HIM-CoT: Reasoning over the Human-Instrument-Music Chain for Musical Performance Deepfake Detection

Anonymous authors
Paper under double-blind review

Abstract

Audio-visual deepfake detection increasingly exploits cross-modal correspondence, yet existing methods predominantly target talking or singing heads, where detection relies on speech–lip synchronization and facial artifacts. However, this formulation breaks down for instrumental performances. Sound production is mediated by physical interaction with an instrument. Human motion drives instrument state, which in turn generates music. This production chain imposes constraints that face-centric detectors cannot exploit. We therefore study the detection and attribution of synthesized modalities in instrumental performances, using consistency with the production chain as evidence: *given structured measurements of music, motion, and performer–instrument contact, can a model determine whether, and in which modality, a performance has been synthesized*? To address this question, we first construct HIM-DF, a benchmark with bidirectional forgeries that isolate audio and visual inconsistencies within the same Human–Instrument–Music structure. We then propose HIM-CoT, a reasoning framework that converts structured performance measurements into queries over pretrained multimodal representations. By representing which measurements are available, modality-specific evidence, and cross-modal discrepancies, HIM-CoT grounds three-stage chain-of-thought generation in forensic cues rather than end-to-end feature fusion. Experiments on HIM-DF show that HIM-CoT detects and attributes forged modalities more accurately than audio-visual, CoT-based, and fine-tuned multimodal baselines, and that its explanations agree with its decisions and cite the measured evidence accurately.

1 Introduction

High-fidelity generative models can now synthesize realistic audio and video of human performances, challenging content authenticity verification. Musical performances, where a visible performer interacts with an instrument to produce sound, are a rich domain for cross-modal forgery detection because structured physical constraints govern the relationship between motion and sound. Judging whether a recorded performance is genuine can therefore draw not only on the plausibility of each modality in isolation, but also on whether the observed actions coherently explain the accompanying music.

Detection becomes challenging when each modality is individually convincing: synthesized video can show natural hand movements over authentic audio, and authentic footage can be paired with plausible synthetic music. Sound follows a chain in which actions change the instrument's state, which generates music, so a forged video may show smooth movements that fail to explain the pitch, register, or dynamics of the sound, and synthesized audio may be inconsistent with the visible instrument state. Evidence about whether the performance coherently produces the sound can therefore complement single-modality artifact cues.

We use "production-chain consistency" to denote whether the observed performance is broadly compatible with the accompanying music in terms of acoustics, motion, contact, and cross-modal evidence. We do not assume reliable frame-level verification of action-to-sound timing.

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.

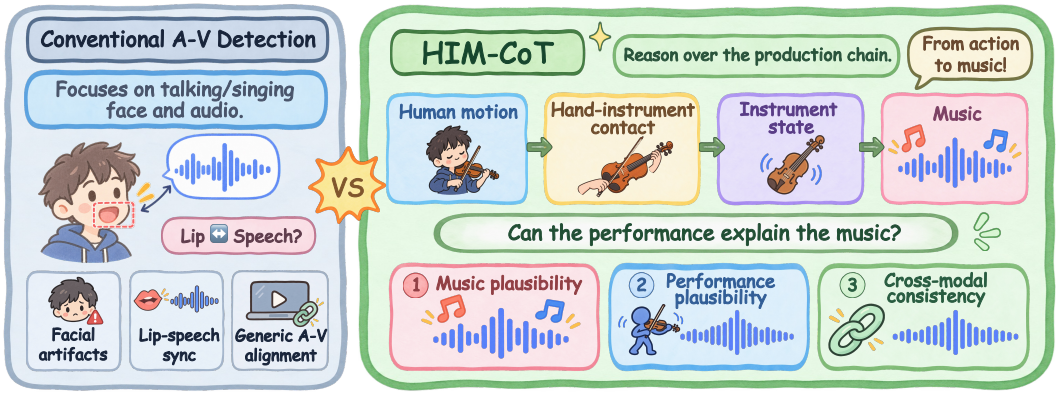

Figure 1: **Comparison with existing methods.** Existing detectors (left) learn generic audio-visual alignment through end-to-end fusion, conflating evidence sources. Our approach (right) grounds detection in structured forensic measurements that query modality-specific representations and generate interpretable three-stage reasoning.

Existing audio-visual deepfake detectors largely sidestep this question. Early methods for talking faces exploited speech–lip synchronization (Haliassos et al., 2022), while recent approaches learn generic audio-visual alignment through unsupervised or contrastive objectives (Smeu et al., 2025; Peng et al., 2026; Kim et al., 2026). These approaches are ill-suited for instrumental performances where audio-visual coupling is mediated by physical interaction: the same hand position can produce different sounds depending on instrument state, and contact dynamics affecting sound are not directly visible. They also conflate evidence sources into learned embeddings, obscuring whether an inconsistency lies in the music, the motion, or their relation.

We argue that this setting calls for three departures from prevailing frameworks: grounding detection in structured measurements of the production process (acoustics, motion kinematics, and performer–instrument contact), using these measurements to query pretrained representations so the same cue can be compared across audio and video, and reasoning in stages that mirror forensic analysis, from music plausibility to performance plausibility to cross-modal consistency.

We propose HIM-CoT, a reasoning framework built on Qwen2.5-Omni-7B (Xu et al., 2025) that projects structured performance measurements into query tokens attending to pretrained audio and video representations. The difference between the modality-specific responses forms an explicit cross-modal disagreement signal, supplemented by an auxiliary comparison of their temporal attention, that guides three-stage chain-of-thought generation. We construct HIM-DF, a benchmark with bidirectional forgeries isolating audio and visual inconsistencies within the Human–Instrument–Music structure. Our contributions are: **(1)** showing that structured production-chain measurements provide useful forensic cues, alongside synthesis artifacts, for detecting and attributing synthesized modalities in instrumental performances, **(2)** an evidence-to-media query mechanism that retrieves modality-specific responses and models the discrepancy between them, and **(3)** HIM-DF, a bidirectionally forged benchmark isolating failure modes within a shared Human–Instrument–Music structure. Experiments show that explicit evidence grounding improves both accuracy and interpretability over end-to-end fusion baselines.

2 Related Work

2.1 Audio-Visual Deepfake Detection

Audio-visual deepfake detection has evolved from unimodal analysis toward exploiting cross-modal correspondence. RealForensics learns temporally dense representations through self-supervised audio-visual correspondence (Haliassos et al., 2022), while recent unsupervised methods emphasize modality alignment: AVH-Align aligns representations trained on real samples (Smeu et al., 2025), HAVIC models intra-modal and inter-modal coherence (Peng et al., 2026), and X-AVDT exploits generator-internal cross-attention (Kim et al., 2026). Existing detection remains largely face-centric: SHDF extends to singing heads using rhythm-aware generation (Liu et al., 2026). HIM-CoT instead

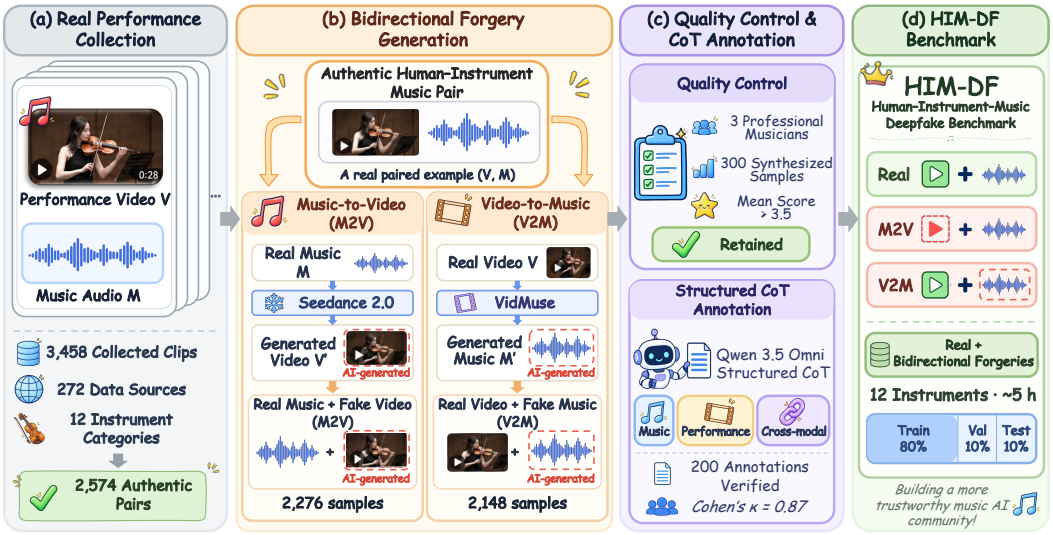

Figure 2: Overview of the HIM-DF benchmark construction and annotation pipeline.

studies instrument-mediated production, grounding reasoning in acoustic, kinematic, and contact measurements rather than learned embeddings.

2.2 Audio-Visual Deepfake Benchmarks

Existing audio-visual deepfake benchmarks predominantly focus on faces. FaceForensics++, DFDC, and Celeb-DF provide large-scale face-swap and reenactment data across diverse manipulation methods and in-the-wild scenarios (Rössler et al., 2019; Dolhansky et al., 2020; Li et al., 2020), FakeAVCeleb and LAV-DF pair synthesized talking faces with speech (Khalid et al., 2021; Cai et al., 2023), and SHDF extends to singing heads but remains face-centric (Liu et al., 2026). Music benchmarks, in turn, are audio-only: SingFake targets multilingual singing voices (Zang et al., 2024), SONICS scales to commercial music generators (Rahman et al., 2025), and M6 broadens coverage across generators, genres, and instruments (Li et al., 2026). No existing benchmark studies instrumental performance forgeries mediated by physical instrument interaction, and HIM-DF fills this gap with bidirectional forgeries that isolate audio and visual inconsistencies.

2.3 Multimodal Reasoning and Chain-of-Thought

Chain-of-thought reasoning improves performance on complex tasks by decomposing problems into intermediate steps (Wei et al., 2022; Zhang et al., 2023). Multimodal foundation models such as LLaVA-NeXT (Li et al., 2025), Gemini 1.5 (Gemini Team, 2024), InternVL 2.0 (Chen et al., 2026), and Qwen2-VL (Wang et al., 2024) have substantially advanced vision-language understanding and reasoning (Li et al., 2024), and vision-language models have been used to generate explanations alongside forgery predictions (Yang et al., 2024). These explanations, however, are not grounded in physical measurements. HIM-CoT instead uses acoustic, kinematic, and contact measurements as queries, tying each reasoning stage to retrievable evidence.

3 HIM-DF Benchmark

We construct HIM-DF, a musical-performance deepfake benchmark with approximately 7K samples (Figure 2). We collect 3,458 real videos from 272 sources covering 12 instrument categories, yielding 2,574 authentic samples after quality filtering.

3.1 Bidirectional Synthesis and Annotation

To construct musical-performance deepfakes, we adopt a bidirectional synthesis strategy that manipulates either the performance video or the music while preserving the other as authentic. For the Music-to-Video (M2V) subset, authentic instrumental music is used to generate corresponding performance videos with Seedance 2.0 (ByteDance Seed Team, 2026), resulting in 2,276 samples with real music and generated performances. For the Video-to-Music (V2M) subset, we retain authentic

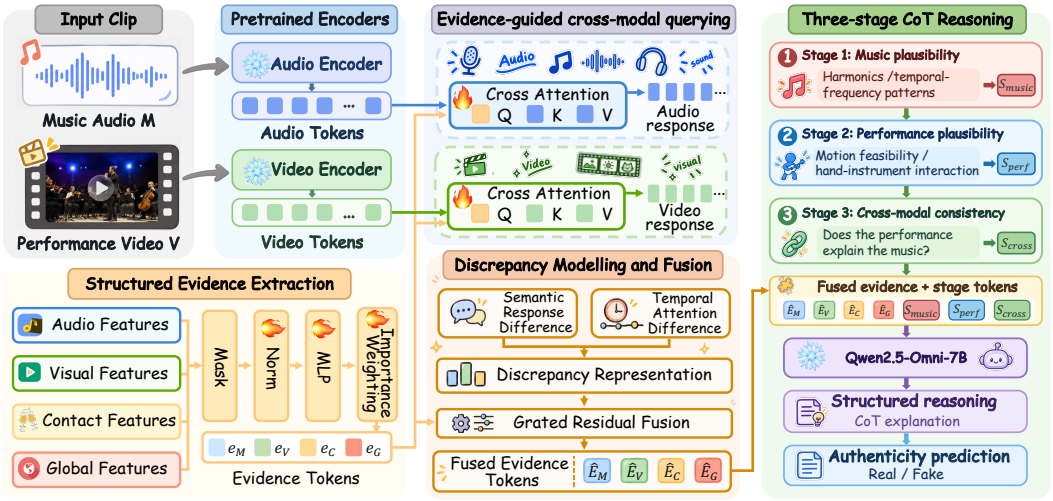

Figure 3: Overview of the HIM-CoT framework with three modules: evidence grounding projects structured measurements into tokens querying pretrained representations, cross-modal discrepancy modeling retrieves modality-specific responses and computes their discrepancy, with temporal attention differences as an auxiliary signal, and three-stage reasoning sequentially assesses music plausibility, performance plausibility, and cross-modal consistency.

performance videos and employ VidMuse to synthesize the corresponding music (Tian et al., 2025), producing 2,148 samples. Together with the original real pairs, HIM-DF covers real performances, visual-performance forgeries, and music forgeries within a unified Human–Instrument–Music setting. Labels encode provenance: a clip is fake whenever one modality is synthesized, however plausible the result. Consistency with the production chain is thus not the label but the evidence HIM-CoT uses to detect and attribute synthesized modalities.

Quality Control and Human Verification All generated samples pass automated screening, and three professional musicians rate a stratified audit set of 300 synthesized samples on visual realism, audio quality, and perceptual plausibility (5-point scale), discarding rated samples with a mean score of at most 3.5 and calibrating the automated thresholds. Two expert annotators verify 200 Chain-of-Thought annotations generated by Qwen 3.5 Omni Plus (Team, 2025) (Cohen's $\kappa = 0.87$ (Cohen, 1960)), and the errors they identify are used to refine the annotation prompt before the full corpus is regenerated. All samples receive structured Chain-of-Thought annotations examining music authenticity, visual performance authenticity, and cross-modal consistency in JSON format, providing explicit supervision for the reasoning module. The benchmark is split into training, validation, and test sets with source-based separation. Full construction details, statistics, the quality-control and annotation protocols, and an analysis of what single-modality detectors can achieve on HIM-DF are provided in Appendix A.

4 Method

Given a musical-performance clip with video V and music M, HIM-CoT (Figure 3) determines its authenticity and the forged modality. Structured measurements become evidence tokens that query the Qwen2.5-Omni-7B (Xu et al., 2025) media tokens, and the resulting evidence drives three-stage reasoning with joint classification.

4.1 Structured Evidence from the Production Chain

For each clip, the audio and video encoders of Qwen2.5-Omni-7B produce music and video token sequences Z^M and Z^V. While these pretrained representations capture rich semantic and temporal information, they do not explicitly encode the structured relationships imposed by the physical production process.

We augment these representations with structured measurements extracted from specialized analysis modules. Features are organized into four groups aligned with the Human-Instrument-Music production chain. The music group x_M encodes pitch and voicing statistics from two pitch estimators and their disagreement. The visual group x_V captures hand kinematics and tracking quality from hand-pose estimation. The contact group x_C represents estimated performer-instrument interaction patterns, including contact topology and temporal dynamics. The global group x_G encodes optical-flow and CLIP temporal statistics and the synchronization between audio onsets and motion.

However, these structured measurements may contain missing or unreliable values due to occlusion, motion blur, or detection failures. To handle missing measurements, we introduce observation masks $m_k \in \{0,1\}^{d_k}$ for each group k, where d_k is the feature dimensionality of group k and $m_{k,i} = 1$ indicates that feature i is valid. Each group is then projected to a common representation space while preserving which values were observed:

$$e_k = \mathrm{MLP}_k\left(\mathrm{Concat}(\mathrm{Norm}(x_k \odot m_k), m_k)\right), \quad k \in \{M, V, C, G\}. \tag{1}$$

By concatenating the mask m_k with the masked features, the projection network learns to distinguish genuinely zero-valued measurements from missing observations. On HIM-DF all groups are observed for every sample (Appendix A.8), so the masks matter mainly for recordings where extractors fail.

Evidence groups vary in reliability, e.g., contact dynamics become less informative under occlusion. To adaptively weight evidence groups based on their reliability, we employ a softmax gating network:

$$\alpha = \mathrm{Softmax}(\Phi_{\mathrm{gate}}(\mathrm{Concat}(e_M, e_V, e_C, e_G))), \tag{2}$$

where Φ_{gate} is a two-layer MLP and $\alpha \in \mathbb{R}^4$ assigns adaptive weights to the four groups. Each token is rescaled to $\alpha_k e_k$, projected to the Qwen hidden space, and added to a learned group embedding. We keep writing e_k for the result.

4.2 Modality-Specific Retrieval and Discrepancy Modeling

Standard multimodal fusion concatenates or pools features across modalities (Zhou & Lim, 2021; Oorloff et al., 2024), conflating different sources of forensic information. Unlike prior methods, we instead employ evidence tokens as queries to independently retrieve modality-specific support from pretrained audio and video representations.

For each evidence token e_k, we compute separate audio and visual responses r_k^M, r_k^V via cross-attention:

$$r_k^M = \mathrm{CrossAttn}(e_k, Z^M, Z^M), \tag{3}$$

$$r_k^V = \mathrm{CrossAttn}(e_k, Z^V, Z^V), \tag{4}$$

where e_k serves as query and the media tokens Z^M, Z^V serve as keys and values. These responses encode how the same forensic cue is manifested in each modality.

In authentic samples, audio and video responses to a given evidence query should be consistent, as both modalities reflect the same underlying performance. However, forged samples often exhibit disagreement: a forensic cue may be strongly supported by one modality but weakly by another. We quantify such cross-modal discrepancy mainly through semantic disagreement, captured by the response difference $\Delta r_k = r_k^V - r_k^M$ for each evidence group. As an auxiliary signal, we also compare the attention distributions of the two modalities over time. This quantity is not a direct estimate of physical audio–video offset. Instead, it measures whether the two modalities assign their strongest evidence to the same temporal regions. For each evidence token, we take its cross-attention weights, average them across attention heads, and sum the spatial locations within each video frame. The audio and video weights are then interpolated to a shared temporal grid of $T = 32$ bins and normalized, producing distributions $a_k^M, a_k^V \in \mathbb{R}^T$ that indicate when each modality supports evidence group k most strongly. The temporal attention discrepancy $\Delta_{t,k} = a_k^V - a_k^M$ captures whether audio and video attend to consistent temporal regions, even when the media have low measurable physical offset. It is only weakly discriminative on its own (Appendix C.10). HIM-CoT is only partly sensitive to temporal misalignment (Appendix C.7), so we treat this signal as an auxiliary input rather than as a reliable mechanism for verifying timing. Both discrepancy signals are fused into a per-group representation:

$$e_{D,k} = F_D(\mathrm{Concat}(\Delta r_k, \Delta_{t,k})), \tag{5}$$

where F_D is a lightweight projection (a linear layer followed by GELU and layer normalization).

At this point, we have obtained three complementary sources of information: the original structured evidence $E = [e_M; e_V; e_C; e_G]$, modality-specific responses $\{r_k^M, r_k^V\}$ that reflect how each modality interprets each evidence group, and the discrepancy representations $e_{D,k}$ that capture cross-modal disagreement. To adaptively integrate these sources based on their relevance, a learned gate combines them:

$$\tilde{E}_k = \mathrm{LN}\big(e_k + \beta_{k,M} \cdot r_k^M + \beta_{k,V} \cdot r_k^V + \beta_{k,D} \cdot e_{D,k}\big), \tag{6}$$

$$\beta_{k,M}, \beta_{k,V}, \beta_{k,D} = \sigma(\Phi(\mathrm{Concat}(e_k, r_k^M, r_k^V))), \tag{7}$$

where Φ is a two-layer gating MLP, σ is the sigmoid function, and LN is layer normalization. This produces grounded representations that encode both forensic content and cross-modal agreement.

4.3 Three-Stage Reasoning with Auxiliary Supervision

End-to-end fusion obscures which cues drive the decision and skips the analysis human experts follow: check whether the music is plausible, then whether the performance is feasible, and finally whether audio and video are consistent. We decompose reasoning into three stages that mirror this workflow.

To implement this staged reasoning, we construct three stage-specific embeddings from the grounded representations $\tilde{E}$ obtained in Section 4.2, each aggregating evidence relevant to one stage:

$$\mathbf{s}_{\text{music}} = \mathrm{LN}\big(\tilde{E}_M + r_M^M\big), \tag{8}$$

$$\mathbf{s}_{\text{perf}} = \mathrm{LN}\Big(\frac{1}{2} \sum_{k \in \{V,C\}} (\tilde{E}_k + r_k^V)\Big), \tag{9}$$

$$\mathbf{s}_{\text{cross}} = \mathrm{LN}\big(\tilde{E}_G + e_{D,G}\big). \tag{10}$$

Here, r_M^M is the audio response to the music evidence, r_k^V is the visual response to group k, and $e_{D,G}$ is the discrepancy representation of the global group. The residual terms re-emphasize the stage-relevant response: acoustic for music, visual and contact for performance, and discrepancy for the cross-modal stage. The four evidence tokens $\tilde{E}_k$ and three stage embeddings are inserted as seven special tokens into a prompt that also contains the Qwen audio and video tokens. The measurements enter the model only through these tokens, never as text. The decoder generates text conditioned on these inputs, producing reasoning through three phases: (1) acoustic plausibility, evaluating harmonic structure and spectro-temporal patterns relative to the claimed instrument, (2) performance plausibility, assessing whether motion and contact patterns align with the visible instrument, and (3) cross-modal consistency, integrating prior findings with the discrepancy evidence to determine if the performance coherently explains the music. For classification, final-layer hidden states are mean-pooled within each generated stage and combined by a learned stage-weighting gate, and two linear heads predict authenticity and forgery type, so the decision is conditioned on the model's own reasoning.

4.4 Training Objective

The model is trained with a multi-task objective combining classification and language modeling:

$$\mathcal{L} = \mathcal{L}_{\text{auth}} + \lambda_f \mathcal{L}_{\text{type}} + \lambda_r \mathcal{L}_{\text{reason}}, \tag{11}$$

where $\mathcal{L}_{\text{auth}}$ is binary cross-entropy for authenticity prediction, $\mathcal{L}_{\text{type}}$ is cross-entropy for forgery-type classification (three classes), and $\mathcal{L}_{\text{reason}}$ is autoregressive language modeling loss over generated reasoning tokens. The language loss applies only to the assistant's response, excluding input tokens from gradient computation. We adopt LoRA (Hu et al., 2022) to fine-tune attention projection layers while keeping the base model frozen, preserve the natural class distribution without resampling, and set λ_f and λ_r to balance each objective.

Table 1: Main results on HIM-DF and LAV-DF test sets. We report accuracy (ACC), average precision (AP), and area under ROC curve (AUC) for authenticity detection on both benchmarks. "–": not evaluated.

Method	HIM-DF			LAV-DF		
	ACC (%)	AP (%)	AUC (%)	ACC (%)	AP (%)	AUC (%)
	Unimodal Detectors			*Unimodal Detectors*		
AASIST (audio)	61.7	83.7	71.6	67.6	79.7	58.5
WavLM (audio)	64.4	81.5	69.9	73.5	92.2	83.8
Self-Blended (video)	50.6	62.3	51.3	73.7	74.4	59.2
AltFreezing (video)	62.0	68.7	58.1	72.7	70.2	56.6
	Audio-Visual Methods			*Audio-Visual Methods*		
ICS-AV	57.4	74.9	62.3	63.1	80.8	61.4
T-AVFD	74.3	88.9	71.4	81.8	92.3	79.6
AVFF	56.3	67.7	55.8	73.5	79.2	61.0
	CoT-Based Methods			*CoT-Based Methods*		
FaceCoT	52.7	63.6	54.1	**92.9**	**98.7**	**97.2**
Gemini 3.8 Flash	52.9	68.3	55.8	74.5	78.3	67.2
MiniMax H3	60.9	65.1	55.6	–	–	–
	Late Fusion and Fine-Tuned MLLM			*Late Fusion and Fine-Tuned MLLM*		
Late fusion (WavLM + AltFreezing)	66.7	76.3	71.4	74.7	80.6	76.2
Qwen2.5-Omni LoRA (label only)	63.3	74.3	62.9	79.8	90.6	77.9
HIM-CoT (Ours)	**89.3**	**95.0**	**94.2**	90.6	91.2	90.9

5 Experiments

5.1 Experimental Setup

Datasets and Metrics We evaluate on two datasets. HIM-DF contains 6,998 samples (2,574 real, 2,276 M2V, and 2,148 V2M) across 12 instrument categories (Section 3) and is split 8:1:1 into training, validation, and test sets, grouped by source recording to prevent leakage. LAV-DF (Cai et al., 2023) contains 136,304 talking-face videos of 153 identities with content-driven audio, visual, and audio-visual manipulations. We address video-level detection only, not temporal localization, and adapt the measurements to a speaker–articulator–speech chain for talking faces (Appendix B.2). We report accuracy (ACC), average precision (AP), and area under the ROC curve (AUC), where AP and AUC are threshold-independent and AP is robust to class imbalance.

Baselines We compare against four groups of baselines: (1) unimodal detectors, including the audio detectors AASIST (Jung et al., 2022) and WavLM (Liu et al., 2024b) and the video detectors Self-Blended Images (Shiohara & Yamasaki, 2022) and AltFreezing (Zhao et al., 2023), (2) audio-visual methods, including ICS-AV (Anshul et al., 2025) for intra- and cross-modal synchronization, T-AVFD (Liu et al., 2026) for text-guided talking- and singing-head forgery detection, and AVFF (Oorloff et al., 2024) for cross-attention fusion, and (3) CoT-based methods, including FaceCoT (Zhang et al., 2026) for multi-stage facial forgery reasoning and the general-purpose models Gemini 3.8 Flash (Gemini Team, 2026) and MiniMax H3 (MiniMax, 2026) under zero-shot prompting, and (4) a late-fusion and a fine-tuned MLLM baseline: *late fusion* of the best audio and video detectors (logistic regression fitted on validation scores) and *label-only Qwen*, which fine-tunes the same backbone with the same LoRA setup on labels alone, without measurements or reasoning. All trainable baselines are retrained on the same source-grouped split as HIM-CoT and selected on validation (Appendix B.1).

5.2 Main Results

On HIM-DF (Table 1), HIM-CoT reaches 89.3% accuracy, 95.0% AP, and 94.2% AUC. It outperforms the strongest baseline, T-AVFD, by 15.0% in accuracy. Single-modality detectors do poorly because every forgery keeps one modality real, and HIM-CoT exceeds the best of them by 24.9% and the best CoT-based method by 28.4%. Combining an audio and a video detector does not close the gap. This late fusion is barely better than the audio detector alone, and HIM-CoT outperforms it by 22.6%. Fine-tuning the same backbone on labels alone falls short by 26.0%, so the gain does not come from the backbone. Most forged clips are convincing: for 352 of the 434 forged test clips, expert musicians judged that the visible performance could have produced the sound. HIM-CoT still detects 86.6% of these convincing forgeries, compared with 96.3% of the less convincing ones, and 36.3% more of them than late fusion (Appendix C.4). We also pair real videos with real music from a different performance. Neither modality is synthesized, but the sound does not match the perfor-

Table 2: Ablation study on HIM-DF. The first three blocks ablate the design within each component (evidence-group rows are MLP classifiers on structured features alone), and the last block removes entire components. *w/o Three-Stage Reasoning* is a reasoning-free variant: it keeps the backbone and all evidence modules but generates no text, drops $\mathcal{L}_{\text{reason}}$, and applies the classification heads directly to the evidence representation. *Observation-Only Reasoning* uses verdict-free targets (Appendix C.3).

Model Variant	**ACC (%)**	**AP (%)**	**AUC (%)**
Evidence Group Contribution			
Audio Only	51.0	61.6	52.7
Visual Only	58.1	63.8	62.4
Audio + Visual + Contact	68.3	79.7	73.5
Audio + Visual + Contact + Global	70.1	80.5	74.2
Cross-Modal Discrepancy			
w/o Discrepancy ($\beta_D = 0$)	84.1	91.3	89.5
w/o Modality-Specific Retrieval	79.7	88.2	84.7
Reasoning Structure and Content			
w/o Music Stage	85.4	90.4	90.0
w/o Performance Stage	85.7	90.6	89.8
w/o Cross-Modal Stage	84.3	89.5	88.7
Single-Stage Generation	86.1	92.8	93.5
Observation-Only Reasoning	87.6	93.7	92.9
Component-Level Ablation			
w/o Structured Evidence	66.3	77.9	67.5
w/o Cross-Modal Discrepancy	78.7	87.3	81.4
w/o Three-Stage Reasoning	82.4	87.2	87.8
HIM-CoT (Full)	**89.3**	**95.0**	**94.2**

mance. The fraction HIM-CoT labels fake rises from 9.4% on genuine clips to 61.7% on these pairs. This rise is 30.8% larger than that of late fusion, which judges each modality on its own, so HIM-CoT uses whether the sound fits the performance and not only synthesis artifacts (Appendix C.6). Shifting the audio by 4 s within the same recording raises this fraction to 22.9%, twice the rise of late fusion but far less than for mismatched performances, so HIM-CoT checks timing only in part (Appendix C.7). When one of three generators per direction is held out from training, HIM-CoT still exceeds the strongest baseline by 14.8% on average (Appendix C.8). On the talking-face benchmark LAV-DF, HIM-CoT reaches 90.6% accuracy and 90.9% AUC. It exceeds the strongest audio-visual baseline by 8.8% in accuracy and trails only FaceCoT, a detector designed for faces, by 2.3%.

5.3 Ablation Study

Table 2 reports the ablations. Classifiers trained on the structured features alone are weak, and the best of them is 19.2% below HIM-CoT. The measurements are therefore far more useful as queries into the pretrained representations than as standalone features. Comparing each cue across the two modalities matters. Removing modality-specific retrieval lowers accuracy by 9.6%, and removing the discrepancy branch ($\beta_D = 0$) lowers it by 5.2%. All three reasoning stages help. Removing the cross-modal stage hurts most, because HIM-DF forgeries are exposed mainly by comparing the two modalities. A single undivided rationale is 3.2% worse than three stages. Removing whole components lowers accuracy by 23.0% for structured evidence, 10.6% for discrepancy modeling, and 6.9% for three-stage reasoning. For the last two, each drop is larger than any of the corresponding partial ablations. The variant without three-stage reasoning classifies without generating any text. Its 6.9% gap shows that training with staged rationales improves the classifier, but not that the reasoning itself causes each prediction. When the rationales contain only observations and no verdict words, accuracy drops by just 1.7%, so most of this benefit does not come from stating the label in text. Some features, such as tracking jitter and pitch-estimator confidence, may simply detect synthesis artifacts. Removing these 18 features lowers accuracy by only 2.9%, and using them alone adds only 2.8% over no structured evidence. The gain therefore does not rest mainly on such artifacts (Appendix C.5).

5.4 Attribution and Explanation Fidelity

Table 3 reports which modality HIM-CoT identifies as synthesized and how well its explanations reflect its decisions. HIM-CoT reaches 84.6% three-way accuracy (84.4% macro-F1) and attributes

Table 3: Forgery-type attribution and explanation fidelity on the HIM-DF test split. *Attr.*: three-way attribution, i.e., the fraction of forged clips whose predicted type (Real, M2V, or V2M) is their true forgery type, so a forged clip predicted as Real counts as an error. *Agree.*: agreement between the verdict stated in the generated reasoning and the classification head. *Fidelity*: fraction of numerical values cited in the reasoning that match the structured measurements. "–": not applicable (no generated reasoning).

Method	3-way ACC	Macro-F1	Attr.	Agree.	Fidelity
Late fusion (WavLM + AltFreezing)	58.1	50.2	42.6	–	–
Qwen2.5-Omni LoRA (label only)	59.1	53.4	51.8	–	–
HIM-CoT w/o Structured Evidence	68.4	61.7	59.7	88.7	73.6
HIM-CoT	**84.6**	**84.4**	**82.7**	**95.4**	**91.8**

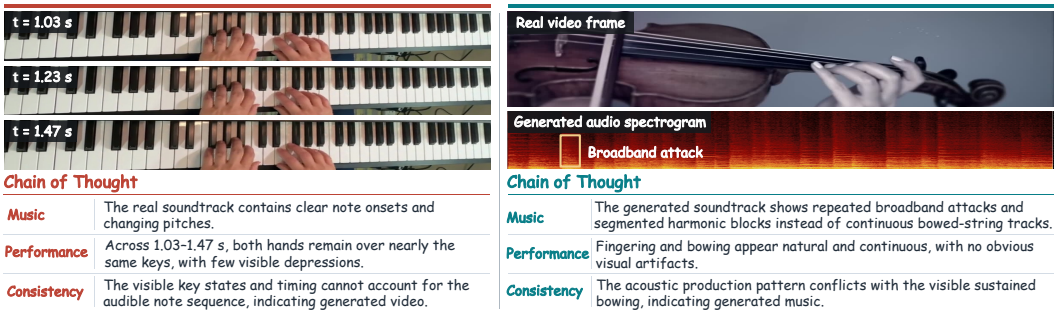

Figure 4: **Reasoning chains generated by HIM-CoT on HIM-DF.** *Left:* an M2V forgery (piano) with a real soundtrack and a generated video. *Right:* a V2M forgery (violin) with a real video and a generated soundtrack. The spectrogram shows repeated broadband attacks (boxed) that conflict with the visible sustained bowing.

82.7% of forged clips to the correct modality, 40.1% more than late fusion and 30.9% more than label-only Qwen. Removing structured evidence lowers attribution by 23.0%, so the measurements are what allow the model to tell which modality was synthesized. The verdict stated in the generated reasoning agrees with the classification head on 95.4% of clips, 6.7% more than without structured evidence, and 91.8% of the numerical values it cites match the structured measurements it receives, 18.2% more than without structured evidence. Only 4.0% of its numerical claims are unsupported by any measurement, 29.0% and 32.0% fewer than Gemini 3.8 Flash and MiniMax H3, which state specific values they cannot have measured. In a blind human evaluation of 100 test clips, raters judged HIM-CoT's reasoning stages factually correct in 88.0% of cases and grounded in evidence in 84.0%, 14.0% and 21.0% more than without structured evidence and 20.0% and 35.0% more than Gemini 3.8 Flash (Appendix C.2).

5.5 Qualitative Analysis

Case Study: Reasoning Chain Visualization In Figure 4, HIM-CoT attributes the M2V forgery to the video: the soundtrack has clear onsets, but almost no key depressions are visible. It attributes the V2M forgery to the audio: natural bowing is paired with broadband attacks atypical of bowed strings.

Appendix C.10 compares HIM-CoT's reasoning with general-purpose MLLMs and analyzes its failure modes.

6 Conclusion

We introduced HIM-DF, a bidirectionally forged benchmark for instrumental performances, and HIM-CoT, which reasons over structured Human–Instrument–Music measurements in three stages. Structured production-chain measurements provide the largest gain, and retrieval, discrepancy modeling, and staged reasoning add further gains. HIM-CoT checks mainly whether a sound fits the performance and only partly its timing, which together with unreliable measurements motivates stronger temporal reasoning and more robust feature estimation.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used large language model assistants in preparing this paper: to organize the literature review, to draft and revise parts of the manuscript text, to check the internal consistency of reported numbers across tables and text, and to help design parts of the evaluation protocol, including the attribution, reasoning-fidelity, human-evaluation, and high-plausibility analyses. Separately, as part of the method, the Chain-of-Thought training targets in HIM-DF are generated by Qwen 3.5 Omni Plus under the label-conditioned protocol and human verification described in Appendix A.7. The research idea, the construction of HIM-DF, the design of HIM-CoT, and all experiments were carried out by the authors, and every reported result comes from the authors' own experiments. The authors reviewed and edited all AI-assisted text for accuracy and take full responsibility for the content of this paper, including all claims, results, and interpretations.

ETHICS STATEMENT

(This section is **required** and does not count toward the page limit.)

This work addresses the societal concern of deepfake content in musical performances, which can undermine trust in digital media and harm artists' reputations. The HIM-DF benchmark was collected with proper licensing and attribution, ensuring that all performance videos used for analysis are publicly available or used with permission. We do not release generation tools or detailed instructions that could be misused to create convincing musical deepfakes. Instead, we focus on detection and interpretability to support forensic analysis. The proposed method provides explainable reasoning, which is crucial for accountability in real-world deployment scenarios where false accusations could harm individuals.

REPRODUCIBILITY STATEMENT

(This section is **required** and does not count toward the page limit.)

To ensure reproducibility, the model architecture is specified in Section 4, and hyperparameters and training procedures are given in Appendix B. The baseline training and evaluation protocol is described in Appendix B.1. The structured measurements, their dimensionality, and the extractors used to compute them are listed in Appendix A.8, and the measurements are released for every sample. We will release the HIM-DF benchmark with train/val/test splits, annotations, and data processing scripts upon acceptance. Model checkpoints and evaluation code will be made publicly available. For LAV-DF experiments, we follow the official data split from Cai et al. (2023) to ensure fair comparison with future work.

REFERENCES

Alibaba Cloud. Wan3.0: 30-second AI video generation from any input. https://www.alibabacloud.com/blog/wan3-0-30-second-ai-video-generation-from-any-input_603452, 2026. Product announcement.

Kumar Anshul, Abhinav Dhall, and Munawar Hayat. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

ByteDance Seed Team. Seedance 2.0 official launch. https://seed.bytedance.com/en/blog/seedance-2-0-official-launch, 2026. Product announcement.

Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization. volume 236, pp. 103818, 2023.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. Internvl 2.0: Scaling up vision-language models

with diverse and high-quality data. In *International Conference on Learning Representations*, 2026.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Luca Comanducci, Paolo Bestagini, and Stefano Tubaro. Fakemusiccaps: A dataset for detection and attribution of synthetic music generated via text-to-music models. *arXiv preprint arXiv:2409.10684*, 2024.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemini Team. Introducing gemini 3.8 flash and 3.8 flash cyber. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/`, September 2026. Released September 2, 2026.

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.

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.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367–6371, 2022.

Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2Music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, pp. 123640, 2024. doi: 10.1016/j.eswa.2024.123640.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

Youngseo Kim, Kwan Yun, Seokhyeon Hong, Sihun Cha, Colette Suhjung Koo, and Junyong Noh. X-avdt: Audio-visual cross-attention for robust deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025. URL `https://arxiv.org/abs/2512.16776`.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-next: Improved reasoning, ocr, and world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, Zicheng Liu, and Jianfeng Wang. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends in Computer Graphics and Vision*, 16(1–2):1–214, 2024.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yupei Li, Hanqian Li, Lucia Specia, and Björn W. Schuller. M6: Multi-generator, multi-domain, multi-lingual and cultural, multi-genres, multi-instrument machine-generated music detection databases. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-36044-w.

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. *arXiv preprint arXiv:2605.27944*, 2026.

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. MuMu-LLaMA: Multi-modal music understanding and generation via large language models. *arXiv preprint arXiv:2412.06660*, 2024a.

Xuechen Liu, Md Sahidullah, and Tomi Kinnunen. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024b.

MiniMax. Minimax h3: A general-purpose omni-modal generative system. Model card, `https://huggingface.co/MiniMaxAI/MiniMax-H3`, 2026. Released July 2026.

Trevine Oorloff, Davide Coccomini, Roberto Caldelli, Fabrizio Falchi, Giuseppe Amato, and Nicola Nikoltsos-Chrysochoidis. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11957–11967, 2024.

Jielun Peng, Yabin Wang, Yaqi Li, Long Kong, and Xiaopeng Hong. Leave no stone unturned: Uncovering holistic audio-visual intrinsic coherence for deepfake detection. *arXiv preprint arXiv:2603.23960*, 2026.

Md Awsafur Rahman, Zaber Ibn Abdul Hakim, Najibul Haque Sarker, Bishmoy Paul, and Shaikh Anowarul Fattah. Sonics: Synthetic or not – identifying counterfeit songs. In *International Conference on Learning Representations*, 2025. arXiv preprint arXiv:2408.14080.

Andreas Rössler, 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*, 2019.

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

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Qwen Team. Qwen 3.5 omni plus: Enhanced multimodal understanding and reasoning. `https://qwenlm.github.io/`, 2025. Technical report.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Haoxing Yang, Ke Liu, Jiahong Li, Liang Wang, and Stan Z. Li. Dawn of vision-language models for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. Singfake: Singing voice deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12156–12160. IEEE, 2024.

Wei Zhang, Jing Liu, Hao Chen, Xiaoming Wang, and Yang Li. Harnessing chain-of-thought reasoning in multimodal large language models for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2341–2350, 2026.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhendong Zhao, Jianqiu Wei, Rui Ni, and Yao Zhao. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4129–4138, 2023.

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.

APPENDIX

A ADDITIONAL DETAILS OF THE HIM-DF BENCHMARK

This section complements Section 3 with the information needed to use, audit, and extend HIM-DF. We first position HIM-DF against existing benchmarks (§A.1), provide visual examples of the released samples (§A.2), and explain why the paired bidirectional construction yields a benchmark that cannot be solved from a single modality (§A.3). We then detail source collection (§A.4), forgery generation (§A.5), quality control (§A.6), Chain-of-Thought annotation (§A.7), the released structured measurements (§A.8), dataset statistics and splits (§A.9), and the release protocol and scope of the benchmark (§A.10).

A.1 COMPARISON WITH EXISTING BENCHMARKS

Table 4 contrasts HIM-DF with representative audio-visual and music deepfake benchmarks. Audio-visual benchmarks focus on talking or singing faces, where the audio–visual coupling is speech–lip synchronization, while music benchmarks are audio-only. HIM-DF is, to our knowledge, the first benchmark in which sound is produced through visible physical interaction with an instrument, and the first to pair every sample with structured production-chain measurements and stage-wise reasoning annotations.

Table 4: Comparison of HIM-DF with existing deepfake benchmarks. *Instrument interaction*: sound is produced by visible performer–instrument contact. *Structured measurements*: per-sample acoustic, kinematic, and contact measurements are released. *Reasoning annotation*: per-sample stage-wise rationales are provided.

Benchmark	Domain	Modality	Instrument interaction	Structured measurements	Reasoning annotation
FakeAVCeleb (Khalid et al., 2021)	Talking face	A+V	✗	✗	✗
LAV-DF (Cai et al., 2023)	Talking face	A+V	✗	✗	✗
SHDF (Liu et al., 2026)	Singing head	A+V	✗	✗	✗
SingFake (Zang et al., 2024)	Singing voice	A	✗	✗	✗
FakeMusicCaps (Comanducci et al., 2024)	Text-to-music	A	✗	✗	✗
SONICS (Rahman et al., 2025)	Full songs	A	✗	✗	✗
M6 (Li et al., 2026)	Multi-genre music	A	✗	✗	✗
HIM-DF (Ours)	Instrumental performance	A+V	✓	✓	✓

A.2 VISUAL EXAMPLES AND PAIRING STRUCTURE

HIM-DF is built around a controlled pairing structure, and Figure 5 shows what changes across the three sample types and what remains fixed. Each row is a source-matched example from a different instrument family, and the columns show REAL, M2V, and V2M. The REAL column shows the authentic recording, the M2V column keeps the authentic audio while replacing the performance video, and the V2M column keeps the authentic video while replacing the music. Because each row uses the same source segment, the direction of the manipulation is visually explicit and the figure cannot be read as a comparison of unrelated performances.

A.3 DESIGN RATIONALE: PAIRED BIDIRECTIONAL FORGERIES

Each HIM-DF sample is a tuple (V, M, y, t), where $y \in \{0, 1\}$ denotes authenticity and $t \in \{\text{REAL}, \text{M2V}, \text{V2M}\}$ denotes the forgery type. Starting from an authentic pair (V, M), the two forged subsets are constructed as:

$$\text{M2V}: \ (\hat{V}, M), \ \ \hat{V} = G_{\text{M2V}}(M, c), \qquad \text{V2M}: \ (V, \hat{M}), \ \ \hat{M} = G_{\text{V2M}}(V), \tag{12}$$

where G_{M2V} is Seedance 2.0, G_{V2M} is VidMuse, and c is the visual conditioning described in Appendix A.5. This construction has three consequences that shape how HIM-DF should be interpreted.

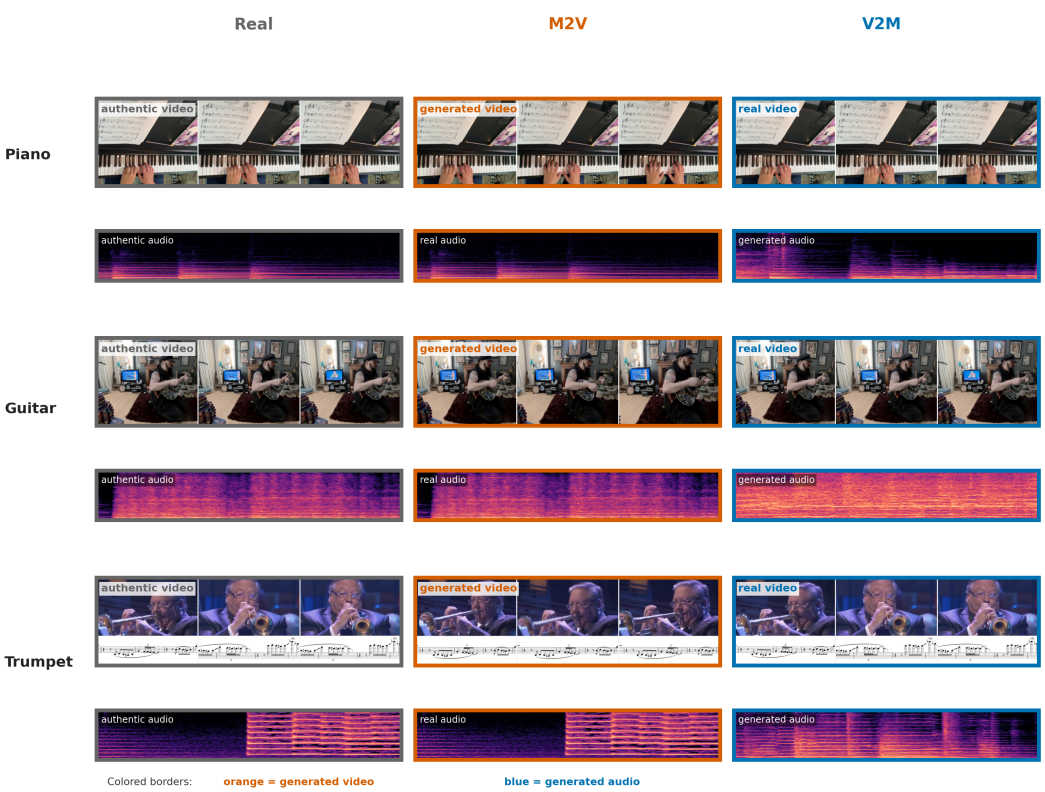

Figure 5: **Source-matched visual examples from HIM-DF.** Each row uses the same source segment across the three columns: piano (keyboard), guitar (plucked string), and trumpet (brass). Each cell shows three continuous video frames above the corresponding 2.5-second spectrogram. M2V keeps the real audio and replaces the video with a generated performance, whereas V2M keeps the real video and replaces the audio with generated music. Orange borders mark generated video and blue borders mark generated audio.

Labels are deterministic provenance labels. Both y and t are determined by which generator produced the sample, not by human or model judgment. The multimodal LLM used in Appendix A.7 only writes rationales for these known labels and never decides them. Consequently, the detection metrics are free of human annotation noise. The labels encode provenance rather than physical plausibility: a synthesized clip is labeled fake even when its performance plausibly explains the sound, as many retained forgeries do according to the expert plausibility ratings (Appendix A.6).

Forgeries are cross-modally conditioned. Exactly one modality is synthesized per forged sample, and it is generated *conditioned on* the authentic one. Forgeries are therefore designed to be plausible with respect to their counterpart, unlike random re-pairing of audio and video. Detecting them therefore cannot rely on coarse audio–visual mismatches, and relating the performance to the sound becomes an informative cue in addition to the artifacts of each modality.

Single-modality detectors have directional limits. Each forgery direction leaves one modality authentic. An audio-only detector can identify V2M samples because their audio is synthesized, but it cannot identify M2V samples from audio alone because their audio remains authentic. The reverse holds for a video-only detector: it can identify M2V samples from the generated video, but it cannot identify V2M samples from video alone. The following ceilings describe an idealized detector that recognizes the forged modality and treats the untouched modality as authentic. Let π_t denote the proportion of type t. For an audio-only detector, this gives

$$\mathrm{ACC}_{\mathrm{audio}} \le \pi_{\mathrm{V2M}} + \max(\pi_{\mathrm{REAL}}, \pi_{\mathrm{M2V}}), \qquad \mathrm{AUC}_{\mathrm{audio}} \le \frac{N_{\mathrm{V2M}} + \frac{1}{2}N_{\mathrm{M2V}}}{N_{\mathrm{M2V}} + N_{\mathrm{V2M}}}, \tag{13}$$

where the AUC expression ranks all V2M samples above the authentic-audio samples and ranks M2V and REAL at chance. The video-only expressions follow by swapping M2V and V2M. Table 5 instantiates these directional ceilings with the class proportions of the HIM-DF test split, on which all results are reported, and compares them with the strongest single-modality baselines. The AP ceiling is the expected AP when ties among authentic-modality samples are broken at random.

Table 5: Directional limits of single-modality detection on HIM-DF. An audio-only detector can identify V2M samples but cannot identify M2V samples from audio alone. A video-only detector has the opposite limitation. The idealized ceilings are computed from the class proportions of the test split (266 Real, 234 M2V, 200 V2M) and are compared with the best single-modality baselines in Table 1 and HIM-CoT. For the audio row, the best value of each metric is taken over AASIST and WavLM separately.

Setting	**ACC (%)**	**AP (%)**	**AUC (%)**
Audio-only oracle ceiling	66.6	85.7	73.0
Best audio-only baseline (WavLM for ACC, AASIST for AP and AUC)	64.4	83.7	71.6
Video-only oracle ceiling	71.4	88.2	77.0
Best video-only baseline (AltFreezing)	62.0	68.7	58.1
HIM-CoT (audio + video + evidence)	**89.3**	**95.0**	**94.2**

Table 5 has two purposes. First, it makes the evaluation protocol explicit: an audio-only model is tested on a benchmark containing both directions, so it can detect the synthesized audio in V2M but must miss M2V unless it exploits an unintended dataset shortcut. A video-only model faces the symmetric limitation. Second, the oracle rows provide reference ceilings for this directional setting. The unimodal baselines remain below their corresponding ceilings, while HIM-CoT exceeds both on all three metrics by using cross-modal evidence. Notably, ICS-AV and AVFF do not reach even the unimodal ceilings, and T-AVFD exceeds them only in accuracy and AP, while its AUC remains below both, indicating that correspondence priors learned from speech largely fail to transfer to instrument-mediated sound production. These ceilings assume that the untouched modality is not statistically distinguishable between the forged and real subsets. We therefore apply the same post-processing pipeline to all subsets (Appendix A.5) and treat any remaining gap as a possible dataset shortcut rather than evidence of cross-modal reasoning. On the full benchmark, whose class proportions differ slightly from the test split, the ceilings are 67.5 / 87.0 / 74.3 (audio) and 69.3 / 87.9 / 75.7 (video) for ACC / AP / AUC. All comparisons above hold under either set of proportions.

A.4 SOURCE COLLECTION AND PREPROCESSING

Sources. We collect 3,458 clips from 272 sources, where a source is a distinct uploader channel or recording session obtained from platform(s) and licensing terms, e.g., Creative-Commons-licensed performance videos. Grouping clips by source is the unit used for leakage-free splitting (Appendix A.9).

Inclusion criteria. A clip is kept only if it satisfies all of the following: (1) it is a solo performance on a single instrument, (2) the performer's hands or mouths and the sound-producing region of the instrument are visible in at least 80% of frames, (3) the audio was recorded together with the video (overdubbed, finger-synced, and post-produced soundtracks are excluded), (4) no voice-over, backing track, or audience noise dominates the instrument, and (5) the clip contains no scene cuts, split screens, or on-screen overlays.

Filtering. Quality filtering reduces the 3,458 collected clips to 2,574 authentic samples, removing 884 clips (25.6%). The main reasons for removal are insufficient hand or instrument visibility, non-live or edited audio, scene cuts or overlays, low resolution or bitrate, and near-duplicates.

Preprocessing. Each clip is trimmed to 2.5 s. Video is resampled to 25 fps at 720P, and audio is resampled to 48 kHz.

A.5 FORGERY GENERATION

Music-to-Video (M2V). For each selected authentic pair, Seedance 2.0 (ByteDance Seed Team, 2026) receives the authentic music track M as an audio reference, a reference frame sampled from the source video that fixes performer appearance, instrument, and scene, and a templated text prompt of the form "*A {instrument} player performing solo, playing in synchrony with the provided music*". Videos are generated at 720P for 2.5 s with default sampling settings. The generated video $\hat{V}$ is combined with the untouched authentic track M, yielding 2,276 M2V samples.

Video-to-Music (V2M). VidMuse (Tian et al., 2025) receives the authentic video V and synthesizes music of the same duration using the official checkpoint and default sampling settings. The generated track $\hat{M}$ is combined with the untouched authentic video V, yielding 2,148 V2M samples.

Generation attempts. For each authentic clip and direction, we generate up to three candidates and keep the first one that passes the automated screening of Appendix A.6. Clips for which no candidate passes, or whose kept candidate is later rejected in the expert rating, receive no forgery in that direction, and rejected candidates are not regenerated. HIM-DF therefore contains 2,276 M2V and 2,148 V2M forgeries for its 2,574 authentic clips, and 298 and 426 clips have no M2V and no V2M forgery, respectively.

Generator choice. Seedance 2.0 accepts audio as a conditioning reference, which makes music-driven performance synthesis possible and produces videos whose motion is intended to follow the given music. VidMuse is a representative open video-to-music model that conditions on both local and global visual content, so its output is intended to match the visible performance. Both therefore produce cross-modally conditioned forgeries rather than unrelated content. We discuss the single-generator-per-direction scope in Appendix A.10.

Format harmonization. To prevent file-level shortcuts, all three subsets, including the untouched modality of forged samples, pass through the same final pipeline: video is re-encoded with H.264 at CRF 23, 25 fps, and 720P, audio is re-encoded at 48 kHz, and all samples share the same container and duration. As a sanity check, a logistic-regression classifier trained on file-level statistics (duration, frame rate, resolution, bitrate, sampling rate, integrated loudness) reaches 60.7% accuracy on the test split, below the 62.0% obtained by always predicting "fake".

A.6 QUALITY CONTROL PROTOCOL

Quality control has two stages. **Automated screening (all generated samples).** We discard generations that fail or are truncated, whose duration differs from the source by more than 0.5 s, or in which the performer's hands or the instrument are detected in fewer than 80% of frames, using the same detectors as in Appendix A.8. We also discard V2M outputs that are near-silent or clipped. Across the up to three attempts per clip, this stage removes 1,236 M2V and 1,876 V2M candidates. **Expert rating.** Three professional musicians independently rate 300 synthesized samples, stratified by forgery direction (150 M2V / 150 V2M) and instrument, on visual realism, audio quality, and perceptual plausibility, using the 5-point rubric in Table 6. A sample is retained if its mean score over raters and criteria exceeds 3.5. 82.7% of rated samples pass (95% Wilson interval 78.0–86.6%). Inter-rater agreement is Krippendorff's α = 0.72, and the mean scores of retained samples are 4.0 / 4.5 / 3.9 for M2V and 4.6 / 4.0 / 3.8 for V2M. Besides filtering the rated samples, the study is used to calibrate the automated-screening thresholds above, and its pass rate estimates the perceptual quality of the unrated remainder.

A.7 STRUCTURED CHAIN-OF-THOUGHT ANNOTATION

Label-conditioned generation. Annotations are generated by Qwen 3.5 Omni Plus (Team, 2025), which receives the audio-visual clip, the instrument category, the ground-truth label y and forgery type t, and the structured measurements (Appendix A.8). The annotator is asked to *explain* a known verdict rather than to reach one. This choice is deliberate: zero-shot multimodal LLMs perform close to chance on HIM-DF (Table 1), so unconditioned annotations would inject the annotator's detection errors into the reasoning supervision.

Table 6: Rating rubric used by the expert musicians (anchors are given for scores 1, 3, and 5, with 2 and 4 as intermediate scores).

Criterion	1 (poor)	3 (acceptable)	5 (excellent)
Visual realism	Obvious artifacts: distorted fingers, deforming instrument, identity drift	Minor artifacts visible on close inspection	Indistinguishable from real footage
Audio quality	Noisy, distorted, or clearly synthetic timbre	Acceptable, with audible artifacts	Natural timbre and recording quality
Perceptual plausibility	Performance clearly cannot produce the sound (e.g., no motion during notes)	Rough match with noticeable timing or gesture mismatches	A musician would accept the performance as producing the sound on casual viewing

Schema. Each annotation is a JSON object with one entry per reasoning stage (music, performance, cross-modal) and a final verdict (Figure 6). Each stage lists observations tied to the measurements and a stage-level verdict.

```
{
  "music_analysis": {
      "observations": [...], "verdict": "plausible | suspicious"},
  "performance_analysis": {
      "observations": [...], "verdict": "plausible | suspicious"},
  "cross_modal_analysis": {
      "observations": [...], "verdict": "consistent | inconsistent"},
  "final": {"label": "real | fake", "forgery_type": "Real | M2V | V2M"}
}
```

Figure 6: Schema of the structured CoT annotations.

Automatic consistency checks Every annotation must pass four checks: (a) it is valid JSON conforming to the schema, (b) its final label and forgery type equal the ground truth, (c) it does not attribute the forgery to the authentic modality (the music stage of an M2V sample and the performance stage of a V2M sample must not claim that modality is synthetic), and (d) numerical values cited in observations match the structured measurements within tolerance. Annotations that fail are regenerated up to 5 times. Check (c) exploits the construction of HIM-DF: because the untouched modality is known to be authentic, rationales that blame it are provably wrong and can be filtered without human effort.

Human verification. Two expert annotators independently verify 200 annotations, stratified by forgery type and instrument, judging each reasoning stage as correct or incorrect with respect to the media and the measurements. Their agreement on these judgments is Cohen's $\kappa = 0.87$. 5% of verified annotations contained at least one incorrect stage. The error categories are used to refine the annotation prompt, and the full corpus is then regenerated with the refined prompt. The 200 verified annotations are replaced by their manually corrected versions.

Use of annotations. CoT annotations serve only as targets for $\mathcal{L}_{\text{reason}}$. Because they are generated conditioned on the known label, they supervise how to explain a decision from the evidence rather than an independent derivation of it. The agreement, fidelity, and human evaluations (Appendix C.2) test whether the generated reasoning matches the model's decision and evidence, but they do not by themselves establish that the reasoning causes the decision. No reported detection metric depends on them. Reasoning quality is examined qualitatively in the case studies and error analysis (Appendix C.10). The released annotations are marked as automatically generated or human-verified.

A.8 Structured Measurements

HIM-DF releases, for every sample, the 69-dimensional structured measurements used by HIM-CoT together with their observation masks (Table 7). All measurements are extracted with the same pipeline for all three subsets.

Table 7: Structured measurements released with HIM-DF and the extractors used to compute them.

Group	Dim.	Content	Extractor
Music x_M	21	F0 and voicing statistics of two pitch estimators and their disagreement	pYIN + CREPE
Visual x_V	20	Hand kinematics, finger opening, and tracking quality	MediaPipe Hand Landmarker
Contact x_C	18	Hand–instrument contact probability, SDF distance, penetration depth, and hand speed	Point-cloud SDF contact field
Global x_G	10	Optical-flow energy statistics, CLIP temporal statistics, and audio onset–motion synchronization	Farneback flow + CLIP + onset–motion correlation

Feature definitions. Table 8 defines all 69 measurements. Each is computed over the 2.5 s window of a clip. Distances marked "norm. dist." are divided by the palm size, measured as the distance from the wrist to the middle-finger MCP joint, so they do not depend on the camera distance. Before entering the model, every dimension is standardized with the mean and standard deviation of the training split. Dimensions 1, 42, and 43 are constant after format harmonization and carry no information. Contact features are computed from a signed distance field (SDF) of the instrument surface. When no surface can be reconstructed, an analytic plane is fitted instead, which dimension 19 records.

Table 8: Definitions of the 69 structured measurements. Statistics in braces denote one dimension each.

Dim.	Feature	Definition	Unit
Audio group x_M (21 dimensions)			
1	`audio_duration_seconds`	Duration of the audio window	s
20–23	`crepe_confidence_`{mean, median, p95, std}	Mean, median, 95th percentile, and standard deviation of the per-frame CREPE periodicity confidence	0–1
24–27	`crepe_f0_voiced_`{mean, median, p95, std}	Statistics of the CREPE fundamental frequency over voiced frames	Hz
28	`crepe_voiced_fraction`	Fraction of frames CREPE labels as voiced	0–1
29	`f0_median_abs_disagreement_hz`	Absolute difference between the voiced F0 medians of pYIN and CREPE	Hz
30	`f0_median_relative_disagreement`	The same difference relative to the CREPE median	–
31–34	`pyin_confidence_`{mean, median, p95, std}	Statistics of the per-frame pYIN voicing probability	0–1
35–38	`pyin_f0_voiced_`{mean, median, p95, std}	Statistics of the pYIN fundamental frequency over voiced frames	Hz
39	`pyin_voiced_fraction`	Fraction of frames pYIN labels as voiced	0–1
Contact group x_C (18 dimensions)			
2	`contact_active_fraction`	Fraction of frames with contact probability above 0.5	0–1
3–6	`contact_contact_probability_`{mean, median, p95, std}	Statistics of the per-frame hand–instrument contact probability	0–1
7–10	`contact_min_abs_sdf_`{mean, median, p95, std}	Statistics of the per-frame minimum absolute SDF distance from hand points to the instrument surface	norm. dist.
11–14	`contact_penetration_`{mean, median, p95, std}	Statistics of the depth of hand points inside the estimated instrument surface	norm. dist.
15–18	`contact_speed_`{mean, median, p95, std}	Statistics of the speed norm of hand keypoints	norm. dist./s
19	`contact_surface_is_analytic_plane`	Whether an analytic plane is used as the fallback instrument surface	0/1
Visual group x_V (20 dimensions)			
40	`visual_bone_length_instability_median`	Median relative deviation of the 20 hand bones from their temporal median length	–
41	`visual_bone_length_instability_p95`	95th percentile of the same deviation	–
42	`visual_duration_seconds`	Time span of the video used for the visual features	s
43	`visual_feature_version`	Version of the visual feature extractor	constant
44	`visual_finger_opening_mean_norm`	Mean fingertip-to-palm-center distance	norm. dist.
45	`visual_finger_opening_std_norm`	Standard deviation of the same distance	norm. dist.
46	`visual_finger_opening_velocity_norm`	Mean absolute rate of change of the finger opening	norm. dist./s
47	`visual_fingertip_range_max_norm`	Largest displacement range among the five fingertips	norm. dist.
48	`visual_fingertip_range_mean_norm`	Mean displacement range of the five fingertips	norm. dist.
49	`visual_fingertip_speed_norm`	Mean speed norm of the five fingertips	norm. dist./s
50	`visual_flag_low_tracking`	Tracking coverage below 50%	0/1
51	`visual_flag_severe_jitter`	Jitter ratio above 8	0/1
52	`visual_flag_unstable_hand_geometry`	95th-percentile bone-length deviation above 0.35	0/1
53	`visual_hand_range_mean_norm`	Mean motion range over all hand keypoints	norm. dist.

Dim.	Feature	Definition	Unit
54	visual_jitter_ratio	Mean acceleration divided by mean speed, measuring high-frequency jitter	–
55	visual_mean_acceleration_norm	Mean acceleration norm over hand keypoints	norm. dist./s^2
56	visual_mean_detection_confidence	Median MediaPipe confidence over frames with a detection	0–1
57	visual_mean_jerk_norm	Mean jerk (third temporal derivative) norm over hand keypoints	norm. dist./s^3
58	visual_mean_speed_norm	Mean speed norm over hand keypoints	norm. dist./s
59	visual_median_speed_norm	Median speed norm over hand keypoints	norm. dist./s
Global group x_G (10 dimensions)			
60	global_flow_energy_mean	Temporal mean of the Farneback optical-flow energy, measuring overall motion intensity	–
61	global_flow_energy_std	Temporal standard deviation of the flow energy	–
62	global_flow_energy_spectral_entropy	Spectral entropy of the flow-energy series, measuring the complexity of the motion rhythm	0–1
63	global_flow_energy_periodicity	Normalized autocorrelation peak of the flow-energy series	–
64	global_clip_temporal_cosine_mean	Mean cosine similarity between CLIP embeddings of adjacent video segments	-1–1
65	global_clip_embedding_velocity_mean	Mean change between adjacent CLIP embeddings	emb. dist./s
66	global_clip_embedding_acceleration_mean	Mean change of the CLIP embedding velocity	emb. dist./s^2
67	global_sync_zero_lag_corr	Zero-lag correlation between the audio onset envelope and visual motion intensity	-1–1
68	global_sync_best_corr	Maximum of the same correlation over the allowed lags	-1–1
69	global_sync_best_lag_ms	Lag at which the maximum correlation is reached	ms

Missing values are not a label shortcut. Pose and contact estimators may fail more often on generated video than on real video, in which case the observation mask m_k alone could reveal the label. Across the released feature files, both groups contain valid observations for every sample and every class. The corresponding masks are therefore constant across the dataset. A classifier trained on the masks alone receives the same input for every sample, so it can only predict the training majority class (fake) and reaches 62.0% test accuracy, exactly the rate of always predicting "fake". This shows that the discriminative signal of the structured evidence (Table 2) comes from the measured values, not from their availability.

A.9 Statistics and Splits

Split protocol. Splits are grouped by source: all clips from one source, together with every forgery derived from them, go to a single split. No performer, recording session, or authentic audio or video track therefore appears in more than one split. Because each forged sample shares one modality with an authentic sample, this grouping is necessary. Without it, a model could recognize a test sample from a training sample that shares its authentic track. Sources are assigned to splits with stratification by instrument, so every category appears in all three splits. Table 9 lists the resulting counts.

Table 9: Sample counts per split and class.

Split	Real	M2V	V2M	Total
Train	2,036	1,798	1,764	5,598
Val	272	244	184	700
Test	266	234	200	700
Total	2,574	2,276	2,148	6,998

Class balance. Forged samples make up 63.2% of HIM-DF and 62.0% of the test split, so a constant "fake" predictor obtains 62.0% test accuracy. We therefore report the threshold-independent AP and AUC alongside accuracy, and use the oracle ceilings in Table 5 as a more informative reference than chance.

Instrument coverage and duration. Table 10 gives the per-instrument breakdown. The benchmark contains approximately 4.86 h of audio-visual content, with each sample represented as a fixed 2.5 s clip.

Table 10: Per-instrument statistics of HIM-DF. Durations are rounded per row, so they may not sum exactly to the total.

Instrument	Family	Sources	Real	M2V	V2M	Duration (min)
Piano	Keyboard	34	322	288	270	36.7
Handpan	Idiophone	17	169	151	142	19.3
Guitar	Plucked string	42	398	350	331	45.0
Harp	Plucked string	20	189	168	158	21.5
Oud	Plucked string	19	173	151	142	19.4
Zither	Plucked string	21	199	176	165	22.5
Violin	Bowed string	32	303	269	253	34.4
Cello	Bowed string	22	208	184	174	23.6
Trumpet	Brass	19	180	159	150	20.4
Clarinet / oboe	Woodwind	16	151	134	126	17.1
Trombone	Brass	15	142	126	118	16.1
Tuba	Brass	15	140	120	119	15.8
Total		272	2,574	2,276	2,148	291.6

A.10 Release, Responsible Use, and Scope

Release. We will release the split files, all generated samples, the structured measurements with observation masks, the CoT annotations with provenance flags, and the expert quality ratings. Authentic clips will be redistributed only when permitted by their source licenses. For all other clips, we will provide source URLs, timestamps, and a retrieval script. Generated media will be distributed under the HIM-DF Research-Only Data License and will carry metadata identifying them as synthetic. We will honor removal requests from performers and rights holders.

Scope. HIM-DF is intended to evaluate whether a detector can verify the Human–Instrument–Music production chain. Several aspects are deliberately out of scope for this version. (1) Each forgery direction uses a single generator. The paired construction prevents detectors from exploiting the authentic modality, but a detector may still learn traces specific to that generator. Appendix C.1 evaluates HIM-CoT on held-out generators for both directions, and the construction pipeline is generator-agnostic, so further generators can be added in future versions. (2) Only solo performances are included, because ensembles introduce source separation and attribution problems. (3) There are no samples in which both modalities are forged. Fully synthetic performances test the realism of each modality separately rather than production-chain consistency, and we leave them to future versions. (4) Coverage of instruments, genres, and recording conditions follows the collected sources and is skewed toward plucked-string instruments and indoor recordings, so results on less-represented instruments and settings should be interpreted with reference to Table 10.

B Implementation Details

Our implementation is based on Qwen2.5-Omni-7B (Xu et al., 2025), a multimodal language model with pretrained audio and video encoders. The structured feature extraction modules produce 69-dimensional measurements organized into four groups: music (21 dimensions), visual (20 dimensions), contact (18 dimensions), and global (10 dimensions), defined in Table 8. Each group is projected to 512-dimensional tokens via group-specific MLPs. Low-rank adaptation (LoRA) (Hu et al., 2022) is applied to all attention projection layers (queries, keys, values, and outputs) with rank $r = 8$, scaling factor $\alpha_{\text{LoRA}} = 16$, and dropout 0.05. The group encoders, cross-attention, gating, and classification modules are trained in full. The model is trained for 20 epochs with AdamW at a constant learning rate of 1×10^{-5}, using one clip per GPU on 4 GPUs with gradient accumulation over 10 steps, for an effective batch size of 40. Loss weights are $\lambda_f = \lambda_r = 1.0$. Within $\mathcal{L}_{\text{reason}}$, the three stages are averaged with weights 1.0, 0.75, and 1.25 for the music, performance, and cross-modal stages. At inference, reasoning is generated greedily with at most 768 new tokens. The

cross-modal stage embedding uses the discrepancy of the global group, whose features summarize cross-modal statistics, while the discrepancies of the other groups reach the decoder through their evidence tokens $\tilde{E}_k$. The measurements enter the model only through the evidence tokens. Textual evidence summaries derived from them are used only to construct the reasoning targets for training (Appendix A.7) and are never part of the model input, so no textual description of the measurements is available at test time.

B.1 Baseline Training and Evaluation Protocol

Split consistency. All trainable baselines use the same source-grouped split as HIM-CoT (5,598 / 700 / 700 clips of 2.5 s, Table 9). Every baseline's training, validation, and test inputs are generated from the same source-to-split assignment as HIM-CoT. We verify for every baseline that its predictions cover exactly the 700 test clips and that none of its training clips comes from a validation or test source.

Training. All trainable baselines are trained from their public pretrained weights with the recipes of the original papers. AltFreezing (Zhao et al., 2023) is trained on 32 uniformly sampled frames per clip at 224×224, with its temporal-drop, temporal-repeat, and clip-blending synthesis applied to visually real clips. Self-Blended Images (Shiohara & Yamasaki, 2022) follows its original protocol: an EfficientNet-B4 classifier is trained with SAM on pseudo-fakes blended from pristine frames of the training split, and the checkpoint is selected by validation AUC. Because SBI operates on detected faces and the performer's face is not visible in every clip, frame scores are averaged over the frames in which a face is detected, and clips with no detected face receive a fake probability of 0.5, so SBI is evaluated on all 700 test clips. The WavLM detector (Liu et al., 2024b) trains its multi-fusion attentive classifier on a frozen WavLM-Base-Plus encoder. AASIST (Jung et al., 2022) is trained on the audio track of each clip with its official configuration. ICS-AV (Anshul et al., 2025) first performs its intra- and cross-modal synchronization pretraining on the training split and then trains its binary detector. AVFF (Oorloff et al., 2024) trains its audio-visual representation stage followed by its classification stage. T-AVFD (Liu et al., 2026) is trained with its official configuration on the same split. FaceCoT (Zhang et al., 2026) is fine-tuned in its two stages on the same clips, using its own reasoning format with HIM-DF labels. Unless stated otherwise, every trained baseline selects its checkpoint on the validation split, and the test split is used only once for the reported numbers. On LAV-DF, all trainable baselines are retrained on the official LAV-DF training split in the same way.

Late fusion of unimodal detectors. This baseline asks how far HIM-DF can be solved by detecting generation artifacts in each modality separately, without relating the two. We take the strongest audio detector (WavLM) and the strongest video detector (AltFreezing) by validation AUC, each trained as described above, and obtain their fake probabilities p_a and p_v on the validation and test splits. The fused probability is $\sigma(w_0 + w_a \operatorname{logit}(p_a) + w_v \operatorname{logit}(p_v))$, where the weights are fitted by L2-regularized logistic regression on the validation predictions only. A clip is predicted as fake when the fused probability is at least 0.5. Because each forged clip contains exactly one synthesized modality, a perfect artifact detector in each modality would make this baseline solve HIM-DF, so its gap to HIM-CoT measures what cross-modal reasoning adds beyond artifact detection.

Label-only Qwen. This baseline isolates the backbone and supervised adaptation from the contributions of HIM-CoT. It uses the same Qwen2.5-Omni-7B backbone, LoRA configuration, optimizer, number of epochs, and audio–video preprocessing as HIM-CoT (Appendix B). The prompt contains only the clip and an instruction to judge its authenticity. There are no structured measurements, evidence tokens, cross-attention modules, or reasoning targets. Authenticity and forgery-type heads, identical to those of HIM-CoT, are applied to the final-layer hidden state of the last prompt token and trained with $\mathcal{L}_{\text{auth}} + \lambda_f \mathcal{L}_{\text{type}}$. The checkpoint is selected by validation AUC. It differs from the media-only ablation of Table 2 ("w/o Structured Evidence"), which keeps the reasoning targets and classifies from the generated reasoning.

Zero-shot models. Gemini 3.8 Flash (Gemini Team, 2026) and MiniMax H3 (MiniMax, 2026) receive each clip with a prompt that asks for a real or fake verdict and a confidence in $[0, 1]$ in JSON format, without any training on HIM-DF. Accuracy is computed from the returned verdict, and AP

and AUC are computed from the confidence converted to a fake probability. Responses that cannot be parsed are retried, and remaining failures are counted as incorrect. For the explanation analyses in Appendix C.2, the value-fidelity measurement and the human evaluation, the same models are additionally prompted to explain their verdicts, and only these explanations are evaluated.

Decision threshold. For every trained model, including HIM-CoT, a clip is predicted as fake when its predicted fake probability is at least 0.5 (equivalently, the arg-max of a two-class output). No threshold is tuned on the test split, so accuracy reflects each model's own calibration, whereas AP and AUC are threshold-independent. Because 62.0% of HIM-DF test clips are fake, a poorly calibrated detector can fall below the accuracy of a constant "fake" prediction while still ranking samples above chance, which is why AP and AUC are reported alongside accuracy. The same applies to LAV-DF, where about 73% of videos are fake: SBI, AltFreezing, and AVFF reach accuracy close to the fake rate, which indicates near-constant predictions at the 0.5 threshold, while their AUC shows limited but above-chance ranking. For these face- and speech-oriented detectors, AUC is therefore the more informative measure.

B.2 Adapting HIM-CoT to LAV-DF

The structured measurements of HIM-DF describe a Human–Instrument–Music chain, but LAV-DF contains talking faces without an instrument. We therefore adapt the measurements to the analogous chain in speech, in which the motion of the face and articulators, such as the lips and jaw, produces the speech signal. The four evidence groups, the model architecture, and the training objective are otherwise unchanged.

Measurements. Table 11 maps each evidence group to its talking-face counterpart. The audio, visual, contact, and global groups have 18, 16, 7, and 28 dimensions. When the face or the voice is not detected in a window, the corresponding entries are marked as missing by the observation masks, which, unlike on HIM-DF, are not constant on LAV-DF (12% of windows have at least one missing group).

Table 11: Evidence groups on HIM-DF and their counterparts on LAV-DF.

Group	HIM-DF (instrumental performance)	LAV-DF (talking face)
Audio x_M	Pitch trajectory, onsets, spectral harmonics, and dynamic envelope (pYIN, CREPE)	Voice F0 (pYIN), energy envelope, syllable and onset rate, formant and spectral statistics, voicing
Visual x_V	Hand and body kinematics (MediaPipe Hand/Pose)	Lip and jaw kinematics (MediaPipe Face Mesh): mouth opening, lip velocity, jaw displacement
Contact x_C	Performer–instrument contact topology and dynamics	Articulatory contact: bilabial closure events and their agreement with bilabial phonemes, and viseme–phoneme consistency
Global x_G	Optical flow, CLIP, and onset–motion synchronization	Audio–visual synchronization (SyncNet offset and confidence), correlation between energy envelope and mouth opening, optical flow and CLIP statistics

Labels and forgery types. LAV-DF contains visual-only, audio-only, and audio–visual manipulations in addition to real videos. We therefore extend the forgery-type head to four classes (Real, visual-only, audio-only, and both), and the three-stage reasoning targets name the manipulated modality or modalities. The authenticity head and $\mathcal{L}_{\text{auth}}$ are unchanged.

Windows and video-level prediction. LAV-DF manipulations are local and often replace only a few words, so a 2.5 s window of a fake video may contain no manipulated content. During training, each window is labeled fake only if it overlaps a manipulated segment according to the official timestamps. At test time, a video's fake probability is the maximum over its windows, so a video is flagged when any window is, and all metrics in Table 1 are computed at the video level.

Reasoning targets. Reasoning targets are generated with the same label-conditioned protocol as on HIM-DF (Appendix A.7), using the same annotator. The stage prompts are adapted to speech plausibility, facial-motion plausibility, and cross-modal consistency. The automatic consistency checks are applied with check (c) adapted accordingly: for a visual-only manipulation, the speech stage must not claim that the audio is synthetic, and vice versa, while no such constraint applies to audio–visual manipulations. Two annotators verify 400 targets (Cohen's $\kappa = 0.86$).

Training data. Generating reasoning targets for all 136,304 videos is costly, so HIM-CoT is trained on a subset of 8,000 training videos sampled uniformly within each manipulation type, and selected on 1,000 validation videos. It is evaluated on the full official test split. Baselines use the full official training split, as described in Appendix B.1. Hyperparameters follow Appendix B.

Contribution of the adapted evidence. Table 12 compares HIM-CoT on LAV-DF with its variant without structured evidence, which uses only the Qwen media tokens with the same reasoning supervision. Removing the adapted measurements lowers accuracy by 2.8% and AUC by 2.3%, so the talking-face evidence contributes beyond the backbone. This drop is much smaller than the 23.0% on HIM-DF, consistent with speech–lip coupling being more directly visible to the pretrained media encoders than instrument-mediated sound production, which leaves less for explicit measurements to add.

Table 12: Contribution of the adapted structured evidence on LAV-DF.

Model	ACC (%)	AP (%)	AUC (%)
HIM-CoT w/o Structured Evidence	87.8	89.4	88.6
HIM-CoT	**90.6**	**91.2**	**90.9**

C Additional Experimental Results

C.1 Cross-Generator Generalization

Because each forgery direction in HIM-DF is produced by a single generator, a detector could separate real from fake by recognizing generator-specific traces rather than violations of the production chain. The key question is whether the detector transfers to a new generator while preserving the same production-chain reasoning. We therefore keep the training set unchanged, with Seedance 2.0 for M2V and VidMuse for V2M, and build two additional test sets from the authentic test-split clips. For unseen M2V, each authentic test track is paired with a video generated by Kling AI (Kling Team et al., 2025) from the same reference frame and prompt template. For unseen V2M, each authentic test video is paired with music generated by MuMu-LLaMA (Liu et al., 2024a). Both sets pass through the same format harmonization and automated quality screening as HIM-DF (Appendices A.5 and A.6), and each is evaluated together with the 266 authentic test samples. No model is retrained or recalibrated on the new generators.

Table 13 compares HIM-CoT with the strongest baseline of each group and with HIM-CoT trained without structured evidence. Each unseen set contains forgeries of only one direction. On unseen M2V, the audio remains authentic, so an audio-only detector cannot observe the changed modality. On unseen V2M, the video remains authentic, so the analogous limitation holds for a video-only detector. The comparison of interest is therefore between HIM-CoT and its variant without structured evidence. If detection relied mainly on generator traces, both models would degrade similarly, whereas evidence grounded in the production chain should transfer across generators. All models degrade on unseen generators, but to very different extents. HIM-CoT reaches 81.4% accuracy and 87.2% AUC on unseen M2V forgeries, and 78.9% accuracy and 84.6% AUC on unseen V2M forgeries. These correspond to accuracy drops of 7.9% and 10.4% relative to the seen setting. Without structured evidence, the variant falls close to chance, and HIM-CoT outperforms it by 26.3% and 26.1% in accuracy. Measured as the AUC margin above chance, HIM-CoT retains 84% and 78% of its margin on the seen generators, about two and nearly three times the retention of the variant without structured evidence. The media-only variant therefore appears to rely more heavily on generator-specific traces, while the structured measurements transfer better across generators.

With a single unseen generator per direction, however, this does not rule out that the measurements also capture artifacts shared across generators. Appendix C.8 therefore adds a leave-one-generator-out evaluation with three generators per direction. HIM-CoT also outperforms T-AVFD by 17.5% and 17.1%. As predicted by the construction, WavLM falls to chance-level AUC on unseen M2V, where the audio is authentic, and AltFreezing on unseen V2M, where the video is authentic. Finally, V2M transfers worse than M2V for each audio–visual model shown, suggesting that the acoustic cues learned from VidMuse outputs are more generator-specific than the kinematic and contact cues learned from Seedance 2.0 videos.

Table 13: Cross-generator generalization on HIM-DF. All models are trained on the original training split, using Seedance 2.0 for M2V and VidMuse for V2M. *Seen* is the original test split (Table 1). *Unseen M2V* replaces the generated videos with Kling AI outputs, and *Unseen V2M* replaces the generated music with MuMu-LLaMA outputs. Each unseen set combines the held-out forgeries with the same authentic test samples.

	Seen		**Unseen M2V**		**Unseen V2M**	
Method	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
WavLM (audio)	64.4	69.9	50.2	50.0	57.3	61.8
AltFreezing (video)	62.0	58.1	55.8	58.9	49.8	51.2
T-AVFD	74.3	71.4	63.9	66.8	61.8	64.5
HIM-CoT w/o Structured Evidence	66.3	67.5	55.1	57.0	52.8	54.9
HIM-CoT	**89.3**	**94.2**	**81.4**	**87.2**	**78.9**	**84.6**

C.2 Attribution and Reasoning Evaluation

This section details the evaluation summarized in Section 5.4. All quantities are computed on the 700 test clips from the same predictions as Table 1, without retraining.

Three-way classification and attribution. The predicted forgery type is the arg-max of the forgery-type head over {Real, M2V, V2M}. We report three-way accuracy and macro-F1 over all clips, and attribution accuracy, defined as the fraction of forged clips (M2V or V2M) whose predicted type is their true forgery type. Label-only Qwen uses its own forgery-type head. Late fusion has no type head, so a clip whose fused probability is below 0.5 is labeled Real. Otherwise it is labeled V2M if the audio detector's logit, standardized on the validation split, exceeds the video detector's, and M2V otherwise. Table 14 gives the confusion matrix of HIM-CoT. Of the 434 forged clips, 359 are attributed to the correct modality, 41 are missed as Real (23 M2V and 18 V2M), and 34 are attributed to the wrong modality. Most of these misattributions are M2V clips blamed on the audio (26), while only 8 V2M clips are blamed on the video. Among the 266 real clips, the type head assigns 33 to a forgery type, split almost evenly between M2V (17) and V2M (16). The authenticity decisions reported elsewhere, for example in Tables 1, 20, and 21, come from the authenticity head, whereas forgery types and attribution come from the separate forgery-type head. The two heads can disagree on a small fraction of clips. For example, the authenticity head flags 25 of the 266 real clips (9.4%) as fake, while the type head assigns 33 of them to a forgery type. For the same reason, a model's three-way accuracy can exceed its binary accuracy, as for the variant without structured evidence in Table 3.

Table 14: Three-way confusion matrix of HIM-CoT on the HIM-DF test split. Rows give the true type and columns the predicted type.

True \ Predicted	**Real**	**M2V**	**V2M**	**Total**
Real	233	17	16	266
M2V	23	185	26	234
V2M	18	8	174	200

Agreement between reasoning and decision. The final stage of each generated chain states an authenticity verdict and, for fake clips, the forged modality. We parse this statement with rule-based matching and compare it with the classification head, counting chains whose verdict cannot

be parsed as disagreements. The authenticity verdict stated in the reasoning agrees with the head for 95.4% of HIM-CoT's clips, 6.7% more than for the variant without structured evidence (Table 3), so the generated explanation almost always states the decision the model actually makes.

Value fidelity. We extract every numerical claim in the generated reasoning that refers to a named measurement, such as a pitch estimate, onset rate, or contact fraction, and compare it with the corresponding entry of the 69-dimensional measurement vector. A value is counted as correct when it lies within 10% of the measurement. Claims about quantities outside the measurement set are counted as unsupported. Table 15 reports the fractions of correct, incorrect, and unsupported claims, including for the zero-shot models when prompted to explain their verdicts. The metric should be read with the models' inputs in mind. HIM-CoT receives the measurements only through its evidence tokens, not as text (Appendix B), and is trained on reasoning targets that cite them. Citing a correct value therefore requires recovering it from the evidence tokens, and for HIM-CoT, fidelity measures whether the model reproduces its evidence faithfully rather than hallucinating values. For models without access to the measurements, incorrect and unsupported claims measure how often they state specific quantities that no measurement supports. HIM-CoT cites 5.1 values per clip, of which 91.8% match the measurements and only 4.0% are unsupported. Its correct rate is 18.2% higher than without structured evidence, and 49.8% and 56.8% higher than Gemini 3.8 Flash and MiniMax H3, which leave about a third of their claims unsupported, consistent with the unsupported fine-grained claims in Figure 7.

Table 15: Fidelity of numerical claims in generated reasoning on the HIM-DF test split. Fractions are over all extracted numerical claims.

Method	Claims / clip	Correct (%)	Incorrect (%)	Unsupported (%)
Gemini 3.8 Flash	4.8	42.0	25.0	33.0
MiniMax H3	5.2	35.0	29.0	36.0
HIM-CoT w/o Structured Evidence	3.6	73.6	12.4	14.0
HIM-CoT	**5.1**	**91.8**	**4.2**	**4.0**

Human evaluation. We sample 100 test clips stratified by forgery type and by whether HIM-CoT's prediction is correct, with correct and incorrect predictions in proportion to their frequency on the test split. For each clip, three raters, including two professional musicians, view the clip and the reasoning produced by HIM-CoT, HIM-CoT without structured evidence, and Gemini 3.8 Flash, shown in random order without method names. Each reasoning stage is rated as factually correct or not with respect to the audio and video, and as grounded or not, that is, supported by perceptible evidence or by the measurements rather than speculation. The final stage is additionally rated for whether it attributes the forgery to the correct modality, and the whole chain receives a usefulness score from 1 to 5. Table 16 reports the results. Inter-rater agreement is Krippendorff's $\alpha = 0.87$. HIM-CoT is rated highest on every criterion: 88.0% of its stages are factually correct and 84.0% are grounded, its final stage attributes 82.0% of forged clips to the correct modality, and its mean usefulness is 4.1. It exceeds the variant without structured evidence by 14.0%, 21.0%, and 17.0% and by 0.9 in usefulness, and Gemini 3.8 Flash by 20.0%, 35.0%, and 27.0% and by 1.2 in usefulness. The largest gain from structured evidence is in groundedness. Although method names are hidden, HIM-CoT's reasoning can be recognized by its references to measured values, so the evaluation is not fully blind. The attribution criterion depends only on the stated verdict and is least affected by this.

Table 16: Blind human evaluation of generated reasoning on 100 HIM-DF test clips. Factual and grounded rates are over reasoning stages. Attribution is over forged clips.

Method	Factual (%)	Grounded (%)	Attribution (%)	Usefulness (1–5)
Gemini 3.8 Flash	68.0	49.0	55.0	2.9
HIM-CoT w/o Structured Evidence	74.0	63.0	65.0	3.2
HIM-CoT	**88.0**	**84.0**	**82.0**	**4.1**

C.3 Observation-Only Reasoning Control

The classifier of HIM-CoT reads the hidden states of the generated reasoning, and the default reasoning targets end with explicit verdicts, such as "fake" or the forged modality. The gain over the reasoning-free variant could therefore come from the model stating the label in text and the classifier reading those words, rather than from the reasoning itself. To rule this out, we train a variant whose reasoning targets contain only objective observations.

Constructing verdict-free targets. Starting from the HIM-DF reasoning targets, we remove the final conclusion field and every judgmental statement about reliability, and rewrite the remaining sentences so that they describe observations without assessing them. The following words and their inflections, synonyms, and negations are removed: *fake*, *real*, *authentic*, *genuine*, *synthetic*, *generated*, *forged*, *manipulated*, *suspicious*, *plausible*, *implausible*, *natural*, *consistent*, *inconsistent*, *mismatch*, *M2V*, and *V2M*. Rewriting combines rule-based substitution with an LLM rewrite under the instruction to preserve every observation and measured value. An automatic lexicon check confirms that no listed word remains in any target. Two annotators inspect 400 rewritten targets and find that 96.5% preserve all observations of the original and 3.5% contain an implicit verdict.

Training and evaluation. The variant is identical to HIM-CoT in inputs, evidence modules, three-stage structure, classification from the pooled hidden states of the generated stages, and training schedule. Only the reasoning targets differ. At test time, 8.5% of the generated chains contain a listed word, compared with 74.0% for HIM-CoT, so a stated verdict is rarely available to the classifier of the observation-only variant.

Table 17: Observation-only reasoning control on the HIM-DF test split. Agreement between reasoning and decision is not applicable to the observation-only variant, whose reasoning states no verdict.

Variant	ACC	AP	AUC	3-way ACC	Attr.	Fidelity
w/o Three-Stage Reasoning	82.4	87.2	87.8	76.9	72.8	–
Observation-Only Reasoning	87.6	93.7	92.9	81.7	78.3	87.8
HIM-CoT	**89.3**	**95.0**	**94.2**	**84.6**	**82.7**	**91.8**

Results. Observation-only reasoning reaches 87.6% accuracy, 5.2% above the reasoning-free variant, and thus recovers about three quarters (75%) of the 6.9% gain of HIM-CoT over that variant. Generating and conditioning on evidence-grounded observations therefore improves the prediction even when the reasoning states no verdict. The remaining 1.7% shows that stating a verdict gives the classifier a smaller additional benefit. The same pattern holds for attribution: the observation-only variant's three-way accuracy and attribution are 4.8% and 5.5% above the reasoning-free variant and 2.9% and 4.4% below HIM-CoT, and 87.8% of the numerical values it cites still match the measurements. Staged reasoning thus helps mainly through the observations it generates and only partly through the verdict it states.

C.4 Performance on High-Plausibility Forgeries

HIM-DF labels encode provenance, so a forged clip is fake even when its performance plausibly explains the sound. To test whether HIM-CoT relies only on obvious cross-modal inconsistencies, we evaluate it separately on forgeries of high and low perceptual plausibility.

Protocol. Three professional musicians rate all 434 forged test clips on the perceptual-plausibility criterion of the quality-control rubric (Table 6), without knowing the forgery direction or any model's prediction. Inter-rater agreement is Krippendorff's $\alpha = 0.82$. A clip is *high-plausibility* when its mean rating is at least 4, that is, when a musician would accept the performance as producing the sound on casual viewing, and *low-plausibility* otherwise. This gives 352 high-plausibility clips (187 M2V, 165 V2M) and 82 low-plausibility clips. Because each subset contains only forged clips, we report the detection rate (the fraction predicted fake at the 0.5 threshold), AUC against all 266 real test clips, and modality attribution. Modality attribution compares only the M2V and V2M scores of the forgery-type head, or, for late fusion, the validation-standardized audio and video logits. It

thus asks which modality the model would blame, independently of whether it flags the clip, and is higher than the three-way attribution in Table 3, which also counts forged clips predicted as Real as errors. All models are evaluated from their existing test predictions without retraining.

Table 18: Detection of forged test clips grouped by expert-rated perceptual plausibility. *Det.*: fraction of forged clips predicted fake. *AUC*: forged clips in the subset against all 266 real test clips. *Attr.*: modality attribution, i.e., the fraction of forged clips whose M2V-versus-V2M prediction matches the forged modality, regardless of the authenticity decision. Unlike the three-way attribution in Table 3, a forged clip predicted as Real can still count as correctly attributed. "–": no forgery-type output.

	Low plausibility ($n = 82$)			**High plausibility** ($n = 352$)		
Method	**Det.**	**AUC**	**Attr.**	**Det.**	**AUC**	**Attr.**
T-AVFD	89.0	91.4	–	65.1	66.7	–
Late fusion (WavLM + AltFreezing)	82.9	86.5	67.1	50.3	67.9	50.0
Qwen2.5-Omni LoRA (label only)	86.6	89.4	70.7	45.5	56.7	55.7
HIM-CoT w/o Structured Evidence	87.8	91.0	75.6	46.0	62.0	62.2
HIM-CoT	**96.3**	**98.1**	**90.2**	**86.6**	**93.2**	**81.8**

Results. Table 18 reports the results. All methods detect fewer high-plausibility forgeries, but HIM-CoT degrades least: its detection rate falls from 96.3% to 86.6% (-9.7%), a drop at least 14.2% smaller than that of any other method. On high-plausibility forgeries, HIM-CoT detects 21.5% more than the strongest baseline, T-AVFD, reaches 93.2% AUC, and identifies the forged modality for 81.8% of them, 31.8% more than late fusion and 19.6% more than without structured evidence. Across all forged clips, the Spearman correlation between the plausibility rating and HIM-CoT's fake probability is $\rho = -0.38$, so more plausible forgeries receive only moderately lower scores. HIM-CoT therefore does not rely only on obvious mismatches between performance and sound. On plausible forgeries, it must draw on subtler cues, including residual artifacts of the synthesized modality that the structured measurements expose. This is consistent with the provenance definition of the labels, and we do not claim that every detected forgery is physically implausible.

C.5 Artifact-Sensitive Features

Several structured measurements describe how reliably an extractor worked rather than what the performance or the sound contains. Generated media can degrade these extractors, so such features may let a model detect synthesis artifacts instead of production-chain inconsistencies. To measure how much of HIM-CoT's gain rests on them, we split the 69 measurements of Table 8 into two sets and retrain HIM-CoT on each.

Feature split. The *artifact-sensitive* set contains the 18 dimensions that quantify extractor reliability: the CREPE and pYIN confidence statistics (dimensions 20–23 and 31–34), the disagreement between the two pitch estimators (29–30), the bone-length instability of the hand skeleton (40–41), the three tracking flags (50–52), the jitter ratio (54), the MediaPipe detection confidence (56), and the analytic-plane fallback indicator of the contact surface (19). The *production-chain* set contains the remaining 51 dimensions, which describe pitch and voicing, hand kinematics, hand–instrument contact, optical flow, CLIP temporal statistics, and onset–motion synchronization.

Protocol. Both variants use the same architecture, evidence groups, reasoning targets, and training schedule as HIM-CoT. Only the input measurements differ, and a removed dimension is treated as missing by the observation masks. We report detection and attribution on the HIM-DF test split, the increase in the fraction of real–real mismatched pairs flagged as fake over genuine pairs (Appendix C.6), and the mean accuracy on the two unseen-generator test sets of Appendix C.1. The last two are evaluated from the trained checkpoints without retraining.

Results. Table 19 reports the results. Artifact-sensitive features alone provide only limited gains over using no structured evidence, increasing accuracy from 66.3% to 69.1%, AUC from 67.5% to 71.8%, and attribution accuracy from 59.7% to 62.4%. Their mismatch sensitivity is 37.2%,

Table 19: Contribution of artifact-sensitive features. *Mismatch*: increase in the fraction of real–real mismatched pairs flagged as fake over genuine pairs. *Unseen gen.*: mean accuracy on the unseen M2V and unseen V2M test sets.

Input measurements	Dims	ACC	AUC	Attr.	Mismatch	Unseen gen.
None (w/o Structured Evidence)	0	66.3	67.5	59.7	34.6	54.0
Artifact-sensitive only	18	69.1	71.8	62.4	37.2	57.1
Production-chain only	51	86.4	91.6	79.7	49.6	76.8
All (HIM-CoT)	69	**89.3**	**94.2**	**82.7**	**52.3**	**80.2**

only 2.6% above the variant without structured evidence, and their unseen-generator accuracy is 57.1%. In contrast, the production-chain measurements alone reach 86.4% accuracy, 91.6% AUC, and 79.7% attribution accuracy. They retain a mismatch sensitivity of 49.6% and 76.8% accuracy on unseen generators, close to the corresponding full-model results of 52.3% and 80.2%. Adding the artifact-sensitive measurements to the production-chain set yields a further 2.9% accuracy gain. These results show that artifact-sensitive features provide a modest complementary signal, while most of the structured-evidence benefit and sensitivity to mismatched performances comes from measurements of the production chain itself.

C.6 Real–Real Mismatched Pairs

HIM-DF labels encode provenance, so every forged clip contains a synthesized modality. A detector could therefore succeed by recognizing synthesis artifacts alone, without using whether the performance explains the sound. To separate the two, we build pairs in which neither modality is synthesized but the production chain is broken.

Protocol. For each of the 266 real test clips, we pair its video with the audio of a real test clip of the same instrument from a different source, chosen at random, and pass the pair through the same format harmonization as HIM-DF. This yields 266 mismatched pairs whose audio and video are both authentic and individually unchanged apart from re-encoding. We evaluate all models from their trained checkpoints without retraining and report the fraction of pairs predicted fake. For models that generate reasoning, we also report the fraction whose cross-modal stage states that the performance does not explain the sound.

Table 20: Real–real mismatched pairs. *Genuine*: fraction of the 266 real test clips predicted fake. *Mismatched*: fraction of real video–real audio pairs from different performances predicted fake. *Stage*: fraction of mismatched pairs whose cross-modal reasoning stage reports an inconsistency. "–": no generated reasoning.

Method	Genuine (%)	Mismatched (%)	Stage (%)
T-AVFD	18.0	42.1	–
Late fusion (WavLM + AltFreezing)	16.5	38.0	–
Qwen2.5-Omni LoRA (label only)	20.3	35.7	–
HIM-CoT w/o Structured Evidence	13.5	48.1	52.6
HIM-CoT	9.4	61.7	77.8

Results. Table 20 reports the results. HIM-CoT flags 61.7% of the mismatched pairs as fake, 52.3% more than genuine pairs, and its cross-modal stage reports an inconsistency for 77.8% of them, compared with 24.1% of genuine clips (Table 21) and 25.2% more than without structured evidence. Because both modalities are authentic, these predictions cannot come from synthesis artifacts in either modality. Late fusion scores each modality separately and cannot detect a broken pairing, yet it also flags 21.5% more mismatched pairs than genuine ones. Part of the increase therefore reflects changes introduced by re-pairing, such as re-encoding, rather than the pairing itself. The increase of HIM-CoT exceeds that of late fusion by 30.8% and that of the variant without structured evidence by 17.7%, so HIM-CoT does use the consistency between performance and sound as evidence, and the structured measurements strengthen this sensitivity. HIM-CoT also flags the fewest genuine pairs, 4.1% fewer than the next-best method, so its sensitivity to mismatches does not come at the cost of more false alarms. Because the two clips of a mismatched pair come

from different sources, they also differ in room acoustics, microphone, and tempo, so a model could flag them as coming from different recordings rather than as physically inconsistent. The within-recording control below removes these differences. HIM-DF itself contains no mismatched pairs, so these probes measure sensitivity to broken production chains rather than detection accuracy on the benchmark labels.

C.7 Within-Recording Temporal Mismatches

The real–real mismatches above pair clips from different sources and therefore also change the recording conditions. To isolate the production chain, we pair each real test video with audio from the same recording, shifted in time. The room, microphone, performer, instrument, and piece are then unchanged, and only the correspondence between the visible performance and the sound is broken.

Protocol. For each of the 266 real test clips, we take the audio of the same source recording starting Δ seconds later, for $\Delta \in \{0.5, 1, 2, 4\}$ s, keep the video unchanged, and pass the pair through the same format harmonization as HIM-DF. Clips whose source recording is too short for a shift are skipped, which leaves 264, 260, 252, and 236 pairs for the four shifts. $\Delta = 0$ corresponds to the genuine clips. Because music often repeats, a small shift may still roughly match the performance, whereas a model that checks precise timing should flag more pairs as Δ grows. All models are evaluated from their trained checkpoints without retraining.

Table 21: Within-recording temporal mismatches. Each real test video is paired with audio from the same recording shifted by Δ seconds. Values are the fraction of pairs predicted fake (%). The last row gives the fraction of pairs for which the cross-modal stage of HIM-CoT reports an inconsistency.

Method	$\Delta = 0$	0.5 s	1 s	2 s	4 s
T-AVFD	18.0	19.3	21.2	21.4	24.2
Late fusion (WavLM + AltFreezing)	16.5	17.4	18.1	19.4	23.3
HIM-CoT w/o Structured Evidence	13.5	14.8	15.0	16.3	18.6
HIM-CoT	9.4	18.2	19.2	22.2	22.9
HIM-CoT cross-modal stage	24.1	26.5	27.7	29.0	29.2

Results. Table 21 reports the results. All methods flag more pairs as the temporal shift grows, but HIM-CoT shows the largest change. Its fake prediction rate rises from 9.4% on genuine clips to 22.9% at a 4 s shift, an increase of 13.5%. The corresponding increases are 6.2% for T-AVFD, 6.8% for late fusion, and 5.1% for HIM-CoT without structured evidence. The increase for the full model is therefore nearly twice that of late fusion and 8.4% larger than that of the variant without structured evidence. Its cross-modal stage reports an inconsistency for 24.1% of genuine clips and 29.2% of the 4 s-shifted pairs, a smaller increase of 5.1%. These results show that the structured measurements make the final decision more sensitive to within-recording temporal displacement. However, even at a 4 s shift, HIM-CoT flags only 22.9% of the pairs, far below the 61.7% obtained when audio comes from a different performance (Appendix C.6). The model therefore uses precise timing as one signal, while broader agreement in instrument timbre, register, playing style, and recording context remains more influential. The modest change in the generated cross-modal stage also indicates that this timing sensitivity is expressed more strongly in the classification decision than in the verbal explanation.

C.8 Leave-One-Generator-Out Evaluation

The main HIM-DF split uses one generator per forgery direction, so a detector may learn traces specific to that generator. Appendix C.1 tests transfer to one unseen generator per direction. Here we extend this to three generators per direction and hold each one out in turn.

Generators and folds. For M2V, we use Seedance 2.0 (ByteDance Seed Team, 2026), Kling AI (Kling Team et al., 2025), and Wan 3.0 (Alibaba Cloud, 2026). For V2M, we use VidMuse (Tian et al., 2025), MuMu-LLaMA (Liu et al., 2024a), and Video2Music (Kang et al., 2024), an affective

multimodal transformer that generates symbolic music (MIDI) from semantic, motion, and emotion features of the video, which we render to audio with a soundfont. Because its audio is rendered from symbolic music rather than generated as a waveform, it differs in kind from the other two V2M generators. Each generator synthesizes forgeries from the same source clips with the same conditioning, format harmonization, and automated quality screening as HIM-DF (Appendices A.5 and A.6), and the source-grouped split is unchanged. In fold k, the k-th M2V and k-th V2M generators are held out: models are trained and selected on forgeries from the other two generators of each direction and tested on the held-out generators together with the 266 real test clips. No model is recalibrated on the held-out generators.

Table 22: Leave-one-generator-out evaluation on HIM-DF. Each fold holds out one M2V and one V2M generator, and models are trained on the other two generators of each direction. Values are ACC / AUC (%) on the held-out generators plus the real test clips.

Method	Fold 1 Seedance / VidMuse	Fold 2 Kling / MuMu-LLaMA	Fold 3 Wan 3.0 / Video2Music	Mean
T-AVFD	68.4 / 72.1	66.9 / 69.8	61.7 / 64.5	65.7 / 68.8
Late fusion (WavLM + AltFreezing)	64.7 / 68.9	62.8 / 66.7	58.6 / 61.5	62.0 / 65.7
Qwen2.5-Omni LoRA (label only)	62.4 / 66.8	60.9 / 64.9	56.8 / 59.7	60.0 / 63.8
HIM-CoT w/o Structured Evidence	66.1 / 70.8	64.5 / 68.4	59.7 / 62.9	63.4 / 67.4
HIM-CoT	**83.1 / 89.2**	**81.6 / 87.5**	**76.8 / 82.4**	**80.5 / 86.4**

Results. Table 22 reports the results. Averaged over the three folds, HIM-CoT reaches 80.5% accuracy and 86.4% AUC on held-out generators, 14.8% and 17.6% above the strongest baseline, T-AVFD, and 17.1% above its variant without structured evidence in accuracy. It leads T-AVFD by at least 14.7% in every fold, and its accuracy varies by 6.3% across folds, so the result does not depend on a single generator. All methods perform worst on the fold that holds out Wan 3.0 and Video2Music. The drop of HIM-CoT from its in-distribution accuracy (8.8%) is comparable to that of the single-generator transfer in Appendix C.1 (7.9% and 10.4%), even though each fold holds out an M2V and a V2M generator at the same time. Training on more generators thus does not close the gap to in-distribution performance, but the large and stable margin over all baselines indicates that the structured measurements capture cues that transfer across generators better than those learned by the baselines.

C.9 Model Scale Analysis

To evaluate the impact of model scale on detection performance, we train and test HIM-CoT on Qwen2.5-Omni-3B (Xu et al., 2025), a smaller variant with fewer parameters. Table 23 presents the results. On HIM-DF, the 3B model reaches 85.4% accuracy compared with 89.3% for the 7B model, a drop of 3.9%, and still exceeds the accuracy of every baseline in Table 1. On LAV-DF, the 3B model reaches 89.8% accuracy, 0.8% below the 7B model, and remains second only to FaceCoT. The framework therefore remains effective with a smaller backbone, which offers flexibility for settings with tighter computational budgets.

Table 23: Model scale comparison on HIM-DF and LAV-DF benchmarks. We evaluate HIM-CoT with Qwen2.5-Omni-3B and Qwen2.5-Omni-7B backbones.

Model	HIM-DF			LAV-DF		
	ACC (%)	AP (%)	AUC (%)	ACC (%)	AP (%)	AUC (%)
Qwen2.5-Omni-3B	85.4	92.6	93.1	89.8	90.4	89.3
Qwen2.5-Omni-7B	89.3	95.0	94.2	90.6	91.2	90.9

C.10 Further Analyses

Comparison with General-Purpose MLLMs Figure 7 compares general-purpose MLLMs and HIM-CoT on an authentic trombone clip. Gemini predicts correctly but makes fine-grained claims that are not explicitly measured. MiniMax mistakes camera-relative motion for inward slide motion and therefore infers a contradiction with the downward pitch glide. The frames show a substantial

Table 24: Zero-shot cross-instrument generalization on HIM-DF. Models are trained on one verified source for each of ten instruments and evaluated on held-out trombone and tuba sources. The split is source-disjoint, so no source segment from either held-out instrument appears during training.

Method	Input	Train	Test	ACC (%)
WavLM	Audio	10	Trombone, tuba	62.8
AltFreezing	Video	10	Trombone, tuba	64.1
ICS-AV	Audio + video	10	Trombone, tuba	70.2
AVFF	Audio + video	10	Trombone, tuba	72.6
T-AVFD	Audio + video	10	Trombone, tuba	77.7
HIM-CoT w/o Structured Evidence	Audio + video	10	Trombone, tuba	68.4
Audio–visual baselines (mean)	Audio + video	10	Trombone, tuba	73.5
HIM-CoT	Audio + video + evidence	10	Trombone, tuba	**82.7**

change in camera–subject distance, making this interpretation unreliable. HIM-CoT instead combines close pYIN and CREPE estimates with plausible hand motion and finds no supported conflict, showing how measured evidence makes explanations easier to audit.

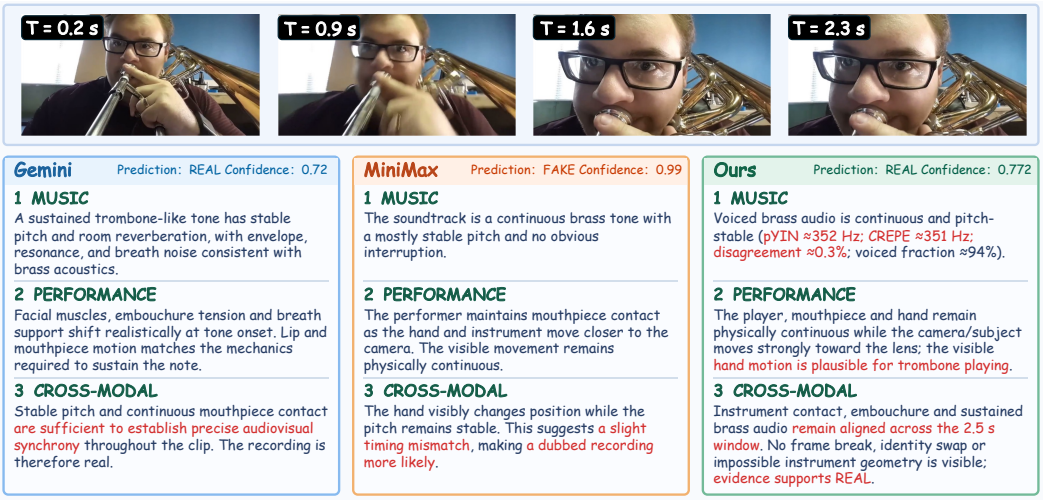

Figure 7: Comparison of CoT reasoning across multimodal language models on a real trombone performance.

Cross-Instrument Generalization Instrument identity affects both the visual motion pattern and the acoustic production process. A detector can therefore obtain a high score by memorizing instrument-specific appearance or timbre instead of learning the production constraints that connect motion, contact, and sound. Cross-instrument generalization tests whether the learned evidence remains useful when the instrument category is never observed during training. We evaluate zero-shot transfer by training on 10 instruments and testing on 2 held-out instruments. The held-out instruments are trombone and tuba, and all source segments from these instruments are reserved for testing. Because this experiment requires instrument labels verified at the source level, it uses a subset of sources whose instrument was manually checked from video frames: one source for each of the ten training instruments (piano, handpan, guitar, harp, oud, zither, violin, cello, trumpet, and oboe, which is grouped with clarinet under woodwinds in Table 10), and 10 trombone and 6 tuba sources for testing. One in five source segments of each training instrument is held out for validation. Table 24 reports the results.

HIM-CoT reaches 82.7% accuracy, 5.0% above the strongest baseline, T-AVFD, and 9.2% above the mean of the three audio–visual baselines. Without structured evidence, its accuracy drops by 14.3%, below every audio–visual baseline, indicating that the transferable signal comes mainly from the production-chain measurements. Because the test instruments are never observed during training, the comparison measures whether each method transfers production constraints across instrument families rather than exploiting instrument-specific appearance or timbre. Absolute accuracies are

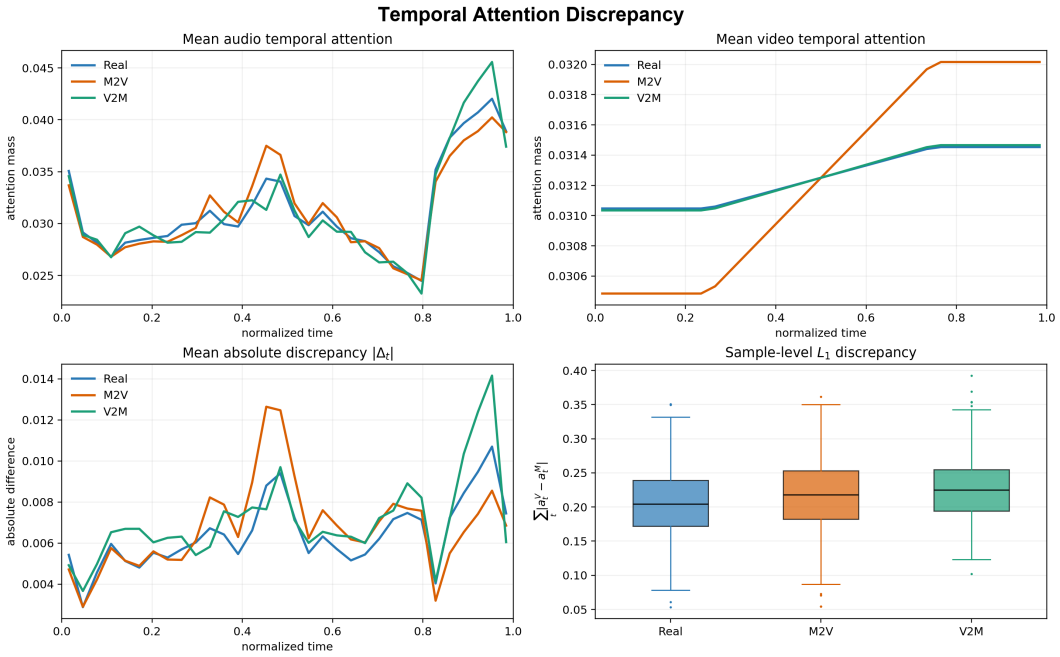

Figure 8: **Temporal Attention Discrepancy under independent synchronization control.** The top panels show mean audio and visual attention over normalized time, the lower-left panel shows their mean absolute difference, and the lower-right panel shows sample-level L_1 discrepancy. The forged groups, especially V2M, retain higher discrepancy despite the absence of a claim about large physical audio–video offsets.

not directly comparable with Table 1, because the training set is a small verified subset and the test set contains only brass instruments. Several baselines score higher here than in Table 1, so methods should be compared only within this table. With a single training source per instrument, this experiment is a small-scale probe of transfer rather than a definitive measurement, and a larger cross-instrument split is left to future versions of the benchmark.

Temporal Attention Discrepancy Modern audio–visual generators can produce media with small explicit temporal offsets, making physical synchronization alone an increasingly weak detection cue. The relevant question is therefore whether the two modalities support the decision using the same temporal evidence, even when the media appear synchronized. We summarize this internal evidence alignment by the L_1 norm of the temporal attention discrepancy, $\|\Delta_t\|_1 = \|a^V - a^M\|_1$ (Section 4.2), rather than treating it as a direct estimate of physical audio–video offset. Figure 8 visualizes Δ_t for real and forged samples, together with the independently measured synchronization features. In the current analysis, the mean $100 \times L_1$ discrepancy is 20.54 for Real, 21.66 for M2V, and 22.77 for V2M. The modest increase for forged samples, clearest for V2M, suggests that modality-specific attention can diverge even when the media remain approximately synchronized. To test this interpretation, we joined the attention measurements with an independent performance-based synchronization estimate using the same sample identifiers. Table 25 shows that the ordering persists after restricting the analysis to samples with small measured lag. With $|\mathrm{lag}| \leq 100$ ms, the means are 20.38, 20.87, and 22.62 for Real, M2V, and V2M, respectively. With $|\mathrm{lag}| \leq 200$ ms and synchronization correlation at least 0.2, the corresponding means are 20.23, 21.48, and 22.69. The attention discrepancy alone has only moderate discrimination (AUC 58.46–60.63% across these controls), so it is best interpreted as explanatory evidence rather than as a standalone detector. Strict source-window matching yields only 2, 7, and 13 matched source groups at 100, 200, and 300 ms, respectively. We therefore treat the matched analysis as a diagnostic with wide uncertainty, while the sample-level controls support the narrower claim that residual attention divergence persists when explicit synchronization quality is relatively good.

Table 25: Temporal Attention Discrepancy after independent synchronization control. Values are mean sample-level $100 \times L_1$ discrepancies. AUC measures whether the discrepancy ranks forged samples above Real samples.

Subset	Real n	M2V n	V2M n	Real	M2V	V2M	AUC (%)
All samples	266	234	200	20.54	21.66	22.77	59.00
$\|\text{lag}\| \leq 100$ ms	65	53	39	20.38	20.87	22.62	58.46
$\|\text{lag}\| \leq 200$ ms, $r \geq 0.2$	87	74	56	20.23	21.48	22.69	60.63

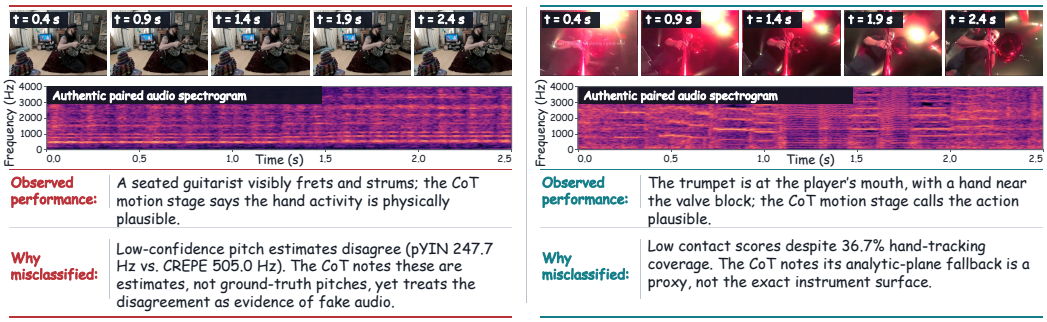

Figure 9: **Representative HIM-CoT failure cases.** *Left:* a guitar case in which low-confidence pitch estimates disagree, causing the model to treat estimator disagreement as evidence of fake audio even though the observed strumming is physically plausible. *Right:* a trumpet case in which low contact-estimation scores are driven by limited hand-tracking coverage and an analytic-plane fallback, while the observed performance remains plausible.

Error Analysis: Failure Modes We manually categorize all 75 misclassified test samples into three failure modes. *Measurement uncertainty* occurs when structured features fail to capture critical forensic cues because of occlusion, motion blur, or recording noise. *Evidence conflicts* occur when audio and video evidence provide contradictory signals that the discrepancy modeling cannot resolve. *Subtle forgeries* are clips in which both modalities are individually plausible and cross-modal inconsistencies are too weak to detect. Measurement uncertainty accounts for 32 errors (43%), evidence conflicts for 23 (31%), and subtle forgeries for 20 (27%). Figure 9 shows two representative cases of the most frequent mode, measurement uncertainty. Both are authentic clips that are misclassified because unreliable measurements are interpreted as forensic evidence. Note that such measurements are still recorded as observed, since the extractor returns a low-confidence value or a fallback estimate rather than no value (Appendix A.8). Uncertainty of this kind is therefore not captured by the observation masks. Failure cases cluster in scenarios with extreme lighting, rapid hand movements, or highly polished synthetic content that minimizes cross-modal attention divergence. This analysis suggests that improving structured feature extraction under challenging recording conditions and incorporating higher-order temporal reasoning could further enhance performance.

C.11 Robustness Evaluation

We evaluate HIM-CoT's robustness across 33 conditions: one clean baseline and 32 perturbations covering audio degradation, video degradation, and temporal manipulation. Each perturbation condition contains 700 samples. The 26 label-preserving perturbations test robustness to common degradations without altering the ground-truth authenticity labels. Six additional sensitivity conditions involve pitch shifts or audio–video temporal offsets that may invalidate the original labels, serving as diagnostics rather than fixed-label robustness tests. We report accuracy (Acc), AUC, and average precision (AP) as percentages, with ΔAcc showing the difference relative to the clean baseline of 89.3%. Results are organized by perturbation category in Tables 26 through 28.

Audio perturbations (Table 26) HIM-CoT is robust to moderate audio degradations. Small gain changes have little impact (-0.5% and -2.4% at ± 3 dB), but attenuation by 6 dB reduces accuracy

Table 26: Robustness to audio perturbations on HIM-DF. All perturbations preserve authenticity labels. Each condition has 700 samples.

Condition	Acc	AUC	AP	ΔAcc
HIM-CoT	89.3	94.2	95.0	0.0
Gain +3 dB	88.8	93.9	94.7	−0.5
Gain +6 dB	87.4	93.7	94.6	−1.9
Gain −3 dB	86.9	92.7	94.1	−2.4
Gain −6 dB	84.4	91.7	93.6	−4.9
Pink noise, SNR 10 dB	82.7	90.0	92.8	−6.6
Pink noise, SNR 20 dB	86.2	92.6	94.1	−3.1
Pink noise, SNR 30 dB	87.8	93.5	94.6	−1.5
Reverb, RT60 0.2 s	85.8	92.6	94.0	−3.5
Reverb, RT60 0.5 s	85.2	90.9	93.0	−4.1
Reverb, RT60 0.8 s	83.1	90.6	92.9	−6.2

Table 27: Robustness to video perturbations on HIM-DF. All perturbations preserve authenticity labels. Each condition has 700 samples. ΔAcc is relative to Clean accuracy of 89.3%.

Condition	Acc	AUC	AP	ΔAcc
HIM-CoT	89.3	94.2	95.0	0.0
Blur $\sigma = 1$	86.9	93.4	94.4	−2.4
Blur $\sigma = 2$	87.2	93.3	94.4	−2.1
Blur $\sigma = 3$	87.5	93.6	94.5	−1.8
Brightness $\times 0.5$	86.7	93.3	94.4	−2.6
Brightness $\times 0.7$	86.7	93.2	94.4	−2.6
Brightness $\times 1.3$	87.4	93.4	94.4	−1.9
Brightness $\times 1.5$	87.7	93.3	94.4	−1.6
H.264 CRF 23	86.9	93.1	94.3	−2.4
H.264 CRF 30	86.9	93.5	94.5	−2.4
H.264 CRF 37	87.5	93.6	94.5	−1.8
Video noise $\sigma = 0.01$	87.5	93.2	94.4	−1.8
Video noise $\sigma = 0.03$	87.1	93.0	94.3	−2.2
Video noise $\sigma = 0.05$	86.2	93.4	94.4	−3.1
Resolution 240p	86.8	92.2	93.8	−2.5
Resolution 360p	87.5	93.4	94.4	−1.8
Resolution 480p	87.5	93.2	94.4	−1.8

to 84.4%, so the model is not fully invariant to input level. Pink noise shows graceful degradation: accuracy improves from 82.7% at 10 dB SNR to 87.8% at 30 dB SNR, demonstrating that structured acoustic features maintain discrimination capability under additive noise while severe noise (10 dB SNR) causes moderate performance drops of 6.6%. Reverberation has modest effects, with accuracy declining from 85.8% (RT60 0.2s) to 83.1% (RT60 0.8s) as room reflections obscure temporal cues, though performance remains substantially above chance even in highly reverberant conditions.

Video perturbations (Table 27) Video degradations have remarkably small effects. Across all 16 video-only conditions, accuracy remains between 86.2% and 87.7%, with ΔAcc within -1.6% to -3.1% relative to baseline. Blur, brightness changes, H.264 compression, and video noise all yield minimal performance variations. The drop is also nearly independent of severity: re-encoding at CRF 23, the setting used to build the benchmark, costs 2.4%, while CRF 37 costs 1.8%. This suggests that the small drop stems mainly from re-processing the video rather than from the degradation itself. Resolution downsampling to 240p, 360p, or 480p maintains 86.8–87.5% accuracy, indicating that motion kinematics and contact patterns remain discriminable at reduced spatial detail. This suggests that the visual evidence groups rely more on kinematic and contact features than on pixel-level detail, although the non-monotonic trends caution against reading them as robustness guarantees.

Table 28: Sensitivity to temporal manipulations on HIM-DF. These conditions may invalidate the original authenticity labels, so results are sensitivity diagnostics rather than fixed-label robustness tests. ΔAcc is relative to Clean accuracy of 89.3%.

Condition	Acc	AUC	AP	ΔAcc
HIM-CoT	89.3	94.2	95.0	0.0
Pitch shift $+0.5$ semitone	70.7	83.8	91.0	-18.6
Pitch shift -0.5 semitone	67.7	77.1	87.2	-21.6
Pitch shift -1 semitone	67.4	78.0	87.9	-21.9
Audio–video offset $+100$ ms	87.8	93.0	94.2	-1.5
Audio–video offset $+200$ ms	86.4	92.6	94.0	-2.9
Audio–video offset $+300$ ms	86.2	92.2	93.7	-3.1

Sensitivity conditions (Table 28) Pitch shifts cause large accuracy drops (67.4–70.7%). Two factors may contribute. A shift can create a genuine mismatch between the heard pitch and the visible fingering, and pitch-shifting also introduces signal-processing artifacts that resemble synthesized audio. Because a half-semitone shift is small relative to the resolution of visible fingering, we do not attribute the drop to pitch–motion reasoning alone. Distinguishing benign audio processing from forgery remains an open challenge. Audio–video temporal offsets of 100–300 ms show modest impact (86.2–87.8% accuracy), with performance declining as offset increases. These imposed offsets test sensitivity to explicit synchronization perturbations and are distinct from the internal attention discrepancy Δ_t analyzed above.

Overall robustness analysis The robustness evaluation reveals consistent vulnerability patterns across modality-specific and cross-modal cues. Video-only perturbations have modest but consistent impact (accuracy remains 86.2–87.7%), with all conditions showing 1.6–3.1% drops, confirming that kinematic and contact features maintain discrimination despite visual degradations. Audio perturbations show similar moderate effects: gain adjustments cause 0.5–4.9% drops, while additive noise demonstrates graceful degradation correlated with SNR (1.5–6.6% drops). Reverberation causes 3.5–6.2% drops as room reflections obscure temporal cues. Audio perturbations cause larger and more severity-dependent drops (up to 6.6%) than video perturbations (at most 3.1%), indicating that the model is more sensitive to degradation of the acoustic evidence than of the visual evidence. Sensitivity conditions show that pitch shifts and imposed temporal offsets affect the prediction, with pitch shifts creating stronger changes and 100–300 ms offsets causing modest but measurable degradation (1.5–3.1%). These results characterize responses to explicit perturbations and do not establish that the clean forgeries contain comparable physical synchronization errors. Overall, HIM-CoT maintains accuracy of at least 82.7% across all label-preserving perturbations, demonstrating robust deepfake detection under realistic degradation scenarios.

C.12 Limitations

While HIM-CoT advances instrumental performance deepfake detection, several limitations remain. First, the method relies on structured feature extraction, which can fail under severe occlusion or low-quality recordings. Some measurements, such as unstable hand tracking or disagreement between pitch estimators, also respond to synthesis artifacts rather than only to the production chain, as the failure cases in Appendix C.10 show, so the gains from structured evidence cannot be attributed to production-chain consistency alone. Second, the current implementation focuses on solo performances, and extending to ensemble settings with multiple performers and instruments requires further investigation. Third, the observation masks capture missing measurements but not low-confidence ones, which account for the largest share of errors (Appendix C.10). Feeding extractor confidence, such as pitch, pose, and contact-estimation confidence, into the evidence gating is a natural extension, and the model's reasoning occasionally hallucinates when measurements are unreliable. Fourth, the reasoning is supervised with label-conditioned rationales, so it should be read as an evidence-grounded explanation aligned with the decision. Because the annotator knows the label, it may also select and emphasize observations that support it, a bias that removing verdict words (Appendix C.3) does not eliminate. Training on rationales written without access to the label, and interventions on the generated reasoning, are needed to establish that the reasoning causally drives the prediction. Fifth, the main HIM-DF split uses one generator per forgery direction, and the leave-one-

generator-out evaluation covers three generators per direction (Appendix C.8). Wider generator coverage is still needed to fully separate production-chain cues from artifacts shared across generators. Sixth, HIM-CoT shows measurable sensitivity to within-recording temporal misalignment, but most temporally shifted pairs are still not detected (Appendix C.7), so fine-grained audio–performance synchronization remains an open challenge. Finally, HIM-DF labels encode provenance. A synthesized clip whose performance plausibly explains the sound is still fake, so the benchmark measures detection of synthesized modalities, for which production-chain consistency is one source of evidence, rather than physical plausibility itself.