

TriDF: Evaluating Perception, Detection, and Hallucination for Interpretable DeepFake Detection

Supplementary Material

Overview

Table of Contents in Supplementary Material

A DeepFake Tasks in TriDF	1
A.1 Partially Manipulated Tasks	1
A.2 Fully Synthesized Tasks	1
B DeepFake Data Generation	2
C Taxonomy of DeepFake Artifacts	4
D Annotation Platform	6
E Distribution of Ground Truth Options	6
F Benchmark Statistics	6
G Templates	8
G.1 Templates for Benchmark Construction . . .	8
G.2 Templates for Artifacts Mapping	8
H Audio Modality Analysis	8
I. Extended Evaluation	9
I.1 . Evaluation Setup	9
I.2 . More Quantitative Results	9
I.3 . More Qualitative Results	10
J. Future Direction of DeepFake Detection	11
K Release Plan and Ethics Statement	11
Bibliography	14

A. DeepFake Tasks in TriDF

DeepFake technologies and synthetic media systems encompass a broad spectrum of manipulation techniques, each targeting distinct aspects of human-centric visual and auditory content. To systematically evaluate this landscape, TriDF organizes these techniques into two functional categories: Partially Manipulated, which encompasses methods that alter specific attributes of an existing subject within a scene, and Fully Synthesized, which covers approaches that generate entirely artificial human appearances or voices. Representative qualitative examples for each category are illustrated in Figs. S1 and S2, respectively. In what follows, we formally define each category included in TriDF and characterize its distinguishing properties, clarifying how

each contributes to the benchmark’s comprehensive coverage of the DeepFake detection problem.

A.1. Partially Manipulated Tasks

Image/Video Face Swapping transfers the identity of a source subject onto a target face while preserving the target’s original scene-consistent attributes, including pose, illumination, and expression.

Facial Attribute Manipulation selectively modifies specific semantic facial attributes, such as age, expression, hair color, or accessories, in a directed and controlled manner, while preserving the subject’s core identity.

Lip Synchronization alters the lip movements of a subject in a video to match a new or substituted audio track, producing the perceptual illusion that the subject is articulating words they did not originally utter.

Face Reenactment transfers the facial expressions, head pose, and eye gaze of a source subject onto a target subject, effectively compelling the target to replicate the source’s performance across a static image or an independent video sequence.

Full-Body Puppetry extends the face reenactment paradigm to the full human body, transferring the complete skeletal pose and motion dynamics of a source actor onto a target subject, thereby enabling the source to drive the target’s movements throughout a video.

Subject-Driven Image/Video Editing applies targeted manipulations to a specific subject within an image or video, typically guided by textual prompts or reference images (e.g., “change the person’s shirt to red”), while preserving both the subject’s identity and the surrounding scene context.

Voice Conversion transforms a speaker’s vocal characteristics to resemble those of a designated target speaker, while strictly preserving the original linguistic content and spoken words.

A.2. Fully Synthesized Tasks

Audio-Driven Talking Head Synthesis generates a fully synthetic video of a human subject in which lip movements, facial expressions, and head pose are produced entirely from scratch and conditioned on an input audio signal, without relying on any real video footage of the subject.

Identity-Preserving Image/Video Generation synthesizes novel images or videos of a specific individual by learning their identity representation from a limited set of reference

photographs, enabling generation of that individual in previously unseen poses, environments, or visual styles.

Text-to-Human Image/Video Generation involves the synthesis of high-fidelity human images or video sequences conditioned exclusively on textual descriptions. Given a text prompt, generative models map semantic concepts to visually coherent representations without the aid of external visual priors.

Human Image-to-Video Generation focuses on animating a static reference image into a continuous video sequence, guided by a textual prompt. The objective is to preserve the identity and fine-grained attributes of the source subject while synthesizing realistic motion and temporal dynamics that align with the provided textual instructions.

Voice Cloning constructs a comprehensive generative model of a specific individual’s voice, often from a minimal audio sample, capturing their unique tonal quality, cadence, and vocal style. The resulting model enables arbitrary speech synthesis in the target speaker’s voice via text-to-speech generation.

B. DeepFake Data Generation

Data Acquisition. We exclusively collect information in accordance with the specific licensing agreements of source websites, avoiding material that is protected against usage for any commercial purposes. The licenses of the existing datasets used in this work are as follows:

- FaceForensics++ [R70]: Non-commercial research and educational purposes.
- FFHQ [R43]: Creative Commons BY-NC-SA 4.0
- CelebAMaskHQ [R45]: Non-commercial research and educational purposes.
- CelebA-HQ [R42]: Non-commercial research and educational purposes
- VGGFace2 [R9]: Unspecified
- Emu Edit [R72]: Creative Commons BY-NC 4.0
- GEdit-Bench [R57]: MIT License
- ImgEdit [R98]: Apache license 2.0
- OmniContext [R89]: Apache License 2.0
- MS-COCO [R53]: Creative Commons BY 4.0
- Flickr30k [R67]: Non-commercial research and educational purposes.
- LAION-Aesthetics [R71]: Creative Commons BY 4.0
- VoxCeleb2 [R16]: Creative Commons BY-SA 4.0
- LRS2 [R76]: Academic Research Purposes.
- TalkingHead-1KH [R82]: Creative Commons BY 3.0
- VPBench [R7]: The CogVideoX License
- FiVE-Bench [R51]: Creative Commons BY-NC 4.0
- HDTF [R109]: Creative Commons BY 4.0
- CelebV-Text [R99]: Non-commercial research purposes only.
- Fashion Video [R103]: Creative Commons BY-NC 4.0
- TED-talks [R75]: Unspecified

- TikTok [R38]: Creative Commons BY-NC 4.0
- A2 Bench [R25]: Apache License 2.0
- OpenS2V-Nexus [R100]: Apache License 2.0
- ConsisID [R101]: Creative Commons BY 4.0
- Panda-70M [R12]: Non-commercial and research purposes.
- HOIGen-1M [R54]: Apache License 2.0
- EMIME [R88]: Open Data Commons Attribution License (ODC-By) v1.0
- VCTK [R93]: Creative Commons BY 4.0
- LibriTTS [R104]: Creative Commons BY 4.0
- LibriSpeech [R65]: Creative Commons BY 4.0

All datasets released with this work are available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). We selected this license to match the terms of several original datasets and to provide our data under the same access conditions.

Data Generation. To ensure comprehensive coverage, we organize our synthesis pipeline into task-oriented sub-domains, as detailed in Tab. S1.

In the image modality, we move beyond traditional Face Swapping to include Subject-driven Editing and Identity-Preserving Generation, utilizing both open-source models, such as PixArt- σ [R11], OmniGen2 [R89], Step1X-Edit [R57], SD3 [R24], and Flux 1 [R6], and proprietary generators like Gemini 2.5 [R28] and GPT-4o [R64].

The video modality represents the most diverse category, addressing the spectrum from facial to full-body synthesis. We include head-centric tasks, such as Face Reenactment and Lip-Syncing (*e.g.*, MuseTalk [R108]), alongside complex body-centric tasks like Full-Body Puppetry via Champ [R114] and ControlNeXt [R66]. Furthermore, we incorporate Human Video Generation utilizing models like LTX-Video [R31], Wan2.2 [R80], Phantom [R55], and HunyuanCustom [R35], covering various conditioning inputs such as reference images and pure text.

Finally, for the audio modality, we target both Voice Cloning and Voice Conversion. By gathering open-source solutions like OpenVoice [R69] and Seed-VC [R56] against commercial APIs like ElevenLabs [R23], we capture the current state-of-the-art across varying acoustic environments.

Quality Control. To increase the high fidelity of our generated DeepFakes, we employ specialized metrics for assessing realism and consistency to ensure automatic quality control before starting the annotation process. *Realism metrics*, namely LPIPS [R105], NIQE [R61], VSFA [R48], and NISQA [R60], evaluate whether the content appears natural and is challenging for humans or algorithms to detect as synthetic. In contrast, *consistency metrics*, including ArcFace [R20], CLIPScore [R33], LSE-C [R68], AED&AKD [R74], SECS [R56], and ViCLIP [R85], mea-

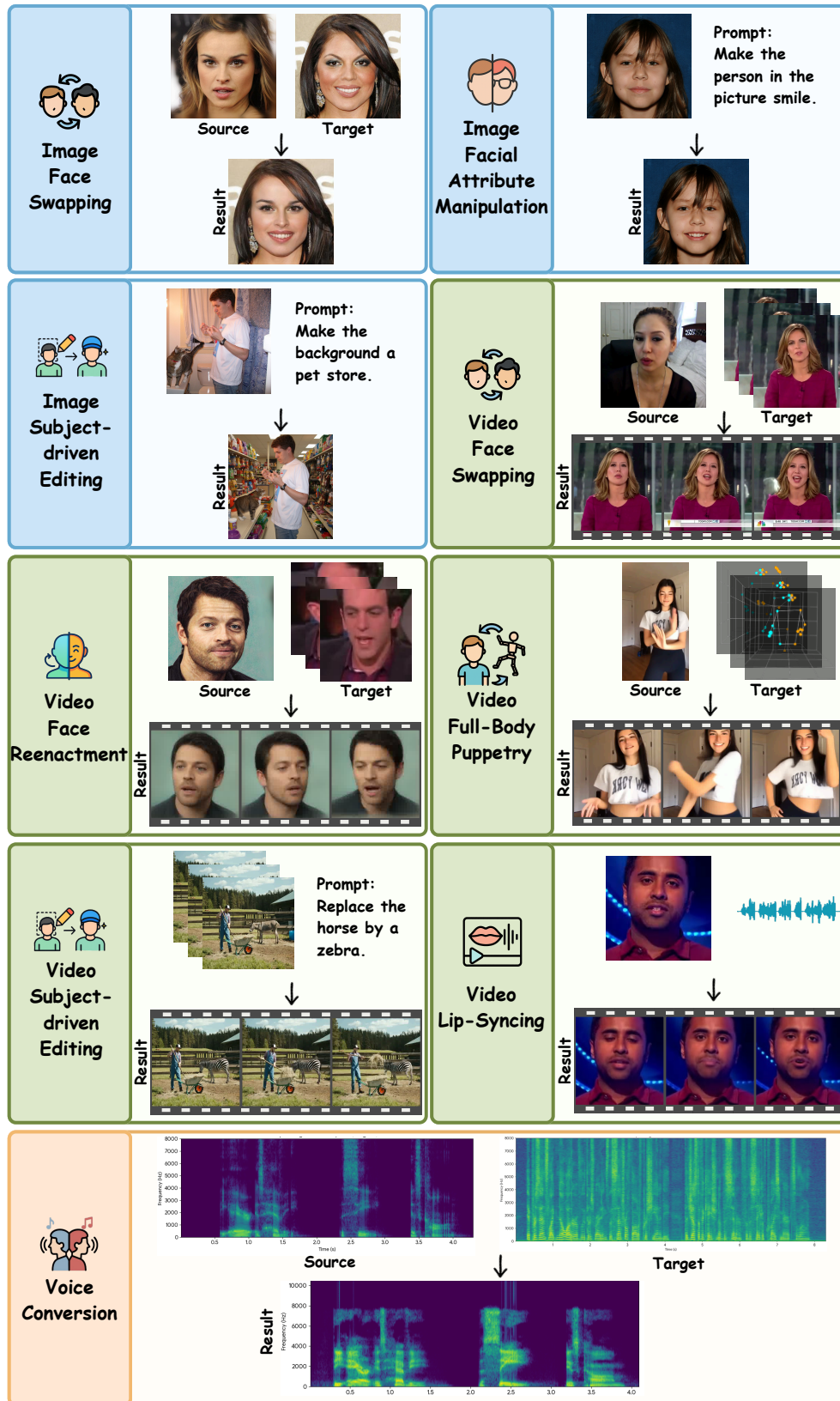


Figure S1. Examples of DeepFakes from Partially Manipulated tasks.

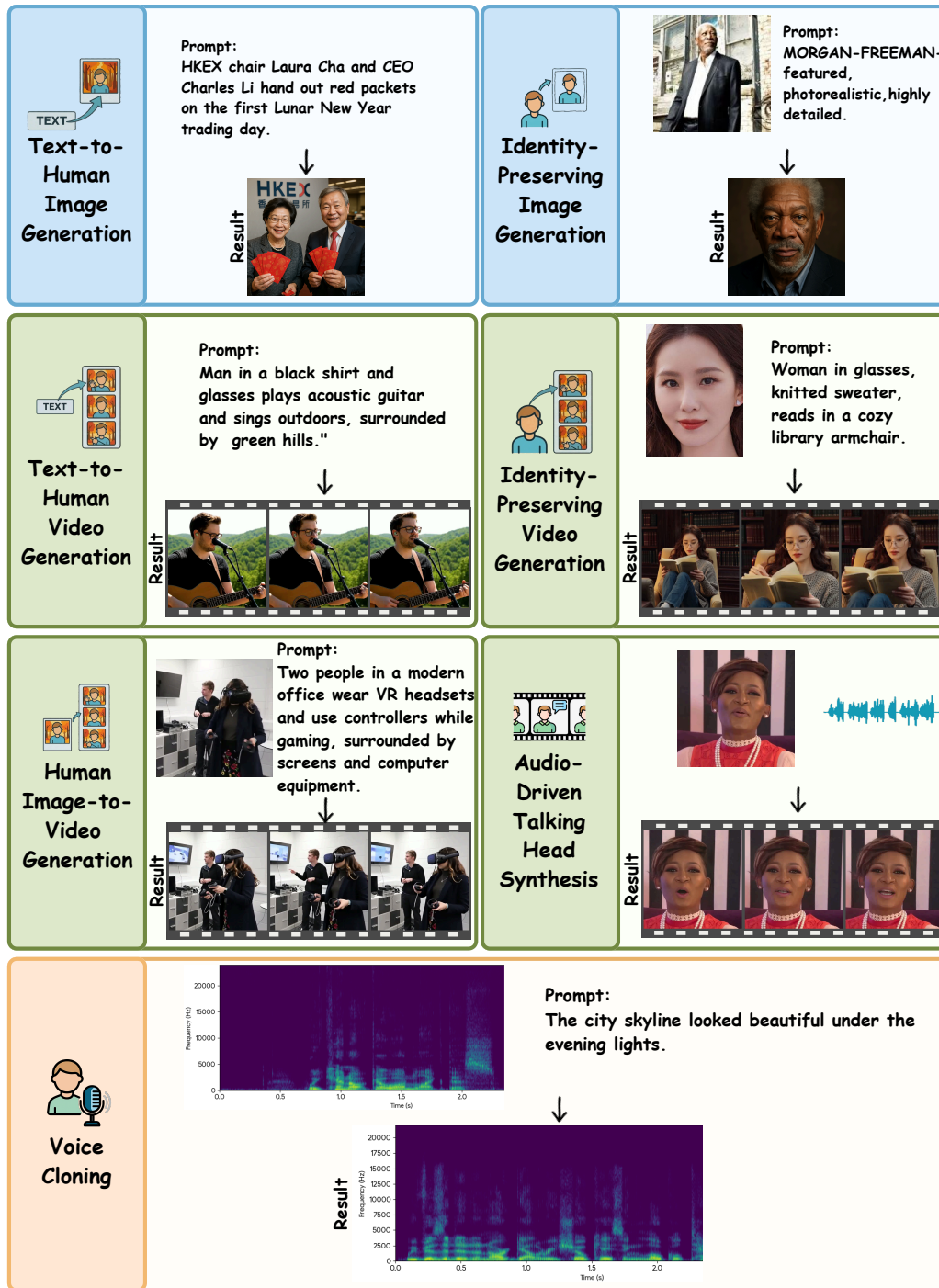


Figure S2. Examples of DeepFakes from Fully Synthesized tasks.

sure how closely the output aligns with input conditions or control signals, such as retaining facial identity, voice characteristics, or movement synchronization. After applying quality control, we form one-to-one real-fake pairs in each DeepFake task, resulting in a total of over 5K high-quality pairs spanning three different modalities.

C. Taxonomy of DeepFake Artifacts

To systematically categorize the artifacts present in DeepFake media, we divide the artifacts into two distinct classes based on the level of analysis required for detection. Tab. S2 outlines *Quality Artifacts*, which encompass low-level sig-

Table S1. Overview of DeepFake tasks, representative synthesis methods, and commonly used public datasets across three modalities. For each task, we select three publicly available code repositories to ensure diversity in generation approaches. To maintain fair evaluation and simulate real-world scenarios, only the testing splits of public datasets or datasets not used for training are employed for generation.

Modality	Tasks	Synthesis Methods	Public Dataset
Image	Face Swapping	DiffSwap [R111]	FaceForensics++ [R70]
		BlendFace [R73]	FFHQ [R43]
		CSCS [R37]	CelebAMaskHQ [R45]
	Facial Attribute Manipulation	PREIM3D [R50]	CelebA-HQ [R42]
		AdaTrans [R36]	VGGFace2 [R9]
		StyleGANEX [R94]	FFHQ [R43]
	Subject-driven Image Editing	Mige [R78]	Emu Edit [R72]
		Step1X-Edit [R57]	GEdit-Bench [R57]
		OmniGen2 [R89]	ImgEdit [R98]
		Gemini 2.5 Flash Image [R28]	
	Identity-Preserving Generation	Mige [R78]	CelebA-HQ [R42]
		UNO [R90]	FFHQ [R43]
		OmniGen2 [R89]	OmniContext [R89]
		Gemini 2.5 Flash Image [R28]	
	Human Scene Generation	SD3 [R24]	MS-COCO [R53]
		PixArt- σ [R11]	Flickr30k [R67]
Flux 1. [R6]		LAION-Aesthetics [R71]	
GPT-4o Image [R64]			
Video	Face Swapping	HifiFace [R84]	CelebA-HQ [R42]
		InfoSwap [R26]	VoxCeleb2 [R16]
		FaceAdapter [R32]	FaceForensics++ [R70]
	Face Reenactment	MCNet [R34]	CelebA-HQ [R42]
		HyperReenact [R8]	VoxCeleb2 [R16]
		LivePortrait [R30]	FaceForensics++ [R70]
	Lip-Syncing	DINet [R110]	LRS2 [R76]
		LatentSync [R47]	VoxCeleb2 [R16]
		MuseTalk [R108]	TalkingHead-1KH [R82]
	Subject-driven Video Editing	VideoPainter [R7]	VPBench [R7]
		VACE [R40]	FiVE-Bench [R51]
		Wan-Edit [R51]	
	Audio-driven Talking-Head Synthesis	SadTalker [R106]	TalkingHead-1KH [R82]
		AniPortrait [R86]	HDTF [R109]
		Hallo2 [R18]	CelebV-Text [R99]
		D-ID [R19]	
	Full-Body Puppetry	Champ [R114]	Fashion Video [R103]
		MotionEditor [R79]	TED-talks [R75]
		MagicDance [R10]	TikTok [R38]
		ControlNeXt [R66]	
	Identity-Preserving Generation	Hunyancustom [R35]	A2 Bench [R25]
		VACE [R40]	OpenS2V-Nexus [R100]
		Phantom [R55]	ConsisID [R101]
		Kling [R21]	
Human Image-to-Video Generation	LTX-Video [R31]	CelebV-Text [R99]	
	CogVideoX [R95]	Panda-70M [R12]	
	Wan2.2 [R80]	HOIGen-1M [R54]	
	Veo3 [R29]		
Human Scene Generation	LTX-Video [R31]	CelebV-Text [R99]	
	Pyramid-Flow [R41]	Panda-70M [R12]	
	SkyReels-A2 [R25]	HOIGen-1M [R54]	
	Veo3 [R29]		
Audio	Voice Cloning	XTTS [R2]	EMIME [R88]
		OpenVoice [R69]	VCTK [R93]
		CosyVoice 2.0 [R22]	LibriTTS [R104]
		ElevenLabs [R23]	
	Voice Conversion	SpeechT5_VC [R4]	LibriSpeech [R65]
		Seed-VC [R56]	VCTK [R93]
		Diff-HierVC [R14]	LibriTTS [R104]

nal distortions and compression errors that are often detectable through traditional image or audio processing techniques. In contrast, Tab. S3 details *Semantic Artifacts*, which represent high-level logical inconsistencies, e.g., violations of physics or anatomy, that require contextual understanding to identify.

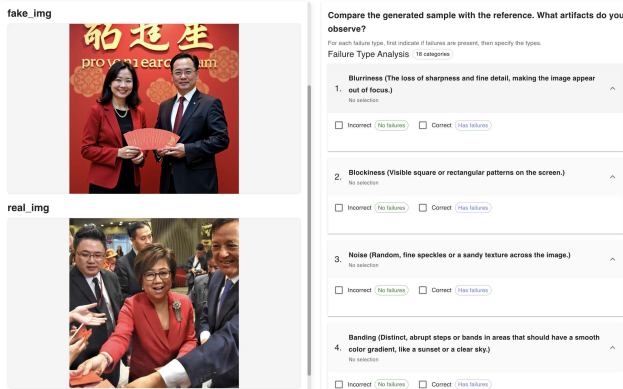


Figure S3. Graphic User Interface of Annotation Platform. It displays paired real and DeepFake samples stacked vertically to facilitate fine-grained comparison and structured artifact labeling for reliable annotation results.

D. Annotation Platform

To implement the unified taxonomy at scale, we have developed a dedicated annotation platform optimized for hierarchical annotation. The annotation process is fully manual, prioritizing accuracy and reliability over automation. In light of the 59% accuracy ceiling observed with GPT-4o [R62] on DeepFake detection, reported by LOKI [R97], we have intentionally excluded AI-assisted pre-annotation. We recruited more than 50 annotators. Each generated DeepFake sample is assigned to at least three annotators, and consensus is reached through majority voting. A key feature of our platform, illustrated in Fig. S3, is the top-down layout for comparing real and fake media pairs, each matched in a strict one-to-one correspondence. This layout enables annotators to systematically compare manipulated samples with their authentic counterparts, facilitating the precise identification of both *Quality* and *Semantic Artifacts*. To accelerate the annotation process and alleviate the burden of typing complete sentences to describe artifacts found in the generated DeepFake samples, we designed an interface that supports a structured checklist in a multiple-choice style, allowing annotators to assign taxonomy-based labels at multiple levels of granularity with ease and efficiency.

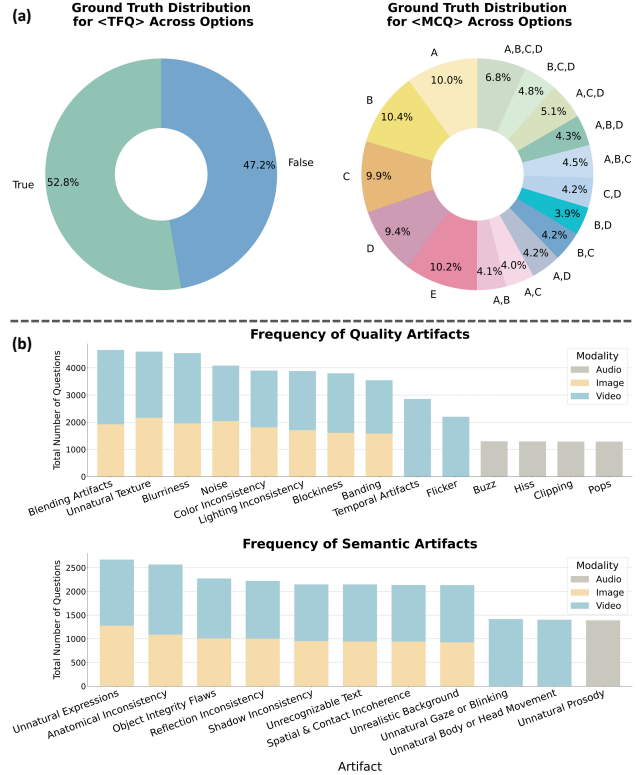


Figure S4. **Statistics of TriDF.** (a) The distribution of ground truth options for $\langle TFQ \rangle$ and $\langle MCQ \rangle$. (b) The frequency of quality artifacts and semantic artifacts.

E. Distribution of Ground Truth Options

As illustrated in Fig. S4, we adopt the approach from [R58, R112] to ensure that the ground truth options, e.g., true-false or multiple-choice options, are distributed as evenly as possible. This step helps alleviate the well-known “selection bias” issues in MLLMs [R59, R112], where they often favor specific option labels as answers.

F. Benchmark Statistics

Comparison with Existing Benchmarks. As shown in Tab. 1 in the main paper, we compare our proposed TriDF with existing benchmarks [R52, R81, R97, R107, R113] for DeepFake detection across several key dimensions, including the size of testing sets, the number of generators, the types of DeepFakes, the data modalities, and the evaluation metrics. Notably, TriDF distinguishes itself with the largest number of questions (65K), generators (51), and DeepFake types (16), spanning three modalities, image, video, and audio, surpassing prior works that often focus on limited generators or types of DeepFake. This extensive collection of generators is a key advantage, providing a far more rigorous test of a detector’s robust-

Table S2. Quality Artifacts: Localized signal errors detectable by traditional processing methods.

Domain	Artifact	Definition
Visual Signal	Blurriness	The loss of sharpness and fine detail, making the image appear out of focus.
	Blockiness	Visible square or rectangular patterns on the screen.
	Noise	Random, fine speckles or a sandy texture across the image.
	Banding	Distinct, abrupt steps or bands in areas that should have a smooth color gradient.
	Color Inconsistency	Colors appear unnatural, with excessive saturation or vibrancy.
	Blending Artifacts	Visible boundaries where elements should merge smoothly.
	Lighting Inconsistency	Illumination that does not agree across the scene.
	Unnatural Texture	The surface is overly smooth, missing natural irregularities.
Temporal	Temporal Artifacts	Inconsistencies across frames that break motion continuity.
	Flicker	Noticeable and often rapid variation in the overall brightness.
Audio Signal	Clipping	Harsh, fuzzy, or crackling sound when audio is too loud.
	Hiss	High-frequency static noise (e.g., “shhhh” sound).
	Buzz	Low-frequency tone, typically caused by electrical interference.
	Pops	Abrupt, short, and sharp sounds that interrupt the audio.

Table S3. Semantic Artifacts: High-level inconsistencies requiring contextual understanding. (Env. = Environment; Lang. = Language)

Context	Artifact	Definition
Physics & Env.	Reflection Inconsistency	Reflections do not match the subject, lighting, or scene geometry.
	Shadow Inconsistency	Shadows do not match the subject, lighting, or scene geometry.
	Spatial Incoherence	Objects or people fail to make contact with surfaces or each other.
	Unrealistic Background	Background lacks plausible detail, perspective, or depth.
Human Biology	Anatomical Inconsistency	Human anatomy is implausible (e.g., distorted limbs).
	Unnatural Expressions	Facial expressions do not align with emotion or context.
	Unnatural Gaze	Eye direction or blink behavior appears robotic.
	Unnatural Movement	Motion lacks physical plausibility.
Objects & Lang.	Object Integrity Flaws	The object is incomplete, broken, or internally inconsistent.
	Unrecognizable Text	Text is unrecognizable, incomplete, broken, or distorted.
	Unnatural Prosody	Speech sounds robotic, monotonous, or flat.

ness and generalization capabilities. It ensures that models are evaluated against a diverse spectrum of generation artifacts, rather than overfitting to the signatures of a few common tools. Crucially, this diversity enables TriDF to simulate real-world “in-the-wild” scenarios by assessing performance against the latest generation models, including state-of-the-art methods such as PixArt- σ [R11], OmniGen2 [R89], Step1X-Edit [R57], Flux 1. [R6], SD3 [R24], Gemini 2.5 Flash Image [R28], GPT-4o Image [R64], Hunyancustom [R35], LTX-Video [R31], Wan2.2 [R80], and Veo3 [R29]. Unlike existing benchmarks, TriDF features a comprehensive suite of metrics to quantify the interpretability of DeepFake detection, including Accuracy and *Cover*

metrics. It also evaluates the perception abilities and hallucination tendencies of MLLMs through strict real-fake pairs, which enable side-by-side comparisons and allow annotators to assign taxonomy-based labels at multiple levels of granularity. This approach provides a more nuanced and robust assessment of model performance in real-world DeepFake scenarios. In designing TriDF, we deliberately avoid using LLM-as-a-judge approaches. As discussed in [R49], employing LLMs as judges inherently introduces biases that can compromise the fairness and reliability of evaluations. Furthermore, LLM judges are susceptible to adversarial attacks, such as prompt injection, thereby raising significant concerns about their reliability in high-stakes

scenarios, including DeepFake detection.

Statistics. TriDF is a meticulously curated benchmark designed to comprehensively evaluate DeepFake detection. It consists of 65K questions that span 16 DeepFake techniques, including modern methods like GANs, SD, and DiT. The benchmark’s scope is intentionally broad, covering 3 distinct modalities (image, video, and audio) and multiple types of forgeries, from partially manipulated content to fully synthetic media. To ensure a thorough evaluation of interpretability in DeepFake detection, perception abilities, and hallucination tendencies in MLLMs, the questions are distributed across 23K <TFQ>, 24K <MCQ>, and 18K <OEQ>. This significant diversity challenges MLLMs, requiring them to demonstrate robust generalization and a more comprehensive capacity for identifying different forms of DeepFakes.

G. Templates

G.1. Templates for Benchmark Construction

Fig. S5 outlines prompt templates designed for benchmark construction across three distinct question formats: <TFQ>, <MCQ>, and <OEQ>. The <TFQ> (True-False Question) section provides templates to verify the observation of specific artifacts, their presence in the background, or their existence in specific locations. The <MCQ> (Multiple-Choice Question) templates ask MLLMs to identify present artifacts or their locations from a list, including instructions to select all that apply or indicate if no options are correct. Finally, the <OEQ> (Open-Ended Question) templates, split into Type A and Type B, establish a persona for a DeepFake forensics analyst, detailing strict guidelines for performing thorough artifact analysis, avoiding false positives, and adhering to a specific output format.

G.2. Templates for Artifacts Mapping

Fig. S6 serves as a structured guide for identifying particular visual flaws in media analysis texts. It offers precise definitions of various artifacts as a reference point, compelling LLMs to assess their occurrence based on these exact standards. The template requires LLMs to deliver straightforward binary judgments of “True” or “False,” formatted in a machine-readable style using only key-value pairs.

H. Audio Modality Analysis

Evaluation of Perception. Tab. S4 presents the audio perception performance of five open-weight Audio-MLLMs and one proprietary multimodal model. Two distinct trends emerge from the results.

Firstly, semantic perception is substantially more challenging than quality perception. On <TFQ>, Gemini-2.5-Pro attains the highest semantic accuracy, yet most audio-specialized models perform near random chance in this

Template for <TFQ>

modality: Is {artifact} observed in the sample? (Answer yes if you think it is true, and answer no otherwise.)

modality: Is the artifact {artifact} present in the background in the sample? (Answer yes if you think it is true, and answer no otherwise.)

modality: Is any visual artifact present in the {location} in the sample? (Answer yes if you think it is true, and answer no otherwise.)

Template for Type A <OEQ>

This is a DeepFake sample generated by an advanced DeepFake technique.

Your Task: Please perform a detailed analysis of artifacts in this sample that seem inauthentic.

Guidelines:

- Be Thorough: Offer a comprehensive analysis covering all noticeable artifacts.
- Be Accurate: Ensure your explanations correctly reflect the information present in the sample.
- Avoid False Positives: Do not incorrectly identify authentic features as inauthentic. Focus only on actual artifacts.
- Organize Your Response: Use clear headings and structure to separate your analysis for each artifact.

Template for Type B <OEQ>

You are a DeepFake forensics analyst. Please analyze the provided sample and identify any artifacts that indicate it may be inauthentic or manipulated.

Scope:

- The data belongs to sample modality. Tailor your analysis to the specific modality.

Your Task:

- Perform a detailed analysis of artifacts that appear inauthentic or indicative of AI generation or manipulation.
- Focus on concrete, observable evidence. Avoid speculation.

Guidelines:

- Be Thorough: Cover all noticeable artifacts and inconsistencies relevant to this modality.
- Be Accurate: Base claims only on what is present in the sample. Explain why each artifact is suspicious in technical terms.
- Avoid False Positives: Do not label authentic features as inauthentic. If uncertain, state the uncertainty and what additional evidence would be needed.
- Organize Your Response: Use clear headings for each artifact and include short evidence quotes.

Output Format:

- 1) Overall Assessment
 - One sentence rating: Likely Authentic, or Likely DeepFake.
- 2) Artifact Findings

For each finding, provide:

- Title of artifact
- Reason: brief technical rationale

Template for <MCQ>

modality: Which of the following artifacts is present in the sample? Select all that apply. If you select 'None of the options are correct,' do not select any other options.

A. {artifact}

B. {artifact}

C. {artifact}

D. {artifact}

E. None of the options is correct

modality: Where in the sample is the {artifact} located? Select all that apply. If you select 'None of the options are correct,' do not select any other options.

A. {location}

B. {location}

C. {location}

D. {location}

E. None of the options is correct

modality: Visual artifacts are present in which of the following locations? Select all that apply. If you select 'None of the options are correct,' do not select any other options.

A. {location}

B. {location}

C. {location}

D. {location}

E. None of the options is correct

Figure S5. Prompt Template Used for Benchmark Construction for <TFQ>, <MCQ>, and <OEQ>

Prompt for Artifacts Mapping

You are an AI evaluation engine. Your task is to process an analysis of a digital media (the Analysis Text) and determine which artifacts from a predefined list (the Artifact Definitions) are present.

Decision Rules

- Decisions must be based on the criteria in "Artifact Definitions".
- Comparisons or synonyms are valid only if their meaning clearly aligns with the concept defined in the Artifact Definitions.
- For each artifact in Artifact Keys, mark:
 - True — only if the Analysis Text contains sufficient evidence that meets the definition.
 - False — if evidence is missing, ambiguous, contradictory, or the definition is absent.

Output Format (machine-parseable)

- Output only key-value pairs, one per line, in the exact order of Artifact Keys.
- Line format: ArtifactName: True or ArtifactName: False
- Use exactly True/False (capitalized).
- No headers, explanations, extra spaces, blank lines, or punctuation beyond the single colon and single space.

1. Analysis Text
(RESPONSE)

2. Artifact Definitions

You must check for the presence of the following artifacts. An artifact is "True" only if the Analysis Text provides evidence that matches its specific Definition.

- **Blurriness**: ["The loss of sharpness and fine detail, making the image appear out of focus."]
- **Blackiness**: ["Visible square or rectangular patterns on the screen."]
- **Noise**: ["Random, fine speckles or a sandy texture across the image."]
- **Banding**: ["Distinct, abrupt steps or bands in areas that should have a smooth color gradient, like a sunset or a clear sky."]
- **Color Inconsistency**: ["Colors appear unnatural, with excessive saturation or vibrancy that makes the sample look too intense or unrealistic."]
- **Blending Artifacts**: ["Visible boundaries where elements should merge smoothly."]
- ...

Begin Evaluation

Figure S6. Prompt Template Used for Artifacts Mapping

regime. By contrast, these models often exhibit strong performance on quality-related artifacts. This divergence suggests that current systems still lean heavily on low-level signal cues rather than forming robust representations of prosody or speaker plausibility. A salient example is the semantic artifact of unnatural prosody: the waveform may appear clean, but subtle irregularities in rhythm, intonation, or stress make the speech sound implausible to human listeners. Such artifacts are notoriously hard for existing models to detect reliably, underscoring the intrinsic difficulty of

semantic perception in audio.

Secondly, we hypothesize that this difficulty is partly driven by an architectural bias. Most MLLMs rely on audio encoders optimized for transcription or high-level semantic understanding, rather than for preserving speaker-identity fidelity or prosodic consistency. As a result, precisely those cues that are critical for judging who is speaking and whether their timing and intonation patterns are human-plausible are under-emphasized in the learned representations, limiting effective DeepFake perception in the audio modality.

Interpretable Detection, Perception and Hallucination.

We analyze interpretable audio deepfake detection using Type-A and Type-B $\langle\text{OEQ}\rangle$ questions, with full results summarized in Tab. S5. For Type-A $\langle\text{OEQ}\rangle$, only Qwen3-Omni-30B-A3B and Gemini-2.5-Pro produce meaningful artifact-level explanations. Qwen3-Omni achieves the highest *Cover* and $F_{0.5}$ scores, albeit with a moderate level of hallucination, whereas Gemini-2.5-Pro attains slightly lower *Cover* and $F_{0.5}$ scores but produces more consistently grounded descriptions. By contrast, audio-focused models such as Qwen2-Audio-7B, SALMONN-7B, and audio-flamingo-3 yield very low *Cover* and near-saturated hallucination rates, resulting in almost zero $F_{0.5}$ scores. These findings indicate that current audio MLLMs still struggle to provide faithful artifact-level explanations and often hallucinate nonexistent distortions.

Type-B $\langle\text{OEQ}\rangle$ highlights a significant disparity between detection accuracy and explanation quality. SALMONN-7B achieves the highest detection accuracy but offers almost no interpretability, often providing the correct label while generating unreliable explanations. In contrast, Gemini 2.5-Pro demonstrates the opposite trend: its detection accuracy is nearly at chance levels, yet it provides the best interpretability, characterized by the highest *Cover*, reduced hallucination, and the strongest $F_{0.5}$ score. Qwen3-Omni-30B-A3B and Phi-4 fall somewhere in between, exhibiting moderate accuracy and $F_{0.5}$ scores, but still suffering from considerable hallucination. Meanwhile, audio-flamingo-3 performs poorly in both detection and interpretability.

Overall, the audio results reinforce the main tri-perspective conclusion that current models rarely achieve both strong detection and low hallucination in this modality. Audio-centric MLLMs often depend on unclear heuristics and provide explanations that are highly prone to hallucination, whereas stronger multimodal models offer more grounded reasoning but show only slight improvements over random guessing. These findings highlight the need for better speech-specific perception modules and enhanced modeling of prosody and identity cues to achieve more reliable audio DeepFake detection.

Table S4. Evaluation of Audio Deepfake Perception

MLLM	$\langle\text{TFQ}\rangle$				$\langle\text{MCQ}\rangle$	
	<i>Semantic Quality</i>	<i>Avg.</i>	<i>Rank</i>		<i>General Rank</i>	
Random Guess	50.00%	50.00%	50.00%	–	0.00	–
Qwen2-Audio-7B	44.50%	67.88%	56.19%	2	0.01	3
Qwen3-Omni-30B-A3B	32.76%	67.37%	50.07%	3	-0.15	5
Phi-4	5.50%	68.45%	36.98%	5	-0.06	4
Audio-Flamingo-3	6.91%	67.88%	37.40%	4	0.10	1
Gemini-2.5-pro	63.65%	50.13%	56.89%	1	0.04	2
Average	30.66%	64.34%	47.51%	–	-0.01	–

Table S5. Evaluation of Interpretable Audio Deepfake Detection, Perception and Hallucination Robustness

MLLM	Type A $\langle\text{OEQ}\rangle$				Type B $\langle\text{OEQ}\rangle$				
	Audio				Audio				
	<i>Cover</i> \uparrow	<i>CHAIR</i> \downarrow	<i>Hal</i> \downarrow	$F^{0.5}$ \uparrow	<i>ACC</i>	<i>Cover</i> \uparrow	<i>CHAIR</i> \downarrow	<i>Hal</i> \downarrow	$F^{0.5}$ \uparrow
Qwen2-Audio-7B	0.0446	0.9342	0.9421	0.0580	0.3799	0.2356	0.6388	0.6756	0.3166
Qwen3-Omni-30B-A3B	0.5278	0.2011	0.2867	0.7031	0.4082	0.3690	0.5279	0.6756	0.4312
SALMONN-7B	0.0012	0.9973	0.9973	0.0021	0.5722	0.0472	0.9225	0.9225	0.0673
Phi-4	0.1983	0.7375	0.7736	0.2360	0.3949	0.3185	0.4967	0.5187	0.4398
Audio-Flamingo-3	0.0811	0.8708	0.8708	0.1129	0.3732	0.0206	0.9635	0.9635	0.0311
Gemini 2.5-pro	0.3065	0.5079	0.5339	0.4279	0.4859	0.5470	0.2106	0.2736	0.7022

I. Extended Evaluation

I.1. Evaluation Setup

Evaluation models and modalities. For visual modalities, we consider open-source MLLMs including InternVL2.5/3_5 [R13, R83], Qwen3-Omni/VL [R5, R91], LLaVA-OV [R46], MiniCPM-V [R96], MiMo-VL [R102], Idefics2 [R44], Mantis [R39], Phi-4 [R1], and the forensic-focused FakeShield [R92] and FakeVLM [R87]. These are compared against proprietary baselines: GPT-5 [R63], Gemini 2.5-Pro [R17], and Claude Sonnet 4.5 [R3]. Audio performance is evaluated using Qwen2-Audio [R15], Qwen3-Omni, Phi, Audio-Flamingo-3 [R27], and SALMONN-7B [R77], with Gemini 2.5-Pro serving as the proprietary reference.

Experimental protocol. All experiments are conducted in a zero-shot setting, where each sample is processed independently without task-specific fine-tuning. For each query, we provide the model with the question prompt together with the corresponding image, video, or audio input. For video tasks, we either use a 16-frame clip (when frame sampling is configurable) or the model’s default frame sampling policy. Unless otherwise noted, the same protocol is applied consistently across all models and modalities.

I.2. More Quantitative Results

Interplay between perception, hallucination, and detection. To understand how the three evaluation dimensions of

TriDF relate to one another, we analyze the correlations between perception, hallucination and detection performances across all 22 evaluated models. For each model m , we compute three macro-averaged scores over all available samples: (i) perception P_m , defined as Type-A *Cover*; (ii) hallucination severity H_m , defined as Type-A *CHAIR*; and (iii) detection D_m , defined as Type-B $\langle\text{OEQ}\rangle$ detection accuracy.

The resulting correlation matrix in Fig. S7 reveals a tightly coupled but non-degenerate triad. Perception and detection are moderately positively correlated ($r(P, D) \approx 0.60$): models that cover more ground-truth artifacts in Type-A explanations tend to achieve higher Type-B detection accuracy. Hallucination severity is also strongly coupled to detection ($r(H, D) \approx -0.60$), with more hallucinated artifacts associated with lower accuracy. Although perception and hallucination are negatively correlated ($r(P, H) \approx -0.44$), the magnitude of this correlation is relatively moderate. It indicates that while models that recognize more genuine artifacts tend to hallucinate less, the two aspects remain far from interchangeable. The overall correlation matrix shows that perception and detection are moderately aligned, while hallucination undermines detection and is moderately anti-correlated with perception.

However, when we further stratify models by hallucination severity, a more revealing pattern emerges. We define hallucination regimes using the empirical sample distribution: all samples with $H = 1$ form a high-hallucination regime (High-H), while samples with $H < 1$ are split at the 33rd and 67th percentiles into Low-H and Mid-H, and analyze the fake-only subset of TriDF. Independently, we discretize perception into five equal-width bins based on Type-A *Cover* (0–0.2, 0.2–0.4, . . . , 0.8–1.0). For each hallucination regime and perception bin, we then compute the average fake detection accuracy and plot the resulting curves in Fig. S8.

The stratified curves reveal a clear three-way interaction. In the Low-H and Mid-H regimes, fake-detection accuracy is high at low *Cover* and rapidly saturates near perfect accuracy as *Cover* increases, indicating that once explanations are largely grounded, additional perceptual coverage yields gains on detection accuracy. In contrast, in the High-H regime, DeepFake detection accuracy remains close to chance across all perception bins and is effectively insensitive to *Cover*. Even when models capture numerous artifacts (high P), severe hallucination in Type-A explanations is associated with systematic failures to flag fakes in Type-B decisions.

Both analyses shown in Fig. S7 and Fig. S8 demonstrate that hallucination can disrupt the natural link between evidence recognition in perception and detection decision-making. The findings reinforce that perception, detection, and hallucination capture fundamentally distinct aspects of

model behavior, and that reliable DeepFake detection requires balanced progress across all three dimensions. Improving only perception or only classification is insufficient. Addressing these intertwined but independent factors is crucial for building trustworthy and human-aligned detection systems capable of withstanding increasingly sophisticated forgeries.

Benefit-Cost Analysis of Localization Hints. As discussed in RQ2 in the main paper, we quantify the efficacy of localization hints and define *Benefit* and *Cost* as the percentages of questions where the hint respectively corrects an initial error or induces a new one. Their difference, *Net Benefit*, serves as the primary indicator of genuine performance gain from spatial guidance. The results are summarized in Tab. S6. Localization hints generally yield a positive *Net Benefit*, though gains vary by architecture. InternVL2.5-8B and Claude Sonnet 4.5 achieve peak efficiency (2.53% and 2.47% *Net Benefit*), demonstrating an effective ability to leverage spatial cues. Conversely, Gemini 2.5-Pro and Qwen3-VL-30B-Instruct exhibit negative *Net Benefit* (−0.30% and −0.32%), suggesting that for certain high-capacity architectures, external hints may introduce disruptive noise. This non-universal efficacy underscores a persistent architectural gap in reconciling external spatial grounding with internal visual representations.

Table S6. RQ2. Benefit and Cost of localization hints.

MLLM	Benefit (%)	Cost (%)	Net Benefit (%)	Rank
InternVL2.5-8B [R13]	3.21	0.68	2.53	1
InternVL2.5-26B [R13]	4.28	1.90	2.38	4
InternVL2.5-38B [R13]	4.22	1.78	2.44	3
InternVL3.5-8B [R83]	<u>10.87</u>	<u>10.10</u>	0.78	7
InternVL3.5-38B [R83]	4.93	3.30	1.63	5
Qwen3-Omni-30B-A3B-Instruct [R91]	6.92	6.55	0.37	9
Qwen3-VL-8B-Instruct [R5]	8.03	7.30	0.73	8
Qwen3-VL-30B-Instruct [R5]	7.01	7.33	-0.32	11
GPT-5 [R63]	6.57	5.65	0.92	6
Gemini 2.5-Pro [R17]	11.67	11.97	-0.30	10
Claude Sonnet 4.5 [R3]	3.17	0.70	2.47	2

I.3. More Qualitative Results

Based on the provided documents, the case studies utilize three distinct evaluation formats, $\langle\text{TFQ}\rangle$, $\langle\text{MCQ}\rangle$, and $\langle\text{OEQ}\rangle$, to assess model performance in detecting synthesis and manipulation artifacts.

$\langle\text{TFQ}\rangle$ focuses on binary verification, prompting models to simply confirm or deny the presence of specific defects, such as detecting “Buzz” in an audio clip or identifying “Temporal Inconsistency” in a video subject’s upper limb. As shown in Fig. S9, Gemini 2.5-Pro outperforms both powerful general-purpose model (e.g., Qwen3-Omni-30B-A3B-Instruct) and specialized model, Audio-Flamingo-3.

Correlation Matrix (P, H, D)

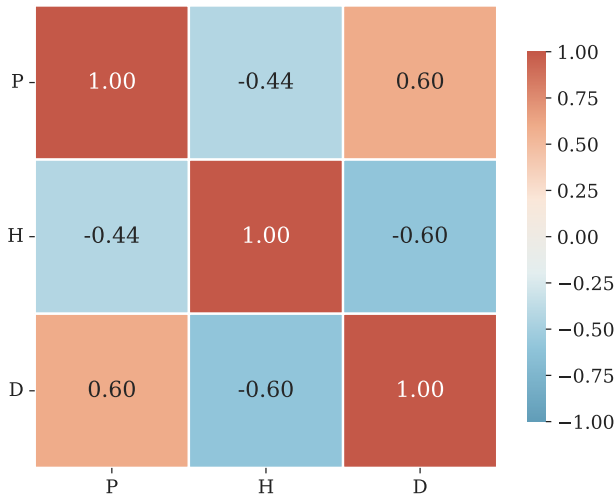


Figure S7. Model-level correlation matrix for perception (P), hallucination severity (H), and detection (D). Perception is positively correlated with detection accuracy, while hallucination is negatively correlated with both, supporting the three-dimensional P–H–D view of MLLM-based DeepFake detection.

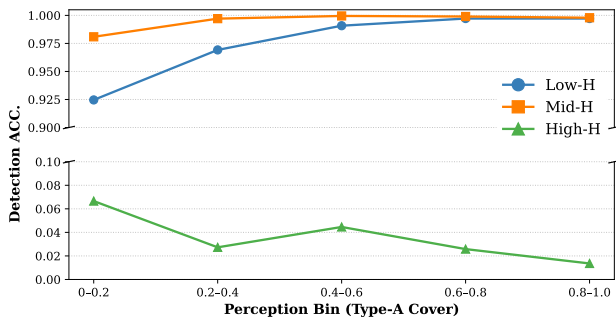


Figure S8. Stratified perception–detection curves on TriDF: fake-detection accuracy vs. binned Type-A *Cover* under three Type-A *CHAIR* regimes, showing that strong hallucination keeps detection near chance even with high perceptual coverage.

Conversely, GPT-5 struggles in this example because it cannot handle raw video inputs without preprocessing, which hinders its ability to understand temporal relationships.

<MCQ> tests the ability to categorize or locate specific errors, asking models to identify semantic issues like “Anatomical Inconsistency” or select specific regions where artifacts appear, such as the “Ear” or “Background”. Within the two examples in Fig. S10, the evaluation metric is strict: models must answer all options correctly to receive the maximum score of 1. Any incorrect selection results in a penalty, preventing a full score.

Finally, <OEQ> requires a more granular, descriptive analy-

sis, asking models to justify a “Likely Manipulated” verdict by detailing observable flaws like “Inconsistent Lighting”, “Unnatural Shadow”, or a “Blurred Background”. Fig. S11 highlights the variance in model perspective: Gemini 2.5-Pro provides a focused, context-aware analysis of lighting physics on a specific object (a cat), whereas InternVL2_5-8B generates a generic list of DeepFake flaws typically associated with human subjects.

J. Future Direction of DeepFake Detection

TriDF fills an important gap in existing evaluation resources by enabling systematic analysis of all three components. Looking forward, TriDF provides several avenues for advancing future DeepFake detection techniques. First, the fine-grained artifact taxonomy offers a structured supervisory signal that can guide new models to focus on meaningful manipulation cues rather than dataset-specific shortcuts. Second, the multimodal and diverse generator design creates a challenging testbed that encourages the development of detectors with stronger generalization across synthesis pipelines. Third, the hallucination evaluation reveals failure modes in explanation generation and provides a foundation for designing models that produce grounded, reliable reasoning. Finally, as new generative techniques and modalities emerge, TriDF can be extended to support evolving research needs, serving as a long-term platform for building trustworthy and deployable DeepFake detection systems.

K. Release Plan and Ethics Statement

All datasets utilized in this benchmark are sourced from publicly available repositories. DeepFake generation was conducted strictly for academic and research purposes to advance the fields of media forensics and authenticity detection. Our research team explicitly opposes the malicious application of this technology and condemns any use of this benchmark or the associated data for deceptive, harmful, or misinformation-related purposes.

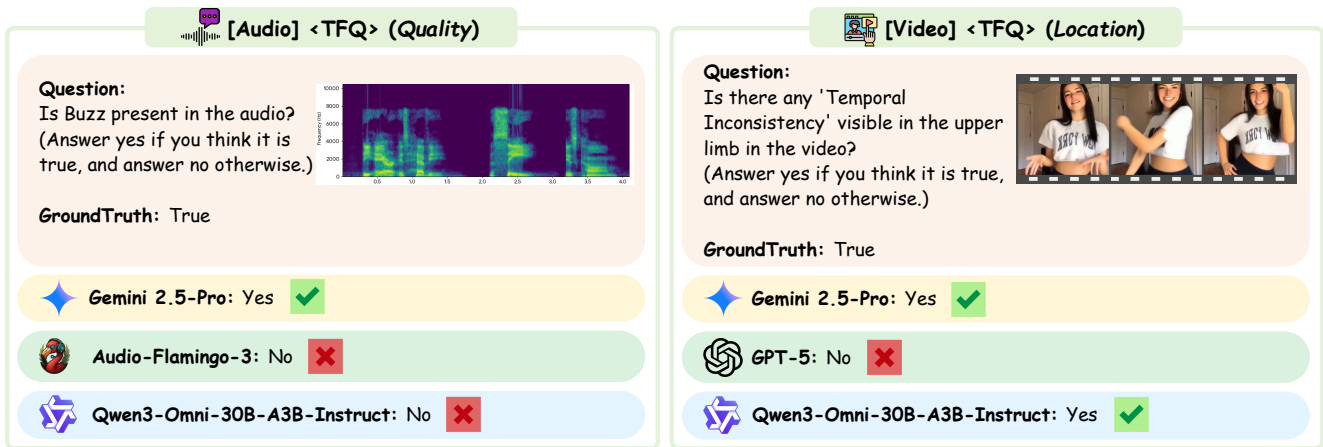


Figure S9. Examples of <TFQ>

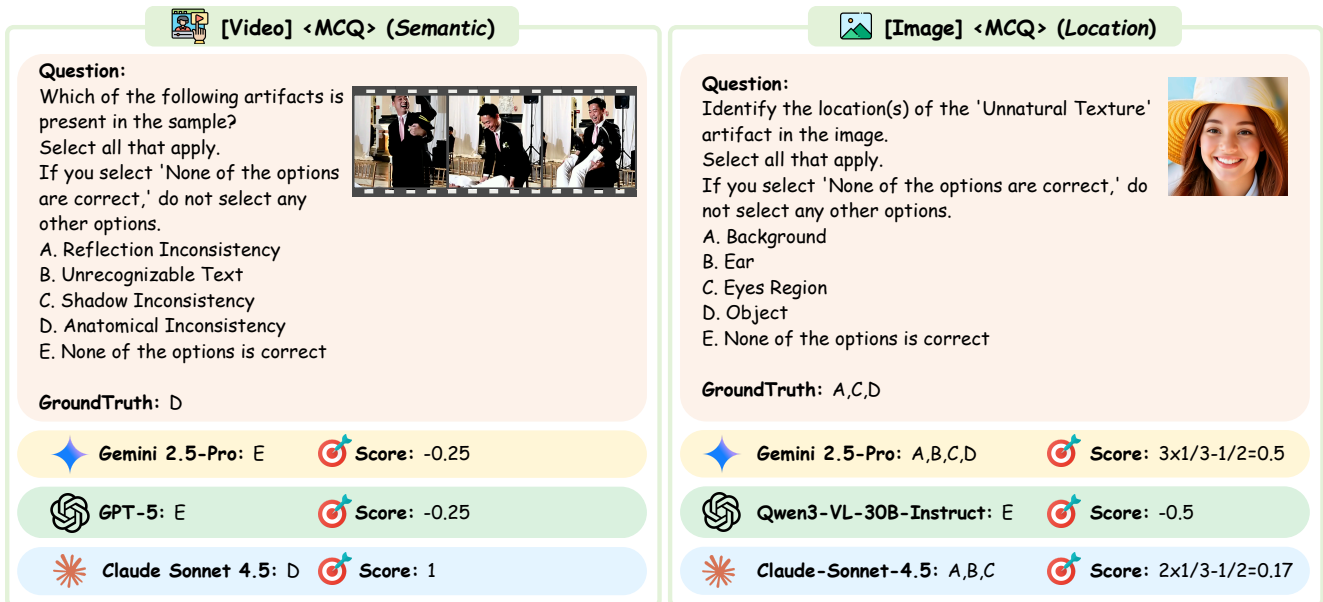


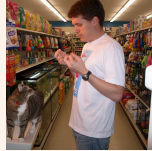
Figure S10. Examples of <MCQ>



[Image] Type-A <OEQ>

Question:

Please perform a detailed analysis of artifacts in this sample that seem inauthentic.



◆ Gemini 2.5-Pro:

The analysis reveals several inconsistencies, primarily related to the cat, which suggest that it may have been digitally added to the scene.

Artifact 1:

- **Inconsistent Lighting and Shadows:** The most significant evidence of manipulation is the incorrect shadowing associated with the cat.
- **Light Source:** The dominant light source in the pet store aisle is the series of long, bright fluorescent lights on the ceiling. This should create distinct, downward-cast shadows.
- **Missing Shadow:** The cat is standing on a bright white surface, which should make any cast shadow clearly visible. However, the cat casts almost no discernible shadow. There are some very faint contact shadows directly under its paws, but there is no larger, softer shadow on the surface that would be consistent with its body blocking the strong overhead light.
- **Lighting Mismatch:** The lighting on the cat itself seems generally consistent with an overhead source, with highlights on its head and back. However, the lack of a corresponding shadow on the surface it occupies creates a strong contradiction, suggesting the cat was photographed under different lighting conditions and composited into this image.

Artifact 2:

- **Unnatural Edges and Composition:** The boundary between the cat and the background shows subtle signs of digital insertion.
- **Soft/Blurry Outline:** The edges around the cat's fur, particularly along its back and head against the fish tanks, appear unnaturally soft and slightly blurred. This halo effect is common when a subject is inexpertly cut out from its original background and placed onto a new one.
- **Lack of Integration:** The cat's paws rest on the white surface, but they don't seem to interact with it naturally. The contact points lack the subtle shadows and pressure points that would make the composition believable. The overall effect is that the cat appears to be floating on the surface rather than standing firmly on it.

Summary:

The primary indicators of manipulation in this image are centered on the cat. The stark inconsistency between the bright, overhead light source and the almost complete absence of a cast shadow from the cat is the strongest piece of evidence. This, combined with the soft and unnatural edges around the cat's outline, strongly suggests that the cat was not originally part of the scene and has been digitally composited into the photograph.



[Video] Type-B <TFQ>

Question:

Decide whether the provided sample is authentic or manipulated. Perform a detailed analysis of artifacts that appear inauthentic or indicative of synthesis/manipulation.



◆ InternVL2_5-8B

Likely Manipulated.

Artifact Findings

- 1) **Unnatural Lighting:** The lighting on the person's face and upper body is inconsistent with the ambient lighting in the background. The face is well-lit, while the upper body has a different lighting quality, suggesting a possible post-production adjustment.
- 2) **Unnatural Shadow:** There is an unnatural shadow on the person's left side (viewer's right side) that does not align with the light source. The shadow appears too sharp and distinct, indicating possible manipulation.
- 3) **Blurred Background:** The background is slightly blurred, which is unusual for a video taken in a static environment. This blurring effect is more typical of video editing software, suggesting post-production adjustments.
- 4) **Inconsistent Clothing Texture:** The texture of the sweatshirt appears inconsistent, with some areas looking smoother and others more textured. This inconsistency is not typical of natural clothing and suggests possible manipulation.
- 5) **Unnatural Hand Position:** The person's hand position appears unnatural, with fingers slightly bent and positioned in a way that does not align with the rest of the body's posture. This could be a result of post-production adjustments.
- 6) **Inconsistent Facial Expression:** The facial expression appears slightly off, with the mouth and eyes not fully synchronized. This inconsistency is not typical of natural facial movements and suggests possible manipulation.
- 7) **Unnatural Hair Movement:** The hair appears to move in a way that is not consistent with the rest of the body's movement. This could be a result of post-production adjustments to the hair.
- 8) **Inconsistent Background Objects:** The background objects, such as the plants and the building, appear slightly out of focus and do not align perfectly with the person's position. This inconsistency is not typical of natural environments and suggests possible manipulation.
- 9) **Unnatural Body Posture:** The person's body posture appears slightly off, with the shoulders and hips not fully aligned. This inconsistency is not typical of natural body movements and suggests possible manipulation.
- 10) **Inconsistent Clothing Color:** The color of the sweatshirt appears slightly off, with some areas looking darker and others lighter. This inconsistency is not typical of natural clothing.

Figure S11. Examples of <OEQ>

Bibliography

- [R1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 9
- [R2] Coqui AI. Coqui X-TTS: A hugging face space for text-to-speech, 2025. 5
- [R3] Anthropic. Introducing Claude 3.5 Sonnet, 2024. 9, 10
- [R4] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *ACL*, 2022. 5
- [R5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 9, 10
- [R6] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 5, 7
- [R7] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *SIGGRAPH*, 2025. 2, 5
- [R8] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. HyperReenact: One-shot reenactment via jointly learning to refine and retarget faces. In *ICCV*, 2023. 5
- [R9] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VggFace2: A dataset for recognising faces across pose and age. In *FG*, 2018. 2, 5
- [R10] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. MagicPose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *ICML*, 2024. 5
- [R11] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024. 2, 5, 7
- [R12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 2, 5
- [R13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 9, 10
- [R14] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Wan Lee. Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. In *Interspeech*, 2023. 5
- [R15] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 9
- [R16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, 2018. 2, 5
- [R17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 9, 10
- [R18] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *ICLR*, 2025. 5
- [R19] D-iD. D-iD, 2024. 5
- [R20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2
- [R21] Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-Avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025. 5
- [R22] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024. 5
- [R23] ElevenLabs. ElevenLabs, 2025. 2, 5
- [R24] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 5, 7
- [R25] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. SkyReels-A2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 2, 5
- [R26] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *CVPR*, 2021. 5
- [R27] Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *NeurIPS*, 2025. 9
- [R28] Google. Gemini 2.5 Flash Image (Nano Banana), 2025. 2, 5, 7

- [R29] Google. Veo 3, 2025. 5, 7
- [R30] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Live-Portrait: Efficient portrait animation with stitching and re-targeting control. *arXiv preprint arXiv:2407.03168*, 2024. 5
- [R31] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 5, 7
- [R32] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *ECCV*, 2024. 5
- [R33] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2
- [R34] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *ICCV*, 2023. 5
- [R35] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. HunyuanCustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 2, 5, 7
- [R36] Zhizhong Huang, Siteng Ma, Junping Zhang, and Hongming Shan. Adaptive nonlinear latent transformation for conditional face editing. In *ICCV*, 2023. 5
- [R37] Ziyao Huang, Fan Tang, Yong Zhang, Juan Cao, Chengyu Li, Sheng Tang, Jintao Li, and Tong-Yee Lee. Identity-preserving face swapping via dual surrogate generative models. *ACM TOG*, 2024. 5
- [R38] Yasamin Jafarian and Hyun Soo Park. Self-supervised 3d representation learning of dressed humans from social media videos. *IEEE TPAMI*, 2022. 2, 5
- [R39] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *TMLR*, 2024. 9
- [R40] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *ICCV*, 2025. 5
- [R41] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2024. 5
- [R42] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2, 5
- [R43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5
- [R44] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, 2024. 9
- [R45] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 2, 5
- [R46] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *TMLR*, 2025. 9
- [R47] Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. LatentSync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024. 5
- [R48] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM MM*, 2019. 2
- [R49] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From Generation to Judgment: Opportunities and challenges of llm-as-a-judge. In *EMNLP*, 2025. 7
- [R50] Jianhui Li, Jianmin Li, Haoji Zhang, Shilong Liu, Zhengyi Wang, Zihao Xiao, Kaiwen Zheng, and Jun Zhu. PREM3D: 3d consistent precise image attribute editing from a single image. In *CVPR*, 2023. 5
- [R51] Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. FIVE-Bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *ICCV*, 2025. 2, 5
- [R52] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. FakeBench: Uncover the achilles' heels of fake images with large multimodal models. *IEEE TIFS*, 2025. 6
- [R53] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 5
- [R54] Kun Liu, Qi Liu, Xinchun Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. HOIGen-1M: A large-scale dataset for human-object interaction video generation. In *CVPR*, 2025. 2, 5
- [R55] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *ICCV*, 2025. 2, 5
- [R56] Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024. 2, 5
- [R57] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2, 5, 7
- [R58] Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing Blind Guessing: Calibration of selection bias in multiple-choice question answering by video language models. In *ACL*, 2025. 6
- [R59] Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer.

- Rethinking the Role of Demonstrations: What makes in-context learning work? In *EMNLP*, 2022. 6
- [R60] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction using crowdsourced datasets. In *Interspeech*, 2021. 2
- [R61] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE SPL*, 2012. 2
- [R62] OpenAI. GPT-4o, 2024. 6
- [R63] OpenAI. GPT-5, 2025. 9, 10
- [R64] OpenAI. GPT-4o image, 2025. 2, 5, 7
- [R65] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, 2015. 2, 5
- [R66] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. ControlNeXt: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2, 5
- [R67] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2, 5
- [R68] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 2
- [R69] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. OpenVoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023. 2, 5
- [R70] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2, 5
- [R71] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 5
- [R72] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *CVPR*, 2024. 2, 5
- [R73] Kaede Shiohara, Xingchao Yang, and Takafumi Takeuchi. BlendFace: Re-designing identity encoders for face-swapping. In *ICCV*, 2023. 5
- [R74] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 2
- [R75] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2, 5
- [R76] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, 2017. 2, 5
- [R77] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *ICLR*, 2024. 9
- [R78] Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. MIGE: Mutually enhanced multimodal instruction-based image generation and editing. In *ACM MM*, 2025. 5
- [R79] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. MotionEditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024. 5
- [R80] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 5, 7
- [R81] Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. Forensics-Bench: A comprehensive forgery detection benchmark suite for large vision language models. In *CVPR*, 2025. 6
- [R82] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 2, 5
- [R83] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 9, 10
- [R84] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. HifiFace: 3d shape and semantic prior guided high fidelity face swapping. In *IJCAI*, 2021. 5
- [R85] Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. 2
- [R86] Huawei Wei, Zejun Yang, and Zhisheng Wang. AniPortrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 5
- [R87] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the Fake: Large multimodal model-based synthetic image detection with artifact explanation. In *NeurIPS*, 2025. 9
- [R88] Mirjam Wester. The emime bilingual database. Technical report, The University of Edinburgh, 2010. 2, 5
- [R89] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyang Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2, 5, 7
- [R90] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-More Generalization:

- Unlocking more controllability by in-context generation. In *ICCV*, 2025. 5
- [R91] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 9, 10
- [R92] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *ICLR*, 2025. 9
- [R93] Junichi Yamagishi, Christophe Veaux, Kirsten MacDon-ald, et al. CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017. 2, 5
- [R94] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. StyleGANEX: Stylegan-based manipulation beyond cropped aligned faces. In *ICCV*, 2023. 5
- [R95] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert trans-former. In *ICLR*, 2025. 5
- [R96] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 9
- [R97] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A comprehensive syn-thetic data detection benchmark using large multimodal models. In *ICLR*, 2025. 6
- [R98] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. ImgEdit: A unified image editing dataset and benchmark. In *NeurIPS*, 2025. 2, 5
- [R99] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023. 2, 5
- [R100] Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *NeurIPS*, 2025. 2, 5
- [R101] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *CVPR*, 2025. 2, 5
- [R102] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025. 9
- [R103] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. DwNet: Dense warp-based network for pose-guided human video generation. In *BMVC*, 2019. 2, 5
- [R104] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A cor-pus derived from librispeech for text-to-speech. In *Inter-speech*, 2019. 2, 5
- [R105] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shecht-man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2
- [R106] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. 5
- [R107] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *ECCV*, 2024. 6
- [R108] Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. MuseTalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. *arXiv preprint arXiv:2410.10122*, 2025. 2, 5
- [R109] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 2, 5
- [R110] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. DInet: Deformation in-painting network for realistic face visually dubbing on high resolution video. In *AAAI*, 2023. 5
- [R111] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. DiffSwap: High-fidelity and con-trollable face swapping via 3d-aware masked diffusion. In *CVPR*, 2023. 5
- [R112] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust mul-tiple choice selectors. In *ICLR*, 2024. 6
- [R113] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards explain-able and generalizable ai-generated image detection via multimodal large language models. In *ICCV*, 2025. 6
- [R114] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 2, 5