TRAINING-FREE EVIDENCE-DRIVEN REFINEMENT FOR UNSUPERVISED TALKING HEAD FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised talking head forgery detection faces severe generalization challenges due to the continuous evolution of generators. By reducing reliance on generator-specific forgery patterns, unsupervised detectors offer an alternative for handling unseen generators. Alongside efforts to build stronger detectors, we investigate whether the anomaly score ranking of the existing detector can be improved without modifying its parameters. In particular, for score-based unsupervised detectors, the limited discriminative ability on hard cases is often reflected in unreliable anomaly ordering, leaving room for further refinement. Motivated by this observation, we propose a **T**raining-**F**ree **E**vidence-driven **R**efinement (TFER) framework to improve anomaly score rankings through selective refinement. TFER first partitions samples into retained and routed subsets using lightweight score-based routing. It then revisits only the routed subset by organizing localized visual evidence and performing semantic analysis to refine its relative ordering while preserving the original score multiset. Extensive experiments demonstrate consistent improvements over the base detector across multiple datasets and perturbation settings. These findings demonstrate the effectiveness of selective evidence-driven refinement in improving unsupervised talking head forgery detection without retraining the base detector.

1 INTRODUCTION

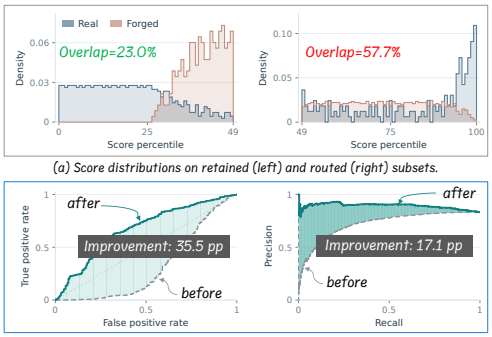
Recent advances in generative models have made talking head forgeries increasingly realistic (Peng et al., 2024; Long et al., 2026; Liu et al., 2025), substantially lowering the barrier to the large-scale creation and dissemination of harmful AI-generated content (Yan et al., 2025; Xie et al., 2025). Reliable talking head forgery detection has therefore become increasingly important. The main difficulty lies in maintaining robust generalization to unseen generators, rather than merely recognizing known manipulations (Guo et al., 2025). This challenge is particularly severe for supervised detectors, as their discriminative capacity is often tied to generator-specific signatures.

Unsupervised methods (Smeu et al., 2025; Feng et al., 2023) aim to reduce reliance on generator-specific forgery patterns by learning authenticity-related regularities. Such methods typically formulate detection as anomaly detection (Liu et al., 2023), assigning anomaly scores according to deviations from the real distribution. However, realistic talking head forgeries can still be difficult to distinguish using these scores. As shown in Figure 1(a), the detector (Smeu et al., 2025) shows stronger real-forged separation in the retained subset and greater overlap in the routed subset. Complementing existing efforts to design stronger detectors (Liu et al., 2024; Kukanov & Ng, 2025), we ask whether an existing detector's anomaly-score ranking can be improved through selective evidence-driven refinement without modifying its parameters.

Based on the above observations, the variation in ranking quality across score regions motivates a selective evidence-driven refinement perspective. We retain the original scores of samples at or below a validation-derived threshold and revisit samples above it for further analysis. Since such detection is fundamentally score-based, we formulate refinement as improving the relative ordering within the routed subset. This shifts the focus from designing an entirely stronger detector to selectively revisiting its anomaly-score ranking. Achieving this goal requires additional evidence beyond

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.

the original detector score. In this regard, Multimodal Large Language Models (MLLMs) (Cai et al., 2026; Cheng et al., 2025) are particularly attractive because of their strength in fine-grained evidence analysis.



(a) Score distributions on retained (left) and routed (right) subsets.

(b) Performance on the routed subset before and after our TFER.

Figure 1: Score distributions and ranking refinement. (a) The base detector shows stronger separation between real and fake samples in the retained subset $\mathcal{C}$ (left) and greater overlap in the routed subset $\mathcal{U}$ (right). (b) ROC and PR curves before and after TFER on the same routed subset, demonstrating improved ranking through selective evidence-driven refinement. *pp* is percentage points.

In recent years, MLLMs have emerged as a promising direction for forgery detection (Gu et al., 2025; Huang et al., 2025), owing to their strong cross-modal modeling capability (Yang et al., 2026). They can provide richer semantic evidence and more interpretable analytical cues, opening new possibilities for fine-grained discrimination in complex forgery scenarios. However, most existing approaches still integrate MLLM capabilities into the detector training pipeline (Yu et al., 2025). Such training-coupled designs inevitably increase training cost and deployment complexity, and remain ill-suited to the rapid evolution of forgery paradigms. In contrast, we do not use large models as detector components that require further training or redesign. Instead, we use them within a training-free refinement stage that is activated only for routed samples, thereby exploiting their strength in fine-grained semantic analysis precisely where it is most needed. As shown in Figure 1(b), this selective refinement substantially improves discrimination within the routed subset.

Building on the above analysis, we develop a Training-Free Evidence-Driven Refinement (TFER) framework for talking head forgery detection. Built on an existing unsupervised detector (Smeu et al., 2025), TFER first estimates a Youden threshold (Fluss et al., 2005) on the labeled validation set and partitions test samples into retained and routed subsets through lightweight score-based routing. Samples in the retained subset keep their original scores, whereas routed samples are forwarded to the evidence-driven refinement stage. Since audio-visual consistency has already been modeled by the base detector, the refinement stage is introduced not to repeat multimodal alignment modeling, but to provide a complementary vision-language perspective for revisiting routed samples. Specifically, frozen CLIP (Radford et al., 2021) and predefined text prototypes are used to score video frames and local patches, identify salient evidence regions, and construct cross-frame patch strips as localized visual evidence. Qwen (Wang et al., 2024a) then performs fine-grained semantic analysis on this evidence to produce textual descriptions, which are converted into ranking scores by a text reranker (Rachidy et al., 2025). The resulting rank scores are used to reorder the routed samples by reassigning their original score slots. This preserves the original score multiset and the ordering between the retained and routed subsets. By supplementing anomaly scores with localized visual-semantic information, TFER improves ranking within the routed subset and overall detection performance in the reported experiments.

In summary, our main contributions are as follows:

- We study selective refinement of anomaly-score rankings for talking head forgery detection, examining how an existing unsupervised detector can be improved without modifying its parameters.
- We propose TFER, a training-free framework that uses lightweight score-based routing to select samples for localized visual evidence mining and semantic analysis, and refines their ordering through slot-preserving reordering while retaining the original score multiset and the ordering between the retained and routed subsets.
- Extensive experiments on multiple benchmarks demonstrate that TFER improves overall detection performance and ranking within the routed subset, with consistent gains over the base detector under the evaluated perturbations.

2 Related Work

2.1 Talking Head Forgery Detection

Talking head generation synthesizes temporally aligned facial motions from speech by modeling cross-modal audio-visual interactions (Chen et al., 2025; Wei et al., 2023). As the realism of such videos improves, corresponding detection methods have evolved from unimodal forgery analysis (Huang et al., 2023; Zheng et al., 2021) to audio-visual collaborative modeling (Liu et al., 2024; Yang et al., 2023). Most existing methods address this challenge by developing increasingly stronger detectors, often with improved multimodal modeling capability. Early methods directly learn audio-visual detectors from labeled real and forged videos (Chugh et al., 2020; Mittal et al., 2020), while later approaches combine unsupervised audio-visual representation learning with labeled adaptation (Zeng et al., 2021; Haliassos et al., 2022; Zhou & Lim, 2021). Although effective in-domain, such methods often struggle to cope with generator shift.

To improve generalization, recent studies have further explored anomaly-based detection strategies trained only on real data (Li et al., 2024; Ricker et al., 2024; Feng et al., 2023; Smeu et al., 2025; Liu et al., 2026). These methods aim to reduce reliance on generator-specific forgery patterns. However, realistic talking head forgeries can still produce anomaly scores with limited separation. Our work therefore investigates whether localized visual-semantic information can improve the score ordering of an existing unsupervised detector through selective refinement without retraining.

2.2 Training-Free Forgery Detection

In recent years, training-free detection (Wang et al., 2024b) has emerged as an alternative to training-based forgery detectors. Instead of learning a new detector, these methods directly exploit the representations or statistical properties already encoded in pretrained foundation models to distinguish real from fake samples. AeroBlade (Ricker et al., 2024) uses reconstruction errors in latent diffusion models, RIGID (He et al., 2024) exploits differences in sensitivity to random perturbations, and ZED (Cozzolino et al., 2024) adapts likelihood-based ideas from AI-generated text detection within a multi-level super-resolution framework. These studies show that effective forgery cues can be derived from reconstructability or statistical consistency without retraining a dedicated detector. More recent work further extends this paradigm. The role of foundation-model robustness in training-free detection has been analyzed in (Tsai et al., 2024). WaRPAD (Choi et al., 2026) constructs a training-free detector from cropping robustness and local patch scoring.

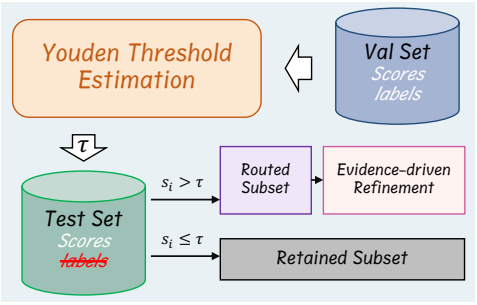


Figure 2: Score-based routing. The threshold τ is estimated on the labeled validation set using the Youden criterion and applied to test anomaly scores without using test labels. Samples with $s_i \leq \tau$ form the retained subset $\mathcal{C}$ and keep their original scores, while samples with $s_i > \tau$ form the routed subset $\mathcal{U}$ and undergo evidence-driven refinement.

However, these methods are mainly developed for fake image detection, and training-free paradigms remain underexplored in talking head forgery detection. More importantly, existing training-free methods typically construct a new detection score directly from raw inputs, whereas our method builds on an existing unsupervised audio-visual detector and uses localized visual-semantic information to refine its ranking within a selected subset.

3 Method

3.1 Overview

Given a test set of videos, let s_i denote the original anomaly score produced by an existing unsupervised audio-visual detector (Smeu et al., 2025) for a sample x_i, where a larger s_i indicates greater suspiciousness. Based on the labeled validation set, a routing threshold τ is estimated using

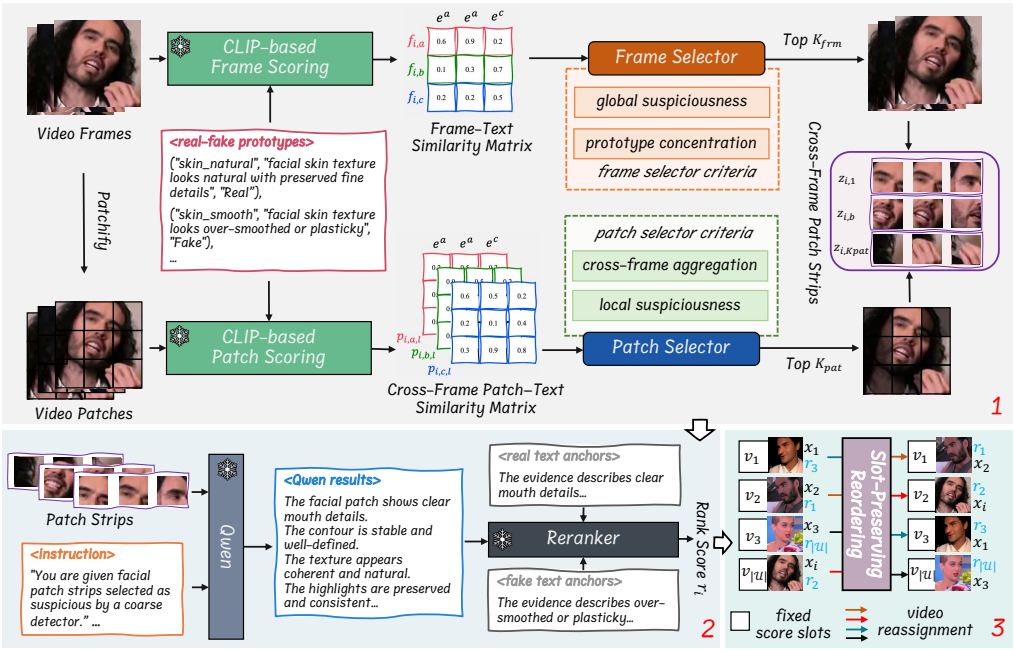


Figure 3: Overview of the evidence-driven refinement stage. (1) For a routed video, frozen CLIP scores sampled frames and local patches using predefined real/fake text prototypes. Frame selection combines global suspiciousness and prototype concentration, while patch selection uses local suspiciousness and cross-frame aggregation. The selected top-K_{frm} frames and top-K_{pat} patch locations form cross-frame patch strips. (2) Qwen analyzes the patch strips with a fixed instruction prompt and generates textual descriptions. A text reranker compares these descriptions against real and fake text anchors to obtain a rank score r_i. (3) Routed samples are reordered according to r_i and reassigned to their original detector score slots, while the scores of retained samples and the ordering between the two subsets remain unchanged.

the Youden criterion (Fluss et al., 2005), which is then used to partition test samples into a retained subset $\mathcal{C}$ and a routed subset $\mathcal{U}$. For each $x_i \in \mathcal{U}$, the evidence-driven refinement stage first mines localized suspicious evidence with a frozen CLIP model (Radford et al., 2021) and predefined real/fake text prototypes generated by GPT-4 (Sanderson, 2023), then performs semantic analysis with Qwen (Wang et al., 2024a), and finally produces a rank score r_i through a text reranker (Rachidy et al., 2025). TFER only refines samples in $\mathcal{U}$ by locally reordering them according to r_i, after which they are reassigned to the original score slots $\mathcal{V}_{\mathcal{U}}$ to obtain the final score $\hat{s}_i$.

3.2 Score-based Routing

As illustrated in Figure 2, the score-based routing stage is built on top of an existing unsupervised audio-visual detector (Smeu et al., 2025) and serves as a lightweight routing module. Its role is to select a subset of samples for evidence-driven refinement while retaining the original scores of the remaining samples.

The routing stage uses the original detector's audio-visual anomaly scores to select samples for further analysis. To select these samples, we estimate a routing threshold τ on the labeled validation set $\mathcal{D}^{\text{val}}$. Let $\mathcal{T}^{\text{val}}$ denote the set of candidate thresholds induced by validation scores. τ is selected using the Youden criterion:

$$\tau = \arg\max_{\tau' \in \mathcal{T}^{\text{val}}} \left(\text{TPR}(\tau') - \text{FPR}(\tau')\right). \tag{1}$$

We adopt the Youden criterion as a simple validation-based rule that maximizes the difference between the true positive rate (TPR) and false positive rate (FPR). This is particularly suitable for our purpose, since this stage is intended to function only as a routing mechanism rather than a separately optimized classifier.

Once τ is obtained, test samples are partitioned into a retained subset $\mathcal{C}$ and a routed subset $\mathcal{U}$ according to:

$$\mathcal{C} = \{x_i \mid s_i \leq \tau\}, \qquad \mathcal{U} = \{x_i \mid s_i > \tau\}. \tag{2}$$

Samples in $\mathcal{C}$ retain the original detector scores, whereas samples in $\mathcal{U}$ are forwarded to the evidence-driven refinement stage for further analysis. This one-sided rule retains the lower-score region and directs additional visual-semantic analysis to higher-scoring candidates. The effect of this selective design is examined in the routing ablation.

3.3 Evidence-driven Refinement

Evidence Mining and Rank Score Estimation. As shown in Figure 3(1) and (2), the evidence-driven refinement stage is activated only for samples in the routed subset $\mathcal{U}$. The base unsupervised detector has already provided an assessment of these samples through audio-visual consistency modeling. The refinement stage supplements the original anomaly scores with localized visual-semantic information for subsequent ranking.

For a routed video $x_i \in \mathcal{U}$, we first construct a frame sequence $\mathcal{F}_i = \{f_{i,t}\}_{t=1}^{T_i}$ by selecting frames according to grayscale frame differences, so as to retain visually informative temporal changes. Each sampled frame is further divided into local patches, where $p_{i,t,\ell}$ denotes the ℓ-th patch of frame $f_{i,t}$. To mine suspicious evidence, we employ a frozen CLIP encoder $\phi(\cdot)$ together with predefined real and fake text prototype sets $\mathcal{E}^r$ and $\mathcal{E}^f$. Here, $\mathcal{E}^r$ and $\mathcal{E}^f$ collect textual prototypes describing real-related and fake-related facial evidence, respectively.

For frame selection, we use two complementary criteria. The first is *global suspiciousness*, which measures the extent to which frame $f_{i,t}$ favors fake-related prototypes over real-related ones:

$$g_{i,t}^{\mathrm{frm}} = \max_{\mathbf{e}^f \in \mathcal{E}^f} \mathrm{sim}\big(\phi(f_{i,t}), \mathbf{e}^f\big) - \max_{\mathbf{e}^r \in \mathcal{E}^r} \mathrm{sim}(\phi(f_{i,t}), \mathbf{e}^r)\,. \tag{3}$$

The second is *prototype concentration*, which measures whether the fake-side response is dominated by a small number of fake prototypes rather than being diffusely distributed over the entire fake prototype set:

$$c_{i,t}^{\mathrm{frm}} = \max_{\mathbf{e}^f \in \mathcal{E}^f} \mathrm{sim}\big(\phi(f_{i,t}), \mathbf{e}^f\big) - \frac{1}{|\mathcal{E}^f|} \sum_{\mathbf{e}^f \in \mathcal{E}^f} \mathrm{sim}\big(\phi(f_{i,t}), \mathbf{e}^f\big)\,. \tag{4}$$

A larger $c_{i,t}^{\mathrm{frm}}$ indicates that the strongest fake-side response is more dominant relative to the average response over the fake prototype set. The final frame score is then defined as:

$$u_{i,t}^{\mathrm{frm}} = g_{i,t}^{\mathrm{frm}} + c_{i,t}^{\mathrm{frm}}. \tag{5}$$

This score combines global suspiciousness and prototype concentration to rank candidate frames.

For patch selection, we instead use *local suspiciousness* and *cross-frame aggregation*. The local suspiciousness of patch $p_{i,t,\ell}$ is defined as:

$$g_{i,t,\ell}^{\mathrm{pat}} = \max_{\mathbf{e}^f \in \mathcal{E}^f} \mathrm{sim}\big(\phi(p_{i,t,\ell}), \mathbf{e}^f\big) - \max_{\mathbf{e}^r \in \mathcal{E}^r} \mathrm{sim}(\phi(p_{i,t,\ell}), \mathbf{e}^r)\,. \tag{6}$$

It measures how strongly this local region favors fake-related evidence over real-related evidence. To improve temporal stability, we aggregate the same patch location across the sampled frames:

$$u_{i,\ell}^{\mathrm{pat}} = \frac{1}{T_i} \sum_{t=1}^{T_i} g_{i,t,\ell}^{\mathrm{pat}}. \tag{7}$$

Here, $u_{i,\ell}^{\mathrm{pat}}$ represents the cross-frame aggregated suspiciousness of the ℓ-th patch location in video x_i. We do not use prototype concentration at the patch level because an individual patch usually contains only partial semantics, making concentration over prototypes substantially less reliable than at the whole-frame level.

Based on the resulting scores, we retain the top-K_{frm} frames according to $\{u_{i,t}^{\mathrm{frm}}\}$ and the top-K_{pat} patch locations according to $\{u_{i,\ell}^{\mathrm{pat}}\}$. The selected patch locations are then organized across the retained frames into cross-frame patch strips:

$$\mathcal{Z}_i = \{z_{i,m}\}_{m=1}^{K_{\mathrm{pat}}}, \tag{8}$$

Table 1: Detection performance on the full test sets of THB, AVLips, and FKAV. We report AP (%) and AUC (%), with the best results in bold and the second-best results underlined. AVH-Align denotes the official checkpoint, and AVH-Align* denotes the version retrained on the AVLips training split. Gains are reported in percentage points relative to the corresponding base detector. We use released checkpoints for supervised baselines and the provided weights for AVAD.

Method	THB		AVLips		FKAV	
	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)
CViT	44.5	42.1	63.5	63.1	91.1	88.5
EfficientViT	31.6	21.7	63.3	64.8	**95.1**	90.9
RealForensics	68.7	74.3	69.9	71.9	94.2	88.2
LipFD	45.0	49.2	85.3	84.7	83.4	77.0
AVAD	43.8	48.1	76.5	73.2	92.1	84.8
AVH-Align	72.6	84.2	76.2	85.8	93.7	93.9
AVH-Align+TFER	**77.5**$_{+4.9}$	**87.4**$_{+3.2}$	**89.6**$_{+13.4}$	<u>87.1</u>$_{+1.3}$	<u>95.0</u>$_{+1.3}$	<u>94.2</u>$_{+0.3}$
AVH-Align*	64.8	82.3	74.3	84.5	93.5	93.0
AVH-Align*+TFER	<u>77.0</u>$_{+12.2}$	<u>87.3</u>$_{+5.0}$	<u>87.5</u>$_{+13.2}$	**89.7**$_{+5.2}$	**95.1**$_{+1.6}$	**94.8**$_{+1.8}$

where each $z_{i,m}$ is formed by the same selected patch location across the retained frames and serves as a localized evidence carrier for subsequent semantic analysis.

The resulting evidence set $\mathcal{Z}_i$ is then fed into Qwen together with a fixed instruction prompt ρ, which explicitly asks the model to describe localized forgery-related evidence rather than directly predicting whether the sample is real or fake. The resulting fine-grained textual description is:

$$q_i = \mathrm{Qwen}(\mathcal{Z}_i, \rho). \tag{9}$$

Here, q_i is treated as an explicit semantic representation of the mined suspicious evidence. To convert it into a comparable ranking signal, we parse q_i into a set of evidence lines $\mathcal{L}(q_i) = \{\ell_{i,n}\}_{n=1}^{N_i}$, where $\ell_{i,n}$ denotes the n-th evidence line in q_i, and N_i is the number of valid lines. We then compare each line against a fake anchor set $\mathcal{A}^f$ and a real anchor set $\mathcal{A}^r$ using a text reranker $h(\cdot,\cdot)$. The line-level margin is defined as:

$$m_{i,n} = \max_{a^f \in \mathcal{A}^f} h(\ell_{i,n}, a^f) - \max_{a^r \in \mathcal{A}^r} h(\ell_{i,n}, a^r). \tag{10}$$

The rank score r_i is computed as the mean line-level margin. A larger r_i indicates higher suspiciousness, and the corresponding video x_i should therefore be assigned to a higher score slot within the routed subset.

Slot-Preserving Reordering. As shown in Figure 3(3), the rank score r_i is not used to define a new detector. It is used only to refine the local ordering of samples in the routed subset $\mathcal{U}$ while preserving the original score structure.

To this end, we collect the original detector scores of all samples in $\mathcal{U}$ and sort them in descending order to form a set of fixed score slots, denoted by $\mathcal{V}_\mathcal{U} = \{v_k\}_{k=1}^{|\mathcal{U}|}$ with $v_1 \geq v_2 \geq \cdots \geq v_{|\mathcal{U}|}$. We then sort the routed samples according to their refinement rank scores and reassign the slots in $\mathcal{V}_\mathcal{U}$ accordingly, so that samples with larger r_i are mapped to higher-valued score slots. In this way, only the relative ordering within $\mathcal{U}$ is refined, while the slot values themselves remain unchanged. For samples in $\mathcal{C}$, the original scores are kept unchanged. For samples in $\mathcal{U}$, the fixed score slots in $\mathcal{V}_\mathcal{U}$ are reassigned according to the descending order of r_i, yielding the final refined scores $\hat{s}_i$. This operation preserves the original score multiset, the scores of samples in $\mathcal{C}$, and the ordering between $\mathcal{C}$ and $\mathcal{U}$. Decisions obtained by thresholding at the routing threshold τ therefore remain unchanged. Because score slots are reassigned within the routed subset of the input collection, the final assigned scores depend on that collection. Our evaluation measures the resulting improvements in anomaly-score ranking.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We conduct experiments on three talking head forgery detection datasets, including the publicly available AVLips (Liu et al., 2024), FakeAVCeleb (FKAV) (Khalid et al., 2022), and TalkingHeadBench (THB) (Xiong et al., 2026). Among them, AVLips is split into training, validation, and test sets with a ratio of 6:1:3. The training set is used to retrain the unsupervised detector, whereas the validation split is used for detector selection during retraining and for threshold estimation. The resulting threshold is used not only to partition the AVLips test set into retained and routed samples, but also to perform the same partition on FKAV and THB, thereby ensuring a consistent routing criterion across different test sets. Following prior work, we use AUC and AP as the main evaluation metrics. AUC reflects the overall discriminative ability of the model, while AP evaluates performance from the precision-recall perspective.

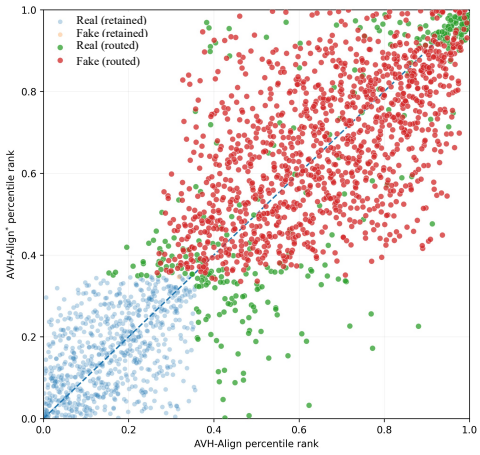


Figure 4: Paired percentile ranks assigned by the official AVH-Align checkpoint and the retrained AVH-Align* on the AVLips test set. Each point denotes the same sample under the two detectors, with color indicating its real/fake label and retained/routed status.

Base Detector Selection. Our goal is not to compare the absolute performance of different unsupervised detectors, but to examine whether TFER can further exploit the latent discriminative information of routed samples in a training-free manner on a detector that remains stable across datasets. Based on this consideration, we adopt the retrained AVH-Align (Smeu et al., 2025), denoted by AVH-Align*, as the default base detector in the main experiments. We also consider another unsupervised detector, AVAD (Feng et al., 2023), but its performance collapses on some datasets. For example, both AUC and AP fall below 50% on THB. It is therefore not suitable as a stable base detector. In addition, we evaluate TFER on top of the official AVH-Align checkpoint to examine whether its gain is sensitive to a particular initialization of the base detector.

Implementation Details. The comparative methods include both supervised and unsupervised talking head forgery detectors. We consider CViT (Wodajo & Atnafu, 2021), EfficientViT (Coccomini et al., 2022), RealForensics (Haliassos et al., 2022), and LipFD (Liu et al., 2024) as supervised baselines, and AVAD (Feng et al., 2023), AVH-Align (Smeu et al., 2025) and AVH-Align* as unsupervised detectors. AVH-Align* is trained using the Adam optimizer on a single NVIDIA A100 GPU, with a learning rate of 9×10^{-4} and a batch size of 1024. We use CLIP ViT-L/14 for evidence mining, Qwen2-VL-7B for fine-grained evidence description, and BGE-Reranker-Large for text-based rank score estimation. All predefined text prompts are generated by GPT-4. *Additional implementation details and extended results are provided in the appendix.*

4.2 EXPERIMENT RESULTS

Cross-Dataset Generalization. Table 1 reports the results on THB, AVLips, and FKAV. Under cross-dataset evaluation, AVH-Align shows more stable base performance than AVAD and outperforms most supervised detectors. It shows that AVH-Align is less tied to specific forgery patterns or dataset biases, and is therefore better suited as the base detector.

To assess whether TFER depends on a particular detector initialization, we evaluate both the official AVH-Align checkpoint and a retrained version AVH-Align*. The retrained model shows weaker base performance, which is expected given its more limited training data. TFER consistently improves both AVH-Align versions across all datasets.

This shows that TFER is not tied to a single initialization, but remains effective across different initial coarse-ranking states. The larger gains on AVH-Align* can be attributed to its weaker initial

score distribution, which leaves greater room for refining routed samples. Accordingly, we use AVH-Align* in the subsequent analysis.

Initial Score Distributions of the Base Detector. Figure 4 compares the percentile ranks between the official AVH-Align checkpoint and the retrained AVH-Align*. Although both checkpoints preserve a usable coarse-grained separation between real and fake samples, they induce noticeably different ranking structures.

This suggests that the official AVH-Align and AVH-Align* correspond to different initial coarse-ranking states. Notably, these deviations are concentrated mainly in routed samples. This means that the main discrepancy between different base detectors does not lie in already well-separated easy samples, but in hard samples.

We can find that the gain of TFER is not tied to a specific initial coarse-ranking state of the detector, since TFER consistently improves both the official AVH-Align checkpoint and AVH-Align* despite their discriminative initial structures.

Table 2: We report AP (%) and AUC (%) on the routed samples selected by score-based routing, where TFER performs evidence-driven refinement. Improvement values indicate the absolute improvement.

Method	THB		AVLips	
	AP (%)	AUC (%)	AP (%)	AUC (%)
AVH-Align*	64.5	36.5	72.2	31.6
with TFER	**77.1**$_{+12.6}$	**55.3**$_{+18.8}$	**89.3**$_{+17.1}$	**67.1**$_{+35.5}$

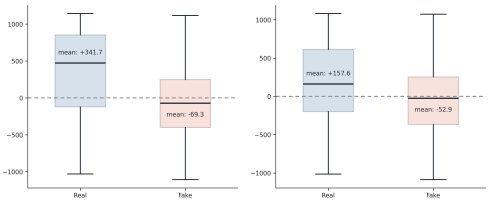


Figure 5: Rank displacement of routed samples on AVLips (left) and THB (right). Negative values indicate that a sample is moved to a more suspicious position after refinement, while positive values indicate that it is moved to a less suspicious position.

Effectiveness on the routed subset. Table 2 evaluates whether TFER effectively improves the samples selected for refinement. After integrating TFER into AVH-Align*, consistent improvements are observed on both THB and AVLips, confirming that evidence-driven refinement effectively enhances the separability of routed samples.

Moreover, the improvements on the routed subset are markedly larger than those on the full test set. This indicates that restricting refinement strictly to the routed subset not only preserves the overall structure of the original detector, but also more fully releases its remaining discriminative potential on hard samples.

Furthermore, we quantify the rank displacement of routed samples before and after TFER, as shown in Figure 5. Negative values indicate that a sample is moved to a more suspicious position after refinement, while positive values indicate that it is moved to a less suspicious position. On both AVLips and THB, real samples exhibit an overall positive shift, whereas fake samples show an overall negative shift. This indicates that TFER tends to push more likely fake samples forward and move more likely real samples backward within the routed subset.

4.2.1 Robustness under Perturbations.

Table 3 reports the detection results under different perturbations. All base detectors degrade under these corruptions, indicating that talking head forgery detection remains challenging under the evaluated perturbations. TFER improves over AVH-Align* under every evaluated perturbation, although absolute performance remains limited under blur. These results show that the gains over the base detector extend beyond clean data. The refinement stage supplements the original anomaly scores with localized visual-semantic information and adjusts ranking within the routed subset.

4.3 Ablation Study

To evaluate the contribution of each component, we conduct ablation experiments on THB and AVLips, with results reported in Table 4.

Table 3: Robustness under different perturbations on THB. We report AP (%) and AUC (%) under the full test sets. Improvement values indicate the absolute improvement brought by TFER over its corresponding base detector.

Method	Inversion		Noise		Blur		Compress	
	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)
AVAD	31.0	20.9	40.1	40.3	39.3	39.3	38.6	38.0
AVH-Align*	41.5	44.0	46.3	53.0	37.4	36.2	38.4	37.9
AVH-Align*+TFER	**67.1**$_{+25.6}$	**72.5**$_{+28.5}$	**67.3**$_{+21.0}$	**74.4**$_{+21.4}$	**55.4**$_{+18.0}$	**50.8**$_{+14.6}$	**55.9**$_{+17.5}$	**63.7**$_{+25.8}$

Effect of Score-based Routing. Removing score-based routing (*w/o Routing*) causes the largest degradation among the reported ablations, demonstrating that TFER benefits from selectively refining routed samples rather than applying evidence-driven refinement to all samples. Without this routing mechanism, the refinement stage is applied to all samples, which can alter the ordering within the retained subset and between the two subsets. These results support the use of selective routing in the proposed framework.

Table 4: Ablation study on the full test sets of THB and AVLips. We report AP (%) and AUC (%), with the best results in **bold**.

Method	THB		AVLips	
	AP	AUC	AP	AUC
w/o Routing	48.9	46.0	50.0	54.6
w/o CLIP-Frame	72.8	85.6	86.1	88.9
w/o CLIP-Patch	71.9	85.1	82.8	87.0
w/o CLIP	70.8	84.6	81.3	86.2
w/o Qwen	69.6	83.8	75.7	83.5
w/o Reranker	74.2	86.1	85.9	88.2
w/o SPR	52.4	46.1	78.1	71.4
Full Model	**77.0**	**87.3**	**87.5**	**89.7**

Effect of Evidence Mining. Removing CLIP-based frame or patch selection (*w/o CLIP-Frame* or *w/o CLIP-Patch*) consistently reduces performance, indicating the contribution of CLIP-guided evidence selection. Frame selection narrows the analysis to informative temporal observations, while patch selection further focuses the refinement process on suspicious local regions. Removing patch selection leads to a larger degradation, suggesting a larger contribution from patch selection under the evaluated configuration. Furthermore, removing CLIP guidance entirely (*w/o CLIP*) results in additional performance drops, supporting the joint use of global frame-level and local patch-level evidence mining for subsequent refinement.

Effect of Semantic Analysis. The contribution of semantic analysis and ranking is further examined by removing Qwen (*w/o Qwen*) and the reranker (*w/o Reranker*). Without Qwen, directly aggregating CLIP-based suspiciousness yields lower performance than the full model, supporting the use of semantic descriptions in the implemented refinement pipeline. Removing the reranker also degrades performance, indicating that text-based scoring contributes to the ranking gains.

Effect of Slot-Preserving Reordering. Removing slot-preserving reordering (*w/o SPR*) leads to a substantial performance drop, supporting its use in the proposed framework. The full method refines the relative ordering within the routed subset by reassigning the original detector score slots, while leaving the scores of retained samples and the ordering between the two subsets unchanged.

5 Conclusion

In this paper, we have proposed TFER, a training-free evidence-driven refinement framework for talking head forgery detection. Built on an existing unsupervised detector, TFER uses score-based routing to select samples for localized visual evidence mining and semantic analysis. It refines their relative ordering through slot-preserving reordering, while keeping the original score multiset, the scores of retained samples, and the ordering between the retained and routed subsets unchanged. Experiments demonstrate improvements over the base detector across the evaluated datasets and perturbation settings.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar checking to improve the clarity and readability of the manuscript. We reviewed all AI-assisted text and revised it where necessary to preserve the intended technical meaning and ensure consistency with our methods and experimental results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The paper and supplementary material describe the proposed method, experimental protocols, and implementation details, including model configurations, key hyperparameters, and prompts. The key code is provided in the supplementary material to facilitate reproducibility. The complete code and relevant model checkpoints will be made publicly available after the review process concludes.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Wei Cai, Shujuan Liu, Jian Zhao, Ziyan Shi, Yusheng Zhao, Yuchen Yuan, Tianle Zhang, Chi Zhang, and Xuelong Li. When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35322–35330, 2026.

Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Pengfei Wan, Di ZHANG, Shuai Li, et al. Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025.

Sungik Choi, Hankook Lee, and Moontae Lee. Training-free detection of ai-generated images via cropping robustness. *Advances in Neural Information Processing Systems*, 38:20587–20610, 2026.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 439–447, 2020.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pp. 219–229. Springer, 2022.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European conference on computer vision*, pp. 54–72. Springer, 2024.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8498–8507, 2024.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10491–10503, 2023.

Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472, 2005.

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11736–11745, 2025.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

Midou Guo, Qilin Yin, Wei Lu, and Xiangyang Luo. Towards open-world generalized deepfake detection: General feature extraction via unsupervised domain adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11572–11580, 2025.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14950–14962, 2022.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection, 2024. URL https://arxiv.org/abs/2405.20112.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4490–4499, 2023.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022. URL https://arxiv.org/abs/2108.05080.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 3677–3685, 2017.

Ivan Kukanov and Jun Wah Ng. Klassify to verify: Audio-visual deepfake detection using ssl-based audio and handcrafted visual features. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13707–13713, 2025.

Xiaolou Li, Zehua Liu, Chen Chen, Lantian Li, Li Guo, and Dong Wang. Zero-shot fake video detection by audio-visual consistency. In *Proc. Interspeech 2024*, pp. 2935–2939, 2024.

Ke Liu, Jiwei Wei, Shiyuan He, Zeyu Ma, Chaoning Zhang, Ning Xie, and Yang Yang. Syncgaussian: Stable 3d gaussian-based talking head generation with enhanced lip sync via discriminative speech features. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1576–1584, 2025.

Ke Liu, Jiwei Wei, Wenyu Zhang, Shuchang Zhou, Ruikun Chai, Yutao Dai, Chaoning Zhang, and Yang Yang. From talking to singing: A new challenge for audio-visual deepfake detection. In *Forty-third International Conference on Machine Learning*, 2026.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Advances in Neural Information Processing Systems*, 37:91131–91155, 2024.

Yang Liu, Zhaoyang Xia, Mengyang Zhao, Donglai Wei, Yuzheng Wang, Siao Liu, Bobo Ju, Gaoyun Fang, Jing Liu, and Liang Song. Learning causality-inspired representation consistency for video anomaly detection. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 203–212, 2023.

Jianzhi Long, Wenhao Sun, Rong-Cheng Tu, and Dacheng Tao. Lightning fast caching-based parallel denoising prediction for accelerating talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7503–7511, 2026.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832, 2020.

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193, 2019.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 666–676, 2024.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

Yassine Rachidy, Youssef Hmamouche, Faissal Sehbaoui, and Amal El Fallah Seghrouchni. Domain adaptive document reranking for retrieval augmented generation. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1381–1387. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Katharine Sanderson. Gpt-4 is here: what scientists think. *Nature*, 615(7954):773, 2023.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18815–18825, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Chung-Ting Tsai, Ching-Yun Ko, I-Hsin Chung, Yu-Chiang Frank Wang, and Pin-Yu Chen. Understanding and improving training-free ai-generated image detections with vision foundation models, 2024. URL https://arxiv.org/abs/2411.19117.

Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14653–14662, 2023.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Tianyi Wang, Mengxiao Huang, Harry Cheng, Xiao Zhang, and Zhiqi Shen. Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10515–10524, 2024b.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Jiwei Wei, Yang Yang, Xing Xu, Jingkuan Song, Guoqing Wang, and Heng Tao Shen. Less is better: Exponential loss for cross-modal matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5271–5280, 2023.

Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer, 2021. URL https://arxiv.org/abs/2102.11126.

Jiajian Xie, Shengyu Zhang, Mengze Li, Zhou Zhao, Fei Wu, et al. Ecoface: Audio-visual emotional co-disentanglement speech-driven 3d talking face generation. In *International Conference on Learning Representations*, volume 2025, pp. 101625–101643, 2025.

Xinqi Xiong, Prakrut Patel, Qingyuan Fan, Amisha Wadhwa, Sarathy Selvam, Xiao Guo, Luchao Qi, Xiaoming Liu, and Roni Sengupta. Talkingheadbench: A multi-modal benchmark & analysis of talking-head deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4139–4149, 2026.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

Yutong Yang, Lifu Huang, Yijie Lin, Xi Peng, and Mouxing Yang. Endowing vision-language models with system 2 thinking for fine-grained visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11802–11810, 2026.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *International Conference on Machine Learning*, pp. 72925–72943. PMLR, 2025.

Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15044–15054, 2021.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14800–14809, 2021.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

A Appendix

A.1 Additional Experimental Details

A.1.1 Comparison Methods

We consider both supervised and unsupervised forgery detectors as baselines in our paper. The unsupervised methods, namely AVAD (Feng et al., 2023) and AVH-Align (Smeu et al., 2025), are trained exclusively on real talking head videos. By contrast, the supervised baselines are trained on both authentic and manipulated talking head samples, including CViT (Wodajo & Atnafu, 2021), RealForensics (Haliassos et al., 2022), EfficientViT (Coccomini et al., 2022), and LipFD (Liu et al., 2024).

AVAD is an audio-visual deepfake detection method that evaluates temporal inconsistencies between speech and facial movements. By examining whether the observed desynchronization follows natural patterns or shows abnormal deviations, it determines the authenticity of a video. AVH-Align identifies a common artifact in forged videos, namely leading silent segments, which can unintentionally bias existing detection models. To address this issue, it introduces an unsupervised alignment strategy that prevents models from relying on such dataset-specific cues, thereby improving the robustness to a wider range of manipulations.

Among supervised detection models, CViT is a visual deepfake detector that integrates convolutional neural networks with Transformer attention, allowing it to model both subtle local artifacts and broader structural patterns in manipulated frames. RealForensics is built on the natural alignment between visual and audio modalities in authentic videos. Using an unsupervised cross-modal learning scheme, it acquires temporally dense video representations that capture facial motion, expressions, and identity-related information. EfficientViT uses a hybrid design that combines several vision transformer variants with a lightweight EfficientNet-B0 backbone for feature extraction, providing strong video forgery detection capability with low computational cost. LipFD is a supervised audio-visual detector tailored to talking head scenarios, where manipulated content is detected by capturing inconsistencies between the audio stream and the associated lip movements.

A.1.2 Talking Head Forgery Datasets

We perform experiments on three representative deepfake datasets, namely FakeAVCeleb (FKAV) (Khalid et al., 2022), AVLips (Liu et al., 2024), and TalkingHeadBench (THB) (Xiong et al., 2026). These datasets cover typical data distributions and synthesis paradigms commonly encountered in talking head videos.

The FKAV dataset contains 500 real videos collected from VoxCeleb2 (Chung et al., 2018) and 19,500 forged videos produced with a range of manipulation methods. These forgeries include visual manipulations generated by Faceswap (Korshunova et al., 2017), FSGAN (Nirkin et al., 2019), and Wav2Lip (Prajwal et al., 2020), together with audio manipulations synthesized by SV2TTS (Jia et al., 2018). To alleviate the imbalance between authentic and forged samples, we build a test set using 500 real videos, which are excluded from all training sets, along with 1,000 fake videos created by diverse forgery techniques. AVLips is a large-scale audio-visual lip-sync dataset developed

Table 5: Corruption settings used in robustness evaluation.

Corruption Type	Parameter	Range	Value
JPEG Compression	quality	0–100	20
Gaussian Blur	ksize	odd integer	5
Gaussian Noise	σ	≥ 0	25
Color Inversion	inv	N/A	N/A

specifically for talking head forgery detection. It includes around 340,000 audio-video samples generated by different lip-sync methods. To produce realistic lip movements, the dataset covers both static generation approaches, such as MakeItTalk (Zhou et al., 2020), and dynamic methods, including Wav2Lip, TalkLip (Wang et al., 2023), and SadTalker (Zhang et al., 2023). Its real samples are mainly collected from LRS3 (Afouras et al., 2018), FaceForensics++ (Rossler et al., 2019), DFDC (Dolhansky et al., 2020), and other real-world sources. The complete dataset is not publicly released. Instead, the authors make available a subset consisting of real samples from LRS3 and synthetic videos generated by unspecified GAN-based methods. We divide this subset into training, validation, and test sets with a ratio of 6:1:3.

THB is a recently proposed dataset for audio-visual forgery detection, mainly built upon diffusion-based generation techniques. It creates manipulated videos by pairing portrait images from FFHQ (Karras et al., 2019) with driving signals from CelebV-HQ (Zhu et al., 2022). The dataset covers samples generated by six representative methods, including Hallo (Xu et al., 2024), Hallo2 (Cui et al., 2024), AniPortrait in both Audio-driven and Video-driven settings (Wei et al., 2024), LivePortrait (Guo et al., 2024), and EMOPortraits (Drobyshev et al., 2024). In addition, THB contains commercial samples produced by MAGI-1 (Teng et al., 2025), which is also a diffusion-based tool. The official repository releases separate training, validation, and test splits for each generation method. For evaluation, we merge all test samples produced by diffusion-based methods into a single unified test set. More specifically, this set contains 117 videos from Hallo, 105 from Hallo2, 146 from AniPortrait (Audio-driven), 108 from AniPortrait (Video-driven), and 66 commercial samples from MAGI-1.

A.1.3 Perturbation Settings

Videos collected in real-world scenarios are frequently affected by various corruptions, which make reliable deepfake detection more challenging. A robust detector should not only exhibit strong cross-dataset generalization but also remain effective under common perturbations when identifying manipulated content. To examine this, we evaluate model performance under four representative types of perturbation:

1. **JPEG Compression** *(Compress)*: Emulates degradation caused by video encoding by introducing compression artifacts and the loss of visual details. As the compression level increases, image quality deteriorates and distortions become more noticeable.

2. **Gaussian Blur** *(Blur)*: Attenuates high-frequency information, which may obscure subtle lip motions and facial expressions. Using larger kernel sizes leads to stronger blurring effects.

3. **Gaussian Noise** *(Noise)*: Simulates sensor imperfections or transmission disturbances by injecting random pixel-level perturbations. A larger standard deviation (σ) corresponds to a higher noise intensity.

4. **Color Inversion** *(Inversion)*: Modifies visual appearance by inverting pixel intensities while keeping the spatial layout unchanged, thereby evaluating the model's robustness to extreme color variations.

The configurations for each perturbation are summarized in Table 5.

Table 6: Key settings of TFER. All configurable parameters are selected on the validation set and then fixed for all test sets, without any test set tuning.

Component	Setting
CLIP backbone	ViT-L/14
CLIP pretrained checkpoint	`laion2b_s32b_b82k`
Qwen model	Qwen2-VL-7B-Instruct
Reranker model	bge-reranker-large
Sampled frames per video	12
Maximum candidate views	12
Patch partition	3×3 grid
Patch size	192
Temporal strip gap	4
Top views sent to Qwen	5
Top suspicious patches sent to Qwen	3
CLIP prototype temperature	0.02
Qwen max new tokens	96
Reranker evidence aggregation	mean line-level margin

A.1.4 Inference Efficiency

Since TFER is a training-free framework, we report only its inference overhead. The score-based routing stage is highly efficient. Within this stage, the main cost comes from the base detector forward pass, which takes approximately 0.02 s per sample. The subsequent threshold-based routing is negligible, requiring only 0.001 ms per sample in our implementation, with no additional GPU memory.

Evidence-driven refinement is activated only for the routed subset. We measure its inference efficiency on routed samples using an NVIDIA A100 GPU. Averaged over 200 routed samples, CLIP-based evidence mining, Qwen-based semantic analysis, and reranker-based score estimation take 2.38 s, 3.05 s, and 0.24 s per sample, respectively. These three stages together take 5.67 s per routed sample. The average peak GPU memory is 21.00 GB. These results indicate that TFER is better suited for offline verification or high-risk auditing than for edge-device or strict real-time deployment. These results are consistent with the design of TFER. Score-based routing remains lightweight and suitable for fast coarse screening, while the additional cost is concentrated in evidence-driven refinement and incurred only on routed samples. This selective activation makes it possible to introduce fine-grained large-model semantic analysis without turning it into a universal cost for all test samples.

A.1.5 Key Hyperparameters

For score-based routing, we do not introduce manually tuned model hyperparameters. Instead, score-based routing uses the original detector scores and estimates a routing threshold on the validation set using the Youden index. The resulting threshold is then fixed and directly applied to the test set, without any test set tuning.

For evidence-driven refinement, we use CLIP ViT-L/14 (Radford et al., 2021) for evidence mining, Qwen2-VL-7B-Instruct (Wang et al., 2024a) for vision-language semantic analysis, and BAAI/bge-reranker-large (Rachidy et al., 2025) for text-based evidence scoring. For each video, we sample 12 frames and keep at most 12 candidate views for CLIP analysis. Spatial evidence is constructed using a $3\times$ 3 patch partition, with the patch size set to 192 and the temporal strip gap set to 4. Based on CLIP scores, the top 5 views and top 3 suspicious patches are selected and sent to Qwen. The CLIP prototype softmax temperature is set to 0.02. Qwen uses a maximum generation length of 96 tokens and bfloat16 inference. In the reranker stage, we use the mean line-level margin as the refinement rank score r_i. All configurable parameters are selected on the validation set and then fixed for all test sets. The detailed parameter settings are shown in Table 6.

A.1.6 PROMPTS IN EVIDENCE-DRIVEN REFINEMENT

We provide the full prompts used in evidence-driven refinement for reproducibility. All prompts are determined on the validation set and then kept fixed for all test sets, without any test set tuning. Specifically, the prompt set used for CLIP-based evidence mining is denoted as **Prompt A**, the instruction used for Qwen-based semantic analysis is denoted as **Prompt B**, and the anchor texts used in the reranker are denoted as **Prompt C**. All prompts are generated by GPT-4 (Sanderson, 2023), and their detailed contents are given below.

Prompt A: CLIP text prototypes. We use paired real/fake text prototypes to score video frames and local facial patches. The full prototype list is given below.

```
1. skin_natural, facial skin texture looks natural with
preserved fine details, Real
2. skin_smooth, facial skin texture looks over-smoothed or
plasticky, Fake

3. mouth_clear, mouth region looks clear with clean lip
contour and visible inner-mouth details, Real
4. mouth_smeared, mouth region looks smeared or muddy with
unclear inner-mouth details, Fake

5. teeth_clean, teeth or inner-mouth appearance looks clear
and well defined, Real
6. teeth_unstable, teeth or inner-mouth appearance looks
unstable, blurred, or inconsistent, Fake

7. boundary_clean, facial boundary looks clean and well
separated from the background, Real
8. boundary_blended, facial boundary looks blended, broken,
torn, or mixed with the background, Fake

9. structure_natural, facial structure looks natural and
well aligned, Real
10. structure_warped, facial structure looks warped,
distorted, or spatially inconsistent, Fake

11. eyes_clear, eye region looks clear with natural contour
and preserved fine details, Real
12. eyes_unstable, eye region looks blurred, distorted, or
visually unstable, Fake

13. detail_preserved, fine details look preserved and
visually coherent, Real
14. detail_degraded, fine details look degraded, blurred,
or visually inconsistent, Fake
```

Prompt B: Qwen instruction prompt. We use a fixed instruction prompt to guide Qwen to describe only fine-grained visible evidence in suspicious facial regions, without directly judging authenticity. The full instruction is given below. In the prompt, $patch_text$ denotes the indices of the suspicious patches selected by CLIP.

```
You are given facial patch strips selected as suspicious by
a coarse detector.

The suspicious patch ids are: {patch_text}.

Each strip shows the SAME facial patch collected from
multiple different views or frames, arranged from left to
right.

Different strips correspond to different suspicious patches.
```

```
Describe only fine-grained visible evidence in these
regions.

Focus on visible artifacts such as:  over-smoothed or
plasticky skin, smeared lip contour, unclear inner-mouth
details, unstable teeth appearance, blurred, distorted, or
visually unstable eye regions, boundary blending or tearing,
warped structure, misplaced highlights, blur, and incon
-sistent fine details.


For normal-looking evidence, use direct positive
descriptions such as:  clear mouth details, stable contour,
coherent texture, preserved detail, clean boundary, natural
highlights.

Do NOT use reverse-negative phrasing such as ``no flicker'',
``not blurred'', or ``not inconsistent''.

Do NOT judge authenticity directly.

Do NOT use words such as real, fake, forged, synthetic, or
deepfake.


Return 3 short bullet points, one evidence item per line.
```

Prompt C: Reranker anchor texts. We use predefined benign and artifact-oriented anchor texts to score the generated evidence description in the reranker stage. The full anchor texts are given below.

```
Real anchors:
1.  The evidence describes clear mouth details, preserved
fine texture, and natural facial appearance.
2.  The evidence indicates stable contour, coherent texture,
clean boundary, and natural highlights.
3.  The description suggests normal facial structure with
clear details, natural eye regions, and visually consistent
appearance.
4.  The text supports a benign interpretation with natural
texture, clear inner-mouth details, and clean separation.


Fake anchors:
1.  The evidence describes over-smoothed or plasticky skin,
degraded fine details, and abnormal facial appearance.
2.  The evidence indicates smeared mouth details, unstable
teeth appearance, broken boundary, and misplaced highlights.
3.  The description suggests warped structure, distorted
eye regions, blurred texture, and visually inconsistent
appearance.
4.  The text supports an artifact-heavy interpretation with
smeared details, blending, distortion, and degraded texture.
```

A.2 Extended Experiments

A.2.1 CLIP Text Prototype Configurations

We further study the effect of CLIP text prototype design in evidence-driven refinement by comparing three configurations with 4, 7, and 10 prototype pairs, respectively. Here, 7 prototype pairs correspond to the default setting used in TFER, while the 10-pair setting additionally includes *lip_sync_natural and lip_sync_misaligned*, *region_coherent and region_incoherent*, and *expression_natural and expression_inconsistent*. The results are reported in Table 7. For reference, we also include the retrained AVH-Align* and our final AVH-Align*+TFER result.

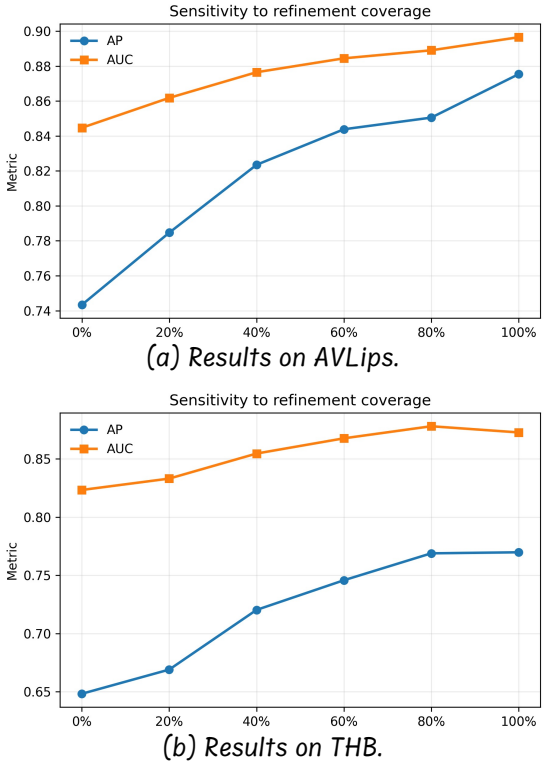


Figure 6: Sensitivity of TFER to refinement coverage within the routed subset. Starting from the fixed partition obtained by score-based routing, we increase the fraction of routed samples selected for refinement from 0% to 100%, while keeping the scores of retained samples unchanged. AP and AUC are reported on the full test sets.

Several observations can be drawn from these results. The CLIP text prototype configuration has a clear influence on the final performance, confirming that prompt design is an important factor in evidence mining. This is consistent with the role of CLIP in our framework, where CLIP is not used as an open-ended classifier but instead serves as a coarse evidence localizer that guides subsequent semantic analysis over suspicious regions. Therefore, the text prototypes directly affect whether the mined evidence is semantically relevant and discriminative for talking head forgery detection.

Using fewer text prompts leads to consistently weaker performance, especially on AVLips, suggesting that an overly small prototype set cannot provide sufficient semantic coverage for diverse forgery artifacts. In contrast, simply introducing more text prompts does not lead to uniformly better results. Although the larger prompt set achieves the best performance on THB, it does not generalize equally well to AVLips, where its performance is lower than that of the default TFER configuration. This indicates that prompt quantity alone is not the key factor. Adding more prototypes may increase semantic coverage, but it may also introduce redundancy or competition among related descriptions, which can weaken the discriminative focus of CLIP-based evidence mining.

Compared with the retrained AVH-Align baseline, all variants that introduce training-free evidence-driven refinement achieve clear improvements on both datasets, regardless of the specific CLIP text prototype configuration. These results show that the gains over the retrained base detector persist across the evaluated CLIP prototype configurations. The default configuration achieves the best results on AVLips while retaining improvements over the base detector on THB.

A.2.2 Sensitivity to Refinement Coverage within the Routed Subset

To study the effect of refinement coverage, we vary the fraction of routed samples selected for refinement from 0% to 100%, while keeping the original score-based routing partition and the scores

Table 7: Effect of different CLIP text prototype configurations in TFER. We report AP(%) and AUC(%). AVH-Align* denotes the retrained baseline, and AVH-Align*+TFER denotes our final method. Bold indicates the best result, and underlining indicates the second-best result.

Method	THB		AVLips	
	AP	AUC	AP	AUC
Fewer text prompts	74.1	87.0	<u>83.7</u>	<u>87.6</u>
More text prompts	**77.8**	**88.1**	83.0	87.4
AVH-Align*	64.8	82.3	74.3	84.5
AVH-Align*+TFER	<u>77.0</u>	<u>87.3</u>	**87.5**	**89.7**

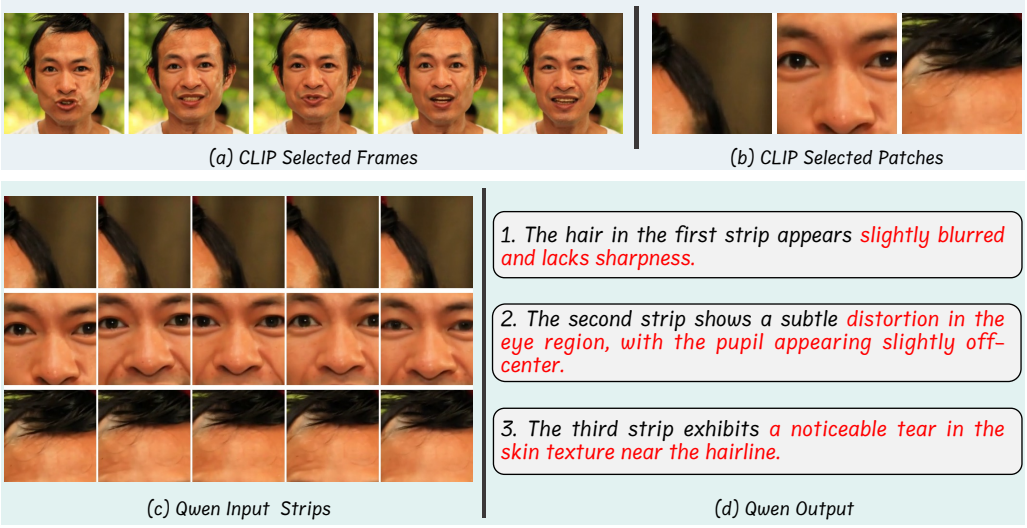


Figure 7: Qualitative example of the evidence mining and semantic analysis process in evidence-driven refinement. (a) Representative frames selected by CLIP. (b) Suspicious local patches identified by CLIP. (c) Cross-frame patch strips constructed from the selected patches and sent to Qwen. (d) Fine-grained evidence descriptions generated by Qwen.

of retained samples unchanged. For each coverage ratio, samples are selected using the refinement ranking signal, and slot-preserving reordering is recomputed accordingly. This experiment evaluates how the extent of refinement within a fixed routed subset affects the final performance.

The results on both AVLips and THB in Figure 6 show that AP and AUC improve as the refinement coverage increases. On AVLips, performance rises from 0.743/0.845 without refinement to 0.875/0.897 when the full routed subset is refined. On THB, AP/AUC similarly increase from 0.648/0.823 to 0.770/0.873. These results demonstrate that applying evidence-driven refinement to a larger fraction of the routed subset improves overall ranking under the evaluated coverage settings.

A.3 Qualitative Results

Figure 7 presents a representative example of the evidence mining and semantic analysis process in evidence-driven refinement. For a sample selected by score-based routing, CLIP selects representative frames and suspicious local patches for subsequent semantic analysis. As shown in Figure 7(a), the selected frames remain comparable as instances of the same face while also covering informative appearance variations, which provides a stable basis for subsequent local evidence analysis. On this basis, CLIP further identifies the three suspicious patches shown in Figure 7(b). These local patches are then reorganized into the cross-frame patch strips shown in Figure 7(c), allowing Qwen

to inspect fine-grained variations of the same local region across multiple frames in a compact and directly comparable format.

The evidence descriptions generated by Qwen in Figure 7(d) further illustrate the role of evidence-driven refinement. For the first strip, Qwen describes the hair region as "slightly blurred and lacks sharpness." For the second strip, Qwen reports "distortion in the eye region, with the pupil appearing slightly off-center." For the third strip, Qwen describes "a noticeable tear in the skin texture near the hairline." These examples illustrate how localized visual content is represented through fine-grained textual descriptions. The reranker uses these descriptions as semantic cues to refine the relative ordering of routed samples.

A.4 Discussion and Future Challenges

Challenge 1. The current TFER uses a fixed refinement budget for all samples routed to evidence-driven refinement, including the number of selected views, suspicious patches, and the overall evidence organization format. This design is simple, stable, and suitable for controlled evaluation. However, the ambiguity level and evidence distribution within the routed subset are not uniform. Some hard samples may be effectively refined with only limited local evidence, whereas others may require richer cross-view or cross-region cues. Therefore, while the present framework has verified the effectiveness of revisiting only the routed subset, it does not yet distinguish the refinement demand among different routed samples.

Future Work 1. A natural next step is to introduce a more adaptive refinement budget allocation mechanism while preserving the training-free property of TFER. For example, the number of selected views, suspicious patches, or analysis depth may be adjusted according to the sample's position in the base detector's anomaly-score ranking, the initial CLIP evidence strength, or the consistency of Qwen-generated descriptions. In this way, evidence-driven refinement can allocate more computation to routed samples requiring more detailed analysis while using less computation on simpler cases.

Challenge 2. The current evidence-driven refinement is designed to provide a complementary vision-language perspective for revisiting routed samples, and therefore its refinement process mainly relies on localized visual evidence. This is consistent with the motivation of the paper, since the original unsupervised detector has already modeled audio-visual consistency, while evidence-driven refinement is introduced to provide additional fine-grained visual evidence analysis. Nevertheless, from a broader perspective, the current framework does not explicitly exploit richer internal signals from the original detector, such as detector-native local responses, modality-specific confidence differences, or more detailed anomaly structures. In other words, the current TFER has demonstrated that training-free evidence refinement can correct the ordering of hard samples without retraining the detector, but there is still room to better connect the refinement stage with the detector's internal discriminative information.

Future Work 2. An important future direction is to explore a tighter detector-reasoner collaboration while still maintaining the training-free setting. For instance, the local anomaly responses, modality uncertainty, or finer-grained score structures from the original detector may be used as additional priors to guide evidence selection and reasoning in evidence-driven refinement. This does not mean training a new joint model, but rather making fuller use of the detector's existing internal signals during inference, so that large-model semantic analysis can operate on more informative and detector-aware evidence.

Overall, our results support the effectiveness of training-free evidence-driven refinement, while motivating future work on adaptive computation allocation and closer integration with the base detector.