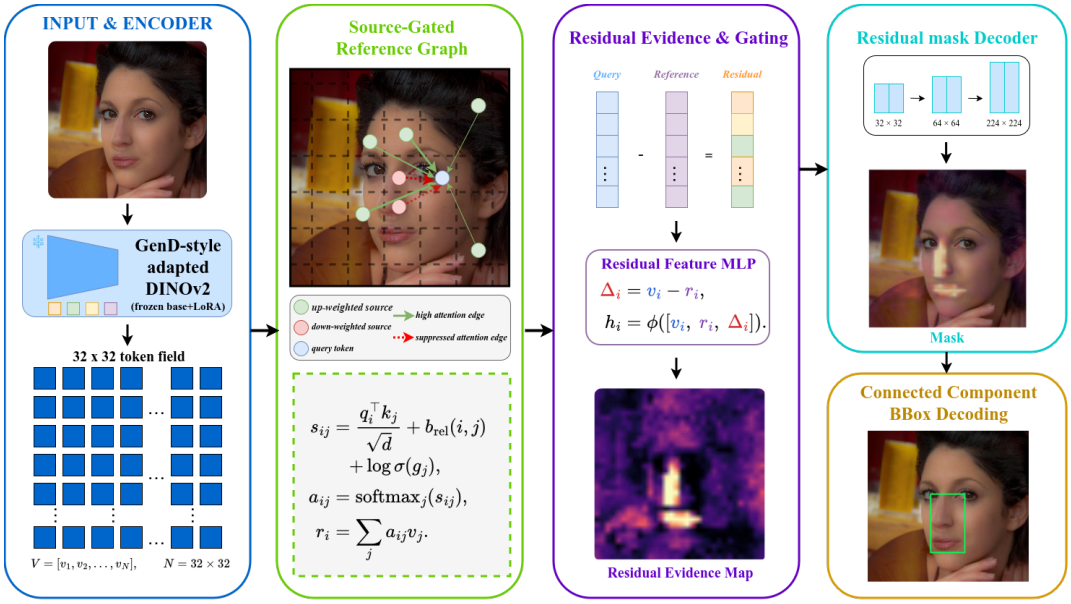

Figure 2: Overview of TRACE. Source-gated attention builds the reference r_i for each token of the 32×32 field, the MLP encodes $[v_i; r_i; v_i - r_i]$ into h_i, and a residual decoder predicts a 224×224 mask whose connected components define boxes.

query, keeping local appearance available while making the comparison explicit and inspectable before dense decoding.

3.2 Source-Gated Reference Graph

Given a 448×448 input, the DINOv2 ViT-L/14 base encoder yields $N = 32{\times}32 = 1024$ spatial tokens. We extract a mid-layer *type* field $z^{\mathrm{t}} = [z_1^{\mathrm{t}}, \ldots, z_N^{\mathrm{t}}]$, $z_i^{\mathrm{t}} \in \mathbb{R}^{d_z}$ (block 11), which forms queries and keys, and a deeper field $z_i^{\mathrm{v}} \in \mathbb{R}^{d_v}$ (block 17), which supplies values. The challenge-adapted base weights stay frozen during TRACE training, leaving only the rank-8 LoRA adapters on the qkv projections of blocks 0–17 and the TRACE head trainable. The preceding classification-only encoder adaptation is documented in Section 4.1.

Affine projections form $q_i = W_Q z_i^{\mathrm{t}} + b_Q$, $k_j = W_K z_j^{\mathrm{t}} + b_K$, and $v_j = W_V z_j^{\mathrm{v}} + b_V$ in a shared space of dimension $d = 128$. The same-image reference is

$$r_i = \sum_{j=1}^{N} a_{ij}\, v_j, \qquad a_{ij} = \frac{\exp(s_{ij}/T)}{\sum_{j'=1}^{N} \exp(s_{ij'}/T)}, \tag{1}$$

with attention score

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d}} + b_{\mathrm{rel}}(i,j) + \log \sigma(g_j), \tag{2}$$

where b_{rel} is a learned relative-position bias, σ is the logistic sigmoid, and $T = \max(\exp \eta, 10^{-2})$ is a positive temperature with learned scalar η ($T = 0.9540$ in the shipped checkpoint). The sum includes $j = i$, so self-reference remains available.

Each token also produces a source scalar,

$$g_j = w_2^\top \,\mathrm{GELU}\big(W_1 z_j^{\mathrm{t}} + b_1\big) + b_2, \tag{3}$$

with hidden width $d_g = 128$. It depends on the source token alone and is shared across queries. The scalar enters the softmax score through the source-side factor $a_{ij} \propto \sigma(g_j)^{1/T} \exp\big((q_i^\top k_j/\sqrt{d} + b_{\mathrm{rel}}(i,j))/T\big)$, so g_j changes how token j is used as a source and leaves its role as a query unchanged. The gate has no separate target and is learned through the final objectives as a task-driven reweighting.

3.3 Residual Evidence and Gated Decoding

The query–reference difference is

$$\Delta_i = v_i - r_i, \tag{4}$$

and a lightweight MLP encodes all three terms:

$$h_i = \phi([v_i;\, r_i;\, \Delta_i]) \in \mathbb{R}^{128}, \tag{5}$$

where brackets denote channel-wise concatenation. Since Δ_i is determined by v_i and r_i, and h_i keeps a direct v_i path, the role of Δ_i is to make the comparison explicit within a representation that still carries local appearance and reference context.

A scalar evidence head ψ modulates each feature:

$$\tilde{h}_i = h_i \cdot \sigma(\psi(h_i) + \beta)\,, \tag{6}$$

where β is a learned bias. Collecting $\tilde{h}_i$ and $\psi(h_i)$ on the token grid gives $\widetilde{H}_{32}$ and Ψ_{32}. The decoder produces

$$M_{224} = D_\theta\Big(\widetilde{H}_{32},\, \Psi_{32}\Big) \in \mathbb{R}^{224\times224}. \tag{7}$$

Internally, the decoder concatenates the gated features and evidence logits, projects them with a 1×1 convolution, and applies residual blocks. It operates at 32×32, upsamples to 64×64, and then to the 224×224 output.

3.4 Training Objectives and Inference

DDL-X provides box annotations but no pixel masks, so we rasterize the union of all ground-truth boxes into a 224×224 binary target. Images without boxes use an all-zero target. TRACE is then trained against this box-derived rasterized mask together with the image-level real/fake label, with the base encoder held frozen and the LoRA adapters and the TRACE head optimized jointly. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, \tag{8}$$

with $\lambda_{\text{cls}} = 1.0$, where the mask loss combines binary cross-entropy and Dice terms,

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}, \tag{9}$$

and $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss of an auxiliary classification branch on the image-level real/fake label.

In the shipped model, the evidence head and attention weights receive no token-level targets. Their supervision comes from gradients that $\mathcal{L}_{\text{mask}}$ propagates through the decoder, while $\mathcal{L}_{\text{cls}}$ trains the classification branch.

At inference, connected components of the thresholded mask probability map are converted to axis-aligned boxes. Section 4.1 gives the exact decoding rule.

4 Experiments

4.1 Experimental Setup

Scope, data, and splits. DDL-X evaluates image classification, region localization, and text explanation, of which this paper covers localization only. Using random seed 20260611, we stratify the 103,531-image development set by real/fake label into 70:15:15 training, validation, and test partitions, giving 72,473 training images (31,835 real / 40,638 fake) and 15,529 images in each of validation and test (6,821 / 8,708). Validation drives model and decoding selection, while the test split is used only for the peer comparison, the qualitative examples, two explicitly labeled controls, and the cross-split duplicate scan. The official hidden set is not used here. Unless stated otherwise, in-domain validation analyses use all 8,708 fake images, and metrics that require authentic images additionally use all 6,821 real images. Sampled diagnostics and sensitivity subsets state their population explicitly.

DDL-X records boxes, an image label, and text, but not identity, source image, manipulation method, or generator, so disjointness at those levels cannot be certified. The split control in Section 4.6 describes what can still be checked.

Metrics. Best-match BBox IoU is the DDL-X project metric used for the held-out peer comparison. Each ground-truth box takes its highest IoU over the predicted boxes, and the scores are averaged over all ground-truth boxes. Because this protocol can reuse a prediction and does not penalize surplus boxes, we also report Hungarian one-to-one IoU together with box precision/recall at IoU 0.5, and we use Hit@0.7, the fraction of ground-truth boxes whose best-match IoU reaches 0.7, and FP/real, the number of predicted boxes per authentic image. Pre-decoder fields are evaluated separately, using token-level Top-1 pointing, AUPRC, and AUROC.

Encoder adaptation. Before TRACE training, public DINOv2-L/14 is adapted in the style of GenD [Yermakov *et al.*, 2026] on the 72,473 training images using real/fake labels only. The objective is cross-entropy plus uniformity and label-alignment losses with weights $1.0/0.5/0.1$ on the L2-normalized CLS embedding. Training runs one head-only warm-up epoch and then at most 12 epochs updating the classification head and backbone LayerNorm affines. AdamW [Loshchilov and Hutter, 2019] uses head/LayerNorm learning rates $4\times10^{-4}/10^{-4}$, weight decay 10^{-4}, cosine decay, batch size 32, BF16 mixed precision, seed 20260527, and early-stopping patience 3. The selected checkpoint is stage-two epoch 7 by validation image AUROC (0.99998), and training stops at epoch 10. No spatial target is used. The base weights are then frozen for TRACE. Neither adaptation nor TRACE checkpoint selection reads the test split.

TRACE training, preprocessing, and decoding. TRACE trains for 12 epochs with AdamW (head learning rate 3×10^{-4}, LoRA learning rate 2×10^{-5}, weight decay 10^{-4}), batch size 16, 448×448 inputs, BF16 mixed precision, and seed 20260527 on one GPU. Rank-8 LoRA is applied to qkv projections in blocks 0–17. Training images are re-encoded with a random JPEG quality from 40 to 95, whereas DDL-X validation and test images use JPEG quality 90. All resized images use ImageNet normalization. The trainer evaluates a fixed validation subset of 2,000 fake and 2,000 real images and selects checkpoints by best-match BBox IoU on the fake portion. At inference, we threshold the 224×224 mask probability at $\tau = 0.9$, discard connected components smaller than 160 pixels, and turn each remaining component into one axis-aligned box. We apply no opening, coordinate trimming, box merging, non-maximum suppression, or image-level gate.

4.2 In-Domain Localization

We first compare the complete localization systems end to end on the held-out test split (Table 1). TRACE reaches 0.7249 BBox IoU, against 0.6886 for DeCLIP [Smeu *et al.*, 2025] and 0.5394 for the fully supervised Patches-C configuration of Dolos [Tantaru *et al.*, 2024], with all three scored on the same 8,708 fake test images under the same rule. Because the systems differ in training recipe, input preprocessing, resolution, and decoding procedure (TRACE, for example, re-encodes its inputs at JPEG quality 90, whereas the peer pipelines read the original files), the comparison applies to complete systems at their validation-selected configurations. Figure 3 shows six test cases selected for visual inspection and ordered by the TRACE-to-peer margin, in each of which TRACE exceeds 0.70 union-region IoU (defined below) while both peers remain below 0.45. The panel is illustrative and does not represent the full test set.

Because the project metric does not penalize surplus boxes and can reuse one prediction across multiple ground-truth boxes, we also evaluate the full validation output under one-to-one matching, in which Hungarian assignment maximizes total IoU and unmatched ground-truth boxes contribute zero before micro-averaging. At IoU 0.5, the same assignment defines box precision, recall, and unmatched predictions per

Figure 3: Qualitative comparison on selected images from the internally held-out DDL-X test split. Rows show the input with ground-truth boxes, DeCLIP, Dolos (Patches-C), and TRACE. Green and red boxes denote ground truth and predictions, respectively.

Method	BBox IoU ↑
Dolos (Patches-C)	0.5394
DeCLIP	0.6886
TRACE	**0.7249**

Table 1: Comparison with peer methods on the internally held-out DDL-X test split, using best-match BBox IoU on the same 8,708 fake images.

Model	Best-match IoU	One-to-one IoU	Hit@0.7	P@0.5	R@0.5
TRACE	**0.7276**	**0.7024**	**0.7366**	**0.9078**	**0.8068**
Dense decoder + LoRA	0.6892	0.6597	0.6571	0.8710	0.7680

Table 2: Strict box evaluation on DDL-X validation (8,708 fake images, 14,614 ground-truth boxes). Best-match is the project BBox IoU, P/R use one-to-one matching at IoU 0.5, and dense decoder + LoRA is a separately trained control.

fake image. Since DDL-X provides no pixel masks, union-region IoU rasterizes the predicted and ground-truth box unions and macro-averages their IoU over fake images.

Under this stricter accounting (Table 2), TRACE retains 0.7024 one-to-one IoU from a best-match score of 0.7276, with precision@0.5 of 0.9078, recall@0.5 of 0.8068, and union-region IoU of 0.8289. TRACE predicts 1.4915 boxes per fake image, whereas the ground truth contains 1.6782 boxes per image, and it leaves 0.1376 unmatched predictions per fake image. Together with the high precision and lower recall, these counts suggest that missed components are more common than surplus detections.

The result also changes little across nearby decoding thresholds, as BBox IoU at $\tau = 0.70/0.85/0.90$ is 0.7115/0.7236/0.7276 and Hit@0.7 is 0.7085/0.7276/0.7366, while FP/real remains 0.0004. Finally, an image-blind control estimates how much of the score can be explained by dataset layout alone. A map formed by averaging the DDL-X training boxes, the mean-training-box control, reaches 0.2553 best-match BBox IoU on the held-out test split, compared with 0.7249 for TRACE, showing that common spatial layout contributes to the score while accurate instance-specific boxes still depend on image content.

4.3 Encoder and Configuration Controls

We then compare the shipped TRACE model with a set of separately trained controls. Without LoRA, the dense-decoder and TRACE configurations obtain 0.5834 and 0.6547 validation BBox IoU, whereas with LoRA they reach 0.6892 and 0.7276, respectively, and Table 2 compares these two systems under the strict metrics. A no-LoRA TRACE control obtains 0.6520 on the held-out test under the same fixed decoding, so LoRA brings a consistent benefit within each recorded configuration. The two systems differ, however, in the encoder blocks they read, decoder architecture, and capacity, so their gap reflects complete configurations, and a matched architectural study would be needed to isolate the effect of the reference-conditioned representation.

A recorded TRACE run initialized from public DINOv2 obtains 0.7327 validation BBox IoU, compared with 0.7276 for the shipped adapted-encoder model, so in this recorded

Signal	Top-1	AUPRC	AUROC	
	Fake	Fake	Fake	Borrowed
Random pointing	11.2	–	–	–
Center prior	67.0	0.512	0.917	0.917
Token norm $\|z_i^{\mathrm{v}}\|$	17.7	0.122	0.532	–
Source gate $\sigma(g_j)$	48.5	0.421	0.868	0.835
$\psi(h_i)$ (TRACE)	**71.6**	**0.681**	**0.935**	0.631

Table 3: Spatial alignment of pre-decoder signals on DDL-X validation, with fake columns over 8,708 fake images with true boxes and the borrowed-box column over 6,821 real images. Top-1 is a percentage (expected for random pointing and the center prior), and dashes mark undefined or unreported quantities.

realization the preceding encoder adaptation was not required to attain this level of validation performance. Because the runs used different script revisions, however, the comparison does not estimate the causal effect of adaptation.

4.4 Pre-Decoder Spatial Signal

We next examine the evidence field before dense decoding. Its logit $\psi(h_i)$ receives no auxiliary token target and is learned through mask-loss gradients propagated by the decoder and multiplicative gate. Table 3 quantifies the resulting field. Metrics are micro-pooled over tokens, and Top-1 is the fraction of fake images whose highest-scored token lies inside a rasterized ground-truth box.

Raw token norm aligns only weakly with the annotated regions, reaching 0.5317 AUROC and 17.7% Top-1. Alignment can also be inflated by position alone when annotations concentrate in similar image locations, and DDL-X indeed has a strong center prior. The fixed center-distance map reaches a tie-aware expected Top-1 of 67.04%, 0.5118 AUPRC, and 0.9167 AUROC. The TRACE evidence logit nevertheless exceeds this prior on all three measures, reaching 71.57% Top-1, 0.6806 AUPRC, and 0.9348 AUROC, while the source gate reaches 48.51% Top-1, 0.4208 AUPRC, and 0.8675 AUROC.

The center prior and evidence field behave differently, however, once the boxes no longer belong to the image. In the borrowed-box control, each real image is scored against the ground-truth boxes of a fake image. We pair the 6,821 real images with the first 6,821 fake-image box sets in sorted-CSV order, using a single deterministic pairing. Under this exchange, the center prior retains 0.9171 AUROC, whereas the evidence AUROC falls from 0.9348 on fake images to 0.6313 on real images. The output-space image-blind control in Section 4.2 likewise produces poor boxes on the test split, so the center-distance and mean-training-box controls together show that fixed spatial structure explains only part of the observed alignment.

4.5 Reference and Gate Behavior

At the token level, an image-blind spatial map reproduces most of the source gate's box alignment, indicating a substantial spatial-prior component. Defining gate contrast as the mean gate value inside the boxes minus the mean outside, we measure, on fixed-seed samples of 2,000 fake and 2,000 real validation images, a contrast of 0.5095 for fakes

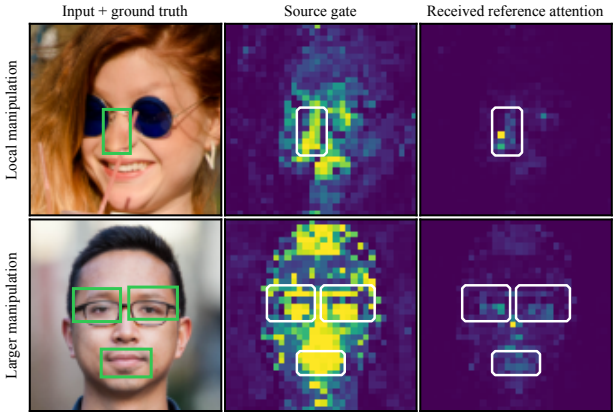

Figure 4: Source-gate and reference-attention examples. Rows show one local and one larger manipulated region. White outlines denote the ground-truth token regions. Received attention is $\sum_i a_{ij}$, the total mass assigned to source token j by all queries.

Quantity	All	Local	Larger
Annotated-token share	0.1123	0.0421	0.1678
Reference mass to annotated tokens			
Queries inside region	0.3277	0.2025	0.4266
Queries outside region	0.4590	0.3037	0.5817

Table 4: Reference composition across manipulation sizes on all 8,708 fake DDL-X validation images, where the local (3,843) and larger (4,865) groups split the ground-truth union at 10% of image area. Reference mass is the mean attention that each row's queries assign to tokens inside that union.

and 0.4534 for reals evaluated with borrowed boxes, while a fixed map formed by averaging the real-image gate maps still gives 0.4543 on the fake-image boxes. These measurements together indicate that the gate acts as a learned source weighting whose box alignment largely follows a spatial pattern shared across images.

Reference attention likewise reaches into the annotated regions. Although the tokens there occupy 0.1123 of the grid, they receive 0.3277 of the reference mass from queries inside the region and 0.4590 from queries outside it. Table 4 reports these quantities by region size, and Figure 4 shows the gate and attention maps for the first image by identifier in each size group. TRACE therefore builds a contextual reference that draws on the whole image, including the annotated regions.

Finally, we evaluate the two scalar gating paths with a single-seed series of separately retrained variants, which are distinct from the shipped checkpoint and use the fixed decoding rule from Section 4.1 on DDL-X validation. The full model reaches 0.7237 BBox IoU and 0.7308 Hit@0.7. Removing the source-score term gives 0.7206 and 0.7271, while removing the decoder-side evidence-logit input gives 0.7254 and 0.7331, marginally above the full model. False alarms remain between 0.0003 and 0.0004 FP/real throughout, and each removal changes the result by only a few thousandths. Because both variants retain the reference-conditioned fea-

ture field, this control evaluates the two scalar paths while keeping the same-image reference mechanism fixed.

4.6 Boundary Analyses

The remaining analyses examine performance on small regions, the extent to which split integrity can be verified, and transfer under the DDL-X-selected operating point.

Small regions. Performance drops sharply for ground-truth boxes below 0.5% of image area, falling to 0.0544 BBox IoU across 19 boxes, whereas BBox IoU ranges from 0.6686 to 0.8868 across the larger-area bins. The drop is consistent with the resolution of the token grid, since even a square covering 0.5% of the image spans only about 2.3 tokens per side on the 32×32 grid and smaller regions occupy fewer tokens. The 160-pixel minimum-component filter, which removes any predicted component below roughly 0.32% of the mask area, further penalizes predictions at this scale. At the other end of the range, the largest manipulation in this split covers 36.9% of image area, so near-global manipulation is not evaluated.

Split control. Because the dataset lacks identity, source-image, manipulation-method, and generator metadata, group-level disjointness cannot be certified. An exhaustive decoded-pixel scan found seven cross-split duplicate groups comprising 14 images, all real. We also ran a more permissive perceptual screen, but because its clusters could not be validated as duplicate or common-source groups, we left them out of the analysis.

Transfer under fixed decoding. The decoding rule of Section 4.1 was selected on DDL-X, so applying the system to other data without re-tuning amounts to a test of one fixed operating point. Performance is poor on the four manipulation sets of the Dolos benchmark [Tantaru *et al.*, 2024], where TRACE falls to 0.1878–0.3461 BBox IoU with 1.2267 FP/real on the real images shared by the four sets and the dense decoder + LoRA control reaches 0.1945–0.3322 with 1.2989. Because the threshold and minimum-area rule were tuned on DDL-X and left unchanged for Dolos, the observed drop reflects both the transferred operating point and the dataset shift. Separating these effects would require target-side tuning, which this test deliberately omits.

5 Discussion

TRACE shows that an explicit same-image comparison can support strong box localization while leaving an intermediate spatial field available for analysis, and it retains most of its best-match score under strict one-to-one matching. The borrowed-box control shows that the exposed field changes with image content, while the center-distance and mean-training-box controls show that dataset layout still contributes to its alignment. The gate and reference analyses further constrain the interpretation. The source gate carries a substantial spatial-prior component, reference attention places mass inside the annotated regions, and the learned reference remains contextual, drawing on those regions as well as their surroundings. Taken together, these direct measurements of the source gate, reference composition, and pre-decoder evidence field provide an inspectable view of TRACE's intermediate processing before dense decoding.

Several factors limit the scope of the evaluation. DDL-X lacks the identity, source-image, manipulation-method, and generator metadata needed to certify group-level disjointness. Performance is poor for extremely small ground-truth boxes, and near-global manipulation is not evaluated. Moreover, performance drops on Dolos under the DDL-X-selected operating point, so the localization claim remains in-domain. Finally, the public-DINOv2 and adapted-encoder runs use different script revisions and therefore do not form a causal comparison of encoder adaptation.

6 Conclusion

We presented TRACE, a supervised deepfake localizer that makes a same-image comparison explicit before dense decoding. Source-gated attention constructs a reference for each token, and an MLP encodes the query, reference, and their difference before a residual decoder predicts the mask, whose connected components define boxes. TRACE achieves 0.7249 BBox IoU under the project metric on the internally held-out DDL-X test split, and a stricter one-to-one evaluation on validation yields 0.7024 IoU. Its pre-decoder field reaches 0.9348 token AUROC on fake validation images and falls to 0.6313 when real images are scored against boxes borrowed from fake images. TRACE therefore combines strong in-domain box localization with an inspectable field, exposed before dense decoding, whose alignment changes with image content. The image-blind controls show that dataset layout accounts for part of that alignment, and the learned reference draws on the annotated regions as well as the wider image. These measurable properties define the scope of our auditability claim.

References

[Bergmann *et al.*, 2020] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4183–4192, 2020.

[Chen *et al.*, 2021] 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)*, pages 14185–14193, 2021.

[Chen *et al.*, 2022] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, 2022.

[Chen *et al.*, 2023] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations (ICLR)*, 2023.

[Cohen and Hoshen, 2020] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[Cozzolino and Verdoliva, 2020] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020.

[DDL-X Challenge Organizers, 2026] DDL-X Challenge Organizers. DDL-X: Deepfake detection, localization, and explainability challenge @ IJCAI 2026 DDL 2.0 workshop. https://www.codabench.org/competitions/15686/, 2026. Accessed 2026-06-10.

[Defard *et al.*, 2021] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition: ICPR International Workshops and Challenges*, 2021.

[Deng and Li, 2022] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, 2022.

[Dosovitskiy *et al.*, 2021] 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.

[Frank *et al.*, 2020] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the International Conference on Machine Learning*, 2020.

[Guo *et al.*, 2025] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20405–20415, 2025.

[Hao *et al.*, 2021] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. TransForensics: Image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15055–15064, 2021.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Huh *et al.*, 2018] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision*, pages 101–117, 2018.

[IJCAI–ECAI 2026, 2026] IJCAI–ECAI 2026. IJCAI–ECAI 2026 competitions: The 2nd challenge on deepfake detection, localization, and interpretability. https://2026.ijcai.org/competitions/, 2026. Accessed 2026-06-10.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitford, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[Kwon *et al.*, 2021] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 375–384, 2021.

[Li *et al.*, 2019] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[Miao *et al.*, 2025a] Changtao Miao, Yi Zhang, Weize Gao, Zhiya Tan, Weiwei Feng, Man Luo, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, Tao Gong, Zhe Li, Weibin Yao, and Joey Tianyi Zhou. DDL: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios, 2025.

[Miao *et al.*, 2025b] Changtao Miao, Yi Zhang, Man Luo, Weiwei Feng, Kaiyuan Zheng, Qi Chu, Tao Gong, Jianshu Li, Yunfeng Diao, Wei Zhou, Joey Tianyi Zhou, and Xiaoshuai Hao. MFFI: Multi-dimensional face forgery image dataset for real-world scenarios, 2025.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski.

DINOv2: Learning robust visual features without supervision, 2023.

[Radford *et al.*, 2021] 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 International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[Ranftl *et al.*, 2021] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

[Roth *et al.*, 2022] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18720–18729, 2022.

[Smeu *et al.*, 2025] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. DeCLIP: Decoding CLIP representations for deepfake localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[Tantaru *et al.*, 2024] Dragos Tantaru, Elisabeta Oneata, and Dan Oneata. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[Wang *et al.*, 2022] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. ObjectFormer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2364–2373, 2022.

[Wu *et al.*, 2018] Yue Wu, Wael Abd-Almageed, and Premkumar Natarajan. BusterNet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European Conference on Computer Vision*, pages 168–184, 2018.

[Wu *et al.*, 2019] 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*, 2019.

[Xu *et al.*, 2023] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2945–2954, 2023.

[Yan *et al.*, 2023] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, 2023.

[Yermakov *et al.*, 2026] Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

[Zavrtanik *et al.*, 2021] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM — a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[Zhai *et al.*, 2023] Yuanhao Zhai, Tianyu Luan, David Doermann, and Junsong Yuan. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22390–22400, 2023.

[Zhang *et al.*, 2024] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen CLIP: A strong backbone for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3806, 2024.

[Zhao *et al.*, 2021] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021.

[Zhou *et al.*, 2018] 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*, 2018.