Multimodal AI Detection In Two Words

Daniel Fein[1] **Arpita Singhal**[1] **Maneesh Agrawala**[1] **Maty Bohacek**[1]
[1]Stanford University

Abstract

AI-generated text, images, audio, and video are now ubiquitous and difficult for humans to reliably identify, making automated detection a practical necessity. Existing detectors are, however, largely unimodal and their predictions are not faithfully explainable. We propose a unified method that addresses both of these issues by extending *neologism learning* from steering of text generation to multimodal classification. We add two tokens, <REAL> and <AIGEN>, to the vocabulary of an otherwise-frozen multimodal LLM (MLLM) and train only their embedding rows (less than 0.001% of the backbone), making the model's next-token distribution encode the real vs. AI-generated prediction. Because all modalities exist in a shared input embedding space, the same token pair may apply across text, image, audio, and video—without the need for modality-specific heads. We show that configurations trained on only some of these modalities do, in fact, transfer to held-out modalities. The trained tokens further admit natural-language descriptions of what <AIGEN> has come to mean, without explicit supervision; we validate their causal faithfulness via plug-in evaluation. Across text, audio, image, and video benchmarks, our method matches or exceeds modality-specific detectors and outperforms MLLM-based explainable detectors, despite training orders of magnitude fewer parameters without human-annotated explanations.

1 Introduction

Generative AI for text, images, video, and audio now produces synthetic content that is both ubiquitous and difficult to identify [Liu et al., 2024b, Cao, 2025, Livernoche et al., 2025, Dolezal et al., 2026], enabling misinformation [Chandra et al., 2025, Livernoche et al., 2025], identity fraud [Khalid et al., 2021], non-consensual imagery, fabricated evidence, and academic dishonesty [Cao, 2025]. Humans are already incapable of reliably distinguishing real from AI-generated content [Cooke et al., 2024, Barrington et al., 2025], and the gap may widen as AI's generation capability improves.

As journalists, educators, and courts increasingly turn to automated detectors to fill this gap, a binary prediction often is not sufficient, since their decisions require verifiable evidence—in the language of machine learning, explainability. Some traditional computer vision explanation techniques [Selvaraju et al., 2017] can highlight regions of interest but generally cannot explain what is wrong with them [Tsigos et al., 2024]. The community has thus moved toward natural-language explanations [Zhang et al., 2024, Li et al., 2025b, Liu et al., 2024a, Xu et al., 2024, Huang et al., 2025b, Hondru et al., 2026, Tariq et al., 2025, Yang et al., 2025], but these rely on human-annotated rationales: the explanations may not reflect what the detector actually learned, and the supervision becomes increasingly unreliable as human ability to spot AI content degrades.

AI detection is mostly fragmented along modality lines. Most methods operate on a single modality, with separate datasets and pipelines for images [Ojha et al., 2023, Khan and Dang-Nguyen, 2024, Cozzolino et al., 2024, Pellegrini et al., 2025, Chandra et al., 2025], video [Bohacek and Farid, 2024a], audio, or text. Existing multimodal methods are almost exclusively *audio-visual*, based on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

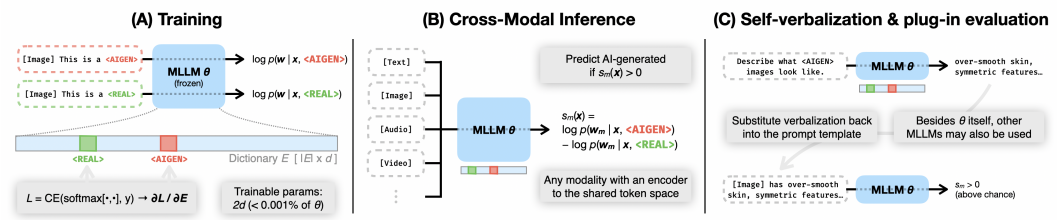


Figure 1: **Method overview. (A) Training.** We add tokens <REAL> and <AIGEN> to a frozen MLLM θ and pair each with input x in contrastive prompts; only the two corresponding rows of E get updated based on log-prob cross-entropy of a fixed continuation w. **(B) Cross-modal inference.** These two rows define a modality-agnostic detector $s_m(x) = \log p(w_m \mid x, \texttt{<AIGEN>}) - \log p(w_m \mid x, \texttt{<REAL>})$, predicting AI-generated when $s_m(x) > 0$. **(C) Self-verbalization & plug-in evaluation.** We prompt θ to describe each token in natural language (top); plug-in evaluation substitutes the description back into the prompt and rescores (bottom), with above-chance s_m indicating faithfulness.

inconsistencies between audio and face video [Khalid et al., 2021, Ilyas et al., 2023, Yang et al., 2023, Wang et al., 2025a, Zhong et al., 2024, Bohacek and Farid, 2024b, 2022, Barrington et al., 2024, Agarwal et al., 2020].

Recent work has begun adapting vision-language models (VLMs) and multimodal LLMs (MLLMs) to AI detection. Zero-shot prompting of frontier MLLMs underperforms task-specific detectors [Jia et al., 2024, Shi et al., 2025, Ojha et al., 2023, Khan and Dang-Nguyen, 2024, Cozzolino et al., 2024]. MLLM-based explainable detectors [Zhang et al., 2024, Li et al., 2025b, Liu et al., 2024a, Xu et al., 2024, Huang et al., 2025b, Hondru et al., 2026] are uniformly trained on human-annotated rationales, so their explanations are supervised rather than derived from what the detector has learned. The closest cross-modality analog, Huang et al. [2025a], applies a shared encoder [Girdhar et al., 2023, Zhu et al., 2023a] to text, image, and voice but retains per-modality heads and offers no explanations; Wen et al. [2025] covers image and video only; Li et al. [2025a] covers only watermarking. To our knowledge, no prior work unifies cross-modality detection of AI-generated content in a unified decision representation or provides natural-language explanations without explicit supervision.

This work. We address both the modality fragmentation and supervision problems by extending *neologism learning* [Hewitt et al., 2025a,b, Park et al., 2025]—a parameter-efficient technique that introduces new words into a frozen LM—from steering of text generation to multimodal classification. As illustrated in Fig. 1, we add two tokens, <REAL> and <AIGEN>, to a MLLM's vocabulary and train only their embedding rows, so that the model assigns higher probability to a designated continuation depending on whether the context is real or AI-generated. Our contributions are as follows:

- **A unified and efficient method for AI-generated content detection across modalities (Sec. 3).** We train a single pair of token embeddings in a frozen MLLM ($\sim 2\,d$ parameters, $< 0.001\%$ of the backbone) to discriminate AI-generated from human-produced content. The model hence avoids modality-specific classifier heads while applying the same representation across all input modalities.
- **Competitive detection (Sec. 4.1) and measurable cross-modal transfer (Sec. 4.3) performance through a shared embedding.** While training orders of magnitude fewer parameter, our approach is competitive with unimodal embedding- and MLLM-based detectors on standard benchmarks. Training on a subset of modalities also naturally transfers to held-out modalities at test time.
- **Faithful natural-language explanations without explicit supervision (Sec. 4.7).** We elicit free-form self-verbalizations of <AIGEN> and <REAL>; unlike prior explainable detectors, our model doesn't require explanation supervision at training time, avoiding known faithfulness issues of such approaches. We validate this using *plug-in evaluation* [Hewitt et al., 2025b], replacing the token with its verbalization and comparing performance.

2 Preliminaries: Neologism Learning and Self-Verbalization

We extend neologism learning [Hewitt et al., 2025a,b] to multimodal classification. Below, we briefly review the framework and its self-verbalization mechanism.

Notation. Let $\mathcal{M}_\theta$ denote a pretrained, possibly multimodal LM with parameters θ, vocabulary $\mathcal{V}$, and input embedding matrix $\boldsymbol{E} \in \mathbb{R}^{d\times|\mathcal{V}|}$, where row $\boldsymbol{E}_v \in \mathbb{R}^d$ embeds token v. Given an input sequence $\boldsymbol{x}_{<t}$, the model outputs a next-token distribution $p_\theta(\cdot \mid \boldsymbol{x}_{<t}) \in \Delta^{|\mathcal{V}|}$, with the output projection tied to $\boldsymbol{E}^\top$. When $\mathcal{M}_\theta$ supports non-text modalities, modality-specific encoders project image patches, audio frames, and video frames into the same d-dimensional space as $\boldsymbol{E}$ [Girdhar et al., 2023, Zhu et al., 2023a, Grattafiori et al., 2024, Google DeepMind, 2026]; the backbone thus processes a sequence of token vectors across modalities in a shared space.

Neologism learning. Hewitt et al. [2025a,b] introduce neologism learning as a parameter-efficient method to steer a frozen LM toward a target concept. We add a new token $c \notin \mathcal{V}$ to the vocabulary, $\mathcal{V}' = \mathcal{V} \cup \{c\}$, with a corresponding embedding row $\boldsymbol{E}_c \in \mathbb{R}^d$; keeping θ frozen, we update only $\boldsymbol{E}_c \in \mathbb{R}^d$ via gradient descent on a behavioral loss over input-output examples of the target concept. The trained $\boldsymbol{E}_c$ hence encodes the concept as a single direction in the LM's input embedding space, requiring only d parameters [Park et al., 2025].

Self-verbalization. Because θ remains frozen, the LM can directly describe what c has learned to represent [Hewitt et al., 2025b]. Passing c inside a prompt such as *"List five synonyms for c"* or *"Describe what c responses are characterized by"* yields a natural-language completion: since $\boldsymbol{E}_c$ sits alongside existing vocabulary embeddings, the decoder treats it like any other token. Because these verbalizations are not guaranteed to be faithful, Hewitt et al. [2025b] propose *plug-in evaluation*: replacing c with its verbalization in the original prompt and comparing whether it steers the model similarly.

3 Method

We extend neologism learning from steering text generation to multimodal binary classification. As illustrated in Fig. 1, we add two tokens, <REAL> and <AIGEN>, to a frozen MLLM's vocabulary and train their embedding rows $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}}) \in \mathbb{R}^{2d}$, so that the model's next-token distribution discriminates between AI-generated and human-produced content.

3.1 Detection via Token-Pair Log-Probability Comparison

Let $\mathcal{X}_m$ denote the input space for modality $m \in \mathcal{M} \coloneqq \{\text{text}, \text{image}, \text{audio}, \text{video}\}$, and $y \in \{0, 1\}$ denote the binary label (0 = human-produced, 1 = AI-generated). Each modality uses a *prompt template* $\Pi_m(\tau, \boldsymbol{x})$ that places a token τ alongside an input $\boldsymbol{x} \in \mathcal{X}_m$ in a user turn:

$$\Pi_{\text{image}}(\tau, \boldsymbol{x}) \;=\; \tau\ \langle \text{image}(\boldsymbol{x})\rangle\ \text{“This is a } \tau\text{”},$$
$$\Pi_{\text{text}}(\tau, \boldsymbol{x}) \;=\; \tau\ \boldsymbol{x}\ \text{“This is a } \tau\text{”},$$

with analogous templates for audio and video. Passing $\boldsymbol{x}$ through $\mathcal{M}_\theta$ twice, first with $\tau =$ <REAL> and second with $\tau =$ <AIGEN>, yields two next-token distributions, which we evaluate at a fixed *continuation token* $w_m \in \mathcal{V}$ where the modality is specified. The prediction comes down to the difference between $\boldsymbol{E}_{\texttt{<REAL>}}$ and $\boldsymbol{E}_{\texttt{<AIGEN>}}$, with the resulting modality-m score being

$$s_m(\boldsymbol{x}) \;\coloneqq\; \log p_\theta\big(w_m \mid \Pi_m(\texttt{<AIGEN>}, \boldsymbol{x})\big) \;-\; \log p_\theta\big(w_m \mid \Pi_m(\texttt{<REAL>}, \boldsymbol{x})\big), \tag{1}$$

with binary prediction $\hat{y}(\boldsymbol{x}) = \mathbb{1}[s_m(\boldsymbol{x}) > 0]$. Because s_m is continuous, it directly supports threshold-based ROC analysis out of the box.

Training objective. Given a labeled dataset $\mathcal{D}_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the two log-probabilities from Eq. 1 form a two-class logit vector trained via cross-entropy:

$$\boldsymbol{z}_i \;=\; \big(\log p_\theta(w_m \mid \Pi_m(\texttt{<REAL>}, \boldsymbol{x}_i)),\ \log p_\theta(w_m \mid \Pi_m(\texttt{<AIGEN>}, \boldsymbol{x}_i))\big), \tag{2}$$

$$\mathcal{L}(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}}) \;=\; -\frac{1}{N}\sum_{i=1}^{N} \log \operatorname{softmax}(\boldsymbol{z}_i)_{y_i}. \tag{3}$$

All parameters θ are frozen, and gradients on $\boldsymbol{E}$ are masked so that only $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ get updated. Both $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ are initialized from $w_0 \in \mathcal{V}$, which is a single, modality-appropriate token (e.g., "photograph").

Generalization across modalities. Eq. 1 requires only that $\mathcal{M}_\theta$ supports inputs of modality m and that a suitable continuation token $w_m \in \mathcal{V}$ exists. Because $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ live in the same multimodal embedding space, the trained pair is modality-agnostic. It transfers directly to any modality supported by the backbone by swapping Π_m and w_m—including modalities held out during training.

Joint multimodal training. Given training data across multiple modalities $\mathcal{M}_{\text{train}} \subseteq \mathcal{M}$, we optimize a single embedding pair jointly across all of them, rotating through $\mathcal{M}_{\text{train}}$ in round-robin order and accumulating per-modality losses (Eq. 2) into the shared embeddings. Because the representation is shared across modalities (Sec. 3.1), gradient updates from each modality jointly shape a unified real vs. AI decision boundary.

3.2 Self-Verbalization and Plug-in Evaluation

Once $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ are trained, $\mathcal{M}_\theta$ processes both tokens like any other entry in $\mathcal{V}$ and can be queried about them directly; because the backbone is frozen, the resulting descriptions reflect only what the trained embeddings have learned. Following Hewitt et al. [2025b], we elicit two forms of self-verbalization via forced-start decoding.

Global verbalization. A concept-level probe asks $\mathcal{M}_\theta$ to describe <AIGEN> and <REAL> without any input example in context, constraining the assistant turn to begin with a fixed prefix:

> **User:** *The concept* <AIGEN> *is used to classify* ⟨*modality*⟩*. Describe what* <AIGEN> ⟨*modality*⟩ *look (resp. sound, read) like.*
> **Assistant (forced start):** *Ok,* <AIGEN> ⟨*modality*⟩ *are characterized by* ...

Decoding the continuation yields a free-form description of the general features learned by the embedding.

Instance verbalization. An instance-level probe passes a specific input $\boldsymbol{x} \in \mathcal{X}_m$ alongside the query, prompting $\mathcal{M}_\theta$ under the same forced-start setup to explain why <AIGEN> applies (or does not apply) to $\boldsymbol{x}$.

Plug-in evaluation. We validate faithfulness using plug-in evaluation [Hewitt et al., 2025b], replacing $\tau \in \{\texttt{<REAL>}, \texttt{<AIGEN>}\}$ in Π_m with its global verbalization $\nu(\tau)$ and recomputing $s_m(\boldsymbol{x})$ on a held-out test set. While the verbalization need not exactly match the performance of the method with trained tokens, robust performance above chance would indicate that the natural-language description faithfully captures what the embedding has learned.

4 Results

We evaluate whether our method (i) matches modality-specific detectors across text, audio, image, and video (Sec. 4.1); (ii) generalizes across datasets and unseen generators (Sec. 4.2); (iii) transfers across modalities, including to held-out modalities at test time (Sec. 4.3); and (iv) produces faithful natural-language explanations that reflect the features used by the detector (Secs. 4.6–4.7). All detection metrics are reported as AUROC. Unless stated otherwise, we use `Llama-3.1-8B-AI-Neo` for unimodal evaluations and `Gemma-4-E4B-AI-Neo` for multimodal evaluations.

4.1 A two-row embedding edit matches modality-specific SOTA

We compare against state-of-the-art open-source detectors on standard benchmarks across all four modalities: RAID [Dugan et al., 2024] (text), AUDETER [Wang et al., 2025b] (audio), GenImage [Zhu et al., 2023b] (image), and DFE-2024 [Chandra et al., 2025] (video). For each modality, we select the strongest publicly available baselines that disclose their training mixture and train only on the benchmark's training split: DivEye [Basani and Chen, 2025], Binoculars [Hans et al.,

Table 1: **Detection Performance.** The models are tested on the in-distribution dataset (evaluation split) and held-out OOD dataset; AUROC is reported.

Method	ID AUROC	OOD AUROC
Text — RAID → Beemo		
DivEye	0.91	0.57
Binoculars	0.85	0.84
TMR-Detector	1.00	0.77
Llama-3.1-8B-Classifer	0.99	0.72
Gemma-4-E4B-Classifer	0.97	0.71
Llama-3.1-8B-AI-Neo	0.99	0.79
Gemma-4-E4B-AI-Neo	0.96	0.78
Audio — AUDETER → In-the-Wild		
AASIST	1.00	0.79
RawNet2	1.00	0.77
Gemma-4-E4B-AI-Neo	0.99	0.72
Image — GenImage → CIFAKE		
UnivFD	0.56	0.47
AIDE	0.82	0.54
Gemma-4-E4B-AI-Neo	0.93	0.63
Video — DFE-2024 → DeeptraceReward		
SBI	0.76	0.58
CDFA	0.75	0.60
Gemma-4-E4B-AI-Neo	0.83	0.88

Table 2: **Generator Generalization.** DivEye, Binoculars, and Llama-3.1-8B-AI-Neo (LL-AI-Neo) are tested for generator generalization on the task of AI-generated text detection. For each held-out generator g, models train on human samples plus the remaining $G-1$ generators and evaluate on human plus g; AUROC is reported.

Held-out generator	DivEye	Bino	Ll-AI-Neo
MPT	0.90	0.59	0.94
Cohere	0.80	0.93	0.96
Mistral	0.85	0.71	0.96
Cohere Chat	0.84	0.95	0.98
GPT-3	0.91	0.97	0.98
GPT-2	0.96	0.82	0.99
GPT-4	0.82	0.94	1.00
Mistral Chat	0.88	0.93	1.00
MPT Chat	0.87	0.77	1.00
ChatGPT	0.89	0.97	1.00
Llama Chat	0.87	0.97	1.00
Mean	0.87	0.87	0.98
Worst-case	0.80	0.59	0.94

2024], TMR Oxidane [2025], and a custom linear classifier for text; AASIST [Jung et al., 2022] and RawNet2 [Tak et al., 2021] for audio; UnivFD [Ojha et al., 2023] and AIDE [Yan et al., 2024] for image; and SBI [Shiohara and Yamasaki, 2022] and CDFA [Lin et al., 2024] for video.

Tab. 1 (left) reports in-distribution AUROC. Our models saturate on text and audio, matching the strongest baselines (within 0.01 AUROC of TMR-Detector on text, and matching AASIST and RawNet2 on audio), while outperforming all baselines on image ($+0.11$ over AIDE) and video ($+0.07$ over SBI)—despite training only two embedding rows in a frozen backbone.

4.2 Detection holds across datasets and across unseen generators

Cross-dataset OOD. We evaluate each model out of the box on a second benchmark per modality with disjoint generators and content sources: Beemo [Artemova et al., 2025] (text), In-the-Wild [Müller et al., 2022] (audio), CIFAKE [Bird and Lotfi, 2024] (image), and DeeptraceReward [Fu et al., 2025] (video). As shown in Tab. 1 (right), our pair achieves the highest AUROC on image and video under distribution shift, and comes within 0.07 AUROC of the top baseline on text and audio. By contrast, the modality-specific baselines that match our method's in-distribution performance drop sharply out of distribution; for example, DivEye falls by -0.34 on text and AIDE by -0.28 on image.

Within-dataset leave-one-generator-out. We run a leave-one-generator-out protocol on RAID [Dugan et al., 2024], which contains human-written and AI-generated text across $G = 11$ distinct generators (Tab. 2). Our pair attains a mean held-out AUROC of 0.98 (worst-case 0.94), compared to a mean held-out of 0.87 (worst-case 0.59-0.80) for the two strongest text baselines. The learned embeddings, hence, capture features, which generalize across generator families.

4.3 A pair trained on a subset of modalities transfers to held-out modalities

We next test whether $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ transfer across modalities: does training on a strict subset of modalities $\mathcal{M}_{\text{train}} \subsetneq \mathcal{M}$ detect AI-generated content in held-out modalities? Tab. 3 reports AUROC

Table 3: **Cross-Modality Generalization.** AUROC or every choice of $\mathcal{M}_{\text{train}}$ (single-modality, leave-one-out, all four), tested for ability to generalize to modalities unseen during training time.

Train				Test			
Txt	**Aud**	**Img**	**Vid**	**Txt**	**Aud**	**Img**	**Vid**
✓				0.96	0.71	0.57	0.58
	✓			0.52	0.99	0.58	0.41
		✓		0.61	0.77	0.92	0.46
			✓	0.48	0.45	0.58	0.83
	✓	✓	✓	0.73	0.99	0.79	0.78
✓		✓	✓	1.00	0.48	0.81	0.80
✓	✓		✓	1.00	1.00	0.60	0.80
✓	✓	✓		0.56	0.99	0.41	0.47
✓	✓	✓	✓	0.99	0.99	0.68	0.80

Table 4: **Image Detection Robustness.** AUROC on the in-distribution GenImage test split under JPEG compression, bilinear resize-and-restore, and Gaussian blur, contrasted with the baseline.

Method	**Clean**	**JPEG quality**				**Resize-restore**			**Gaussian blur** (σ)		
		90	75	50	25	75%	50%	25%	0.5	1.0	2.0
UnivFD	0.56	0.54	0.54	0.54	0.55	0.58	0.57	0.58	0.55	0.59	0.55
AIDE	0.82	0.82	0.80	0.79	**0.79**	0.82	**0.77**	**0.69**	0.84	**0.79**	**0.69**
Gemma-4-E4B-AI-Neo	**0.93**	**0.92**	**0.87**	**0.79**	0.73	**0.82**	0.74	0.61	**0.91**	0.78	0.66

across all combinations of $\mathcal{M}_{\text{train}} \in \{\text{single modality}\} \cup \{\text{leave-one-out}\} \cup \{\mathcal{M}\}$, evaluated on each modality.

Training on a single modality already transfers to many unseen ones, as most off-diagonal entries are above chance. Leave-one-out training closes most of the remaining gap on the held-out modality when text is included (e.g., $\{$audio, image, video$\} \rightarrow$ text reaches 0.73, above any single-modality transfer to text), whereas adding modalities other than text does not (e.g., $\{$text, image, video$\} \rightarrow$ audio falls below text-only transfer). Our results show that text contributes most of the cross-modal signal to the shared embeddings. Finally, the most balanced performance can be achieved through joint training across all four modalities yields. In that setting, the performance on text, audio, and video all come within 0.03 AUROC of their single-modality counterparts, with a larger drop only on image (0.68 vs. 0.92). Image artifacts hence appear more orthogonal to the shared direction than those in text, audio, or video (App. C.2).

4.4 Detection degrades gracefully under common perturbations

We next test robustness to common perturbations across the in-distribution test sets, comparing our trained pair against the strongest baseline per modality from Sec. 4.1.

Image perturbations. We apply (i) JPEG compression at quality factors $\{90, 75, 50, 25\}$, (ii) bilinear resizing to $\{75\%, 50\%, 25\%\}$ of the original resolution before re-upsampling, and (iii) Gaussian blur with $\sigma \in \{0.5, 1.0, 2.0\}$ (Tab. 4). Strong bilinear resizing affets both our method the the baseline the most. JPEG compression and mild Gaussian blur cause minor drops in performance. Our method start out at 0.11 and 0.37 AUROC point above AIDE and UnivFD on non-erturbed images, with the gap narrowing only under the strongest perturbations.

Text perturbations. Following RAID, we evaluate across five attack categories on the held-out RAID source IDs: (i) paraphrasing, (ii) misspelling, (iii) whitespace insertions, (iv) casing perturbations, and (v) homoglyph substitutions (Tab. 5). `Llama-3.1-8B-AI-Neo` remains nearly flat across all five attacks. The smaller `Gemma-4-E4B-AI-Neo` backbone drops in performance on paraphrasing ($0.96 \rightarrow 0.86$) but holds up across all attempted character-level attacks. Across attack categories, neither text baseline matches our method; the margins hold or widen under perturbation.

Table 5: **Text Detection Robustness.** AUROC on held-out RAID source IDs, evaluated by attack category. Detection-token methods are trained only on the standard random training split.

Method	Clean	Paraphrase	Misspelling	Whitespace	Casing	Homoglyph
DivEye	0.91	0.83	0.89	0.79	0.82	0.93
Binoculars	0.85	0.93	0.84	0.81	0.83	0.81
Llama-3.1-8B-AI-Neo	**0.98**	**0.97**	**0.98**	**0.99**	**0.99**	**1.00**
Gemma-4-E4B-AI-Neo	0.96	0.86	0.96	0.99	0.96	0.99

Table 6: **Comparison Against MLLM Detectors.** Out-of-distribution AUROC across modalities, using each method's released checkpoint without further adaptation. N/A marks unsupported modalities.

Method	Txt	Aud	Img	Vid	Mean
SIDA	–	–	0.17	–	0.17
FakeShield	–	–	0.61	–	0.61
Ll-AI-Neo	0.79	–	–	–	0.79
Ge-AI-Neo	0.78	0.72	0.63	0.88	0.75

Table 7: **Plug-in Evaluation.** Pairwise AI-vs-real classification accuracy on held-out text pairs, when the trained <AIGEN> embedding is replaced with its verbalization $\nu(\texttt{<AIGEN>})$ and handed to an off-the-shelf MLLM evaluator that has never seen our trained tokens.

Evaluator	Baseline	$+\,\nu(\texttt{<AIGEN>})$	Δ
Gemini 3.1 Pro	0.47	0.72	$+0.25$
Gemini 3 Flash	0.48	0.75	$+0.27$

4.5 Outperforming MLLM-based detectors out of distribution

Recent work also fine-tunes MLLMs end-to-end for AI-generated content detection using human-annotated rationales [Zhang et al., 2024, Li et al., 2025b, Xu et al., 2024, Huang et al., 2025b, Hondru et al., 2026]. Because each method trains on its own dataset and cannot be retrained, we evaluate their released checkpoints on the unseen OOD benchmarks from Sec. 4.2: Beemo, In-the-Wild, CIFAKE, and DeeptraceReward.

Tab. 6 compares our trained pair against SIDA [Huang et al., 2025b] and FakeShield [Xu et al., 2024]. Because both MLLM baselines only support images, direct comparison is limited to CIFAKE, where our `Gemma-4-E4B-AI-Neo` pair reaches 0.63 AUROC, outperforming FakeShield ($+0.02$) and SIDA ($+0.46$). Across the remaining three modalities, the same pair reaches 0.78 on text, 0.72 on audio, and 0.88 on video (0.75 mean across all four modalities), while `Llama-3.1-8B-AI-Neo` reaches 0.79 on text—all while training orders of magnitude fewer parameters and requiring no human-annotated rationales.

4.6 Self-verbalizations are faithful

Verbalized words score consistently with the detector. We first prompt the trained model for lists of words characteristic of <AIGEN> and <REAL> text, denoted W_{AI} and W_{human}, and score each word w in isolation via $s_{\text{text}}(w)$ (Eq. 1). If the verbalization is faithful, words the model identifies as AI-likely should also score higher under the detector: $\mathbb{E}_{w \in W_{\text{AI}}}[s_{\text{text}}(w)] > \mathbb{E}_{w \in W_{\text{human}}}[s_{\text{text}}(w)]$. For comparison, we directly ask the pre-trained backbone (without the tokens from our method) for AI- and human-likely words. Across 480 samples per condition (Tab. 10), the trained tokens yield a $+0.91$ log-probability gap in the expected direction, whereas the baseline yields -0.08, that is to say, in the wrong direction (one-sided Welch $p \approx 0$ for both). Equivalently, the detector classifies 54.2% of words in $W_{\langle\texttt{AIGEN}\rangle}$ as AI-generated, compared to 27.6% for the baseline.

Plug-in evaluation. Following Hewitt et al. [2025b], we also run plug-in evaluation: we replace $\tau \in \{\texttt{<REAL>}, \texttt{<AIGEN>}\}$ in Π_{text} with its model-generated verbalization $\nu(\tau)$ and pass the resulting plain-text prompt to an off-the-shelf frontier MLLM to identify which input in a held-out pair $(\boldsymbol{x}_{\text{real}}, \boldsymbol{x}_{\text{AI}})$ is AI-generated. As shown in Tab. 7, plugging in the verbalization improves pairwise accuracy of both tested models from approximately chance to well above chance: $0.47 \rightarrow 0.72$ for Gemini 3.1 Pro and $0.48 \rightarrow 0.75$ for Gemini 3 Flash. An unrelated MLLM can hence use the free-form description of <AIGEN> to detect AI text, showing that the verbalization reflects what the embedding actually learned.

Text

<AIGEN>

1. Brevity: Short sentences and paragraphs, often with a focus on conveying a single idea or message. **2. Conciseness:** Avoidance of unnecessary words, phrases, or descriptions, with a focus on clarity and efficiency. **3. Directness:** Straightforward and unambiguous language, with minimal use of metaphors, similes, or other literary devices. **4. Objectivity:** A neutral or detached tone, often lacking emotional appeals or personal opinions. **5. Clarity:** Easy-to-understand language, with a focus on ...

<REAL>

1. Complex syntax: The text may have long, convoluted sentences with multiple clauses and phrases, making it hard to follow. **2. Technical jargon:** The text may use specialized vocabulary or technical terms that are not familiar to the average reader. **3. Abstract concepts:** The text may deal with abstract ideas or complex concepts that require a high level of cognitive processing to understand. **4. Lack of clarity:** The text may be written in a way that is ambiguous or unclear ...

Image

<AIGEN>

1. Geometric and Sharp: often features clean lines and precise angles. The shapes are rarely organic or blurry, instead leaning towards a structured, geometric aesthetic. **2. Minimalist and Sparse:** images tend to have a lot of negative space. The composition is often simple, with very few elements competing for attention. **3. Monochromatic or Limited Palette:** often uses a restricted color scheme, frequently relying on black, white, and shades of grey. ...

<REAL>

1. Abstract and Geometric: The images often feature intricate geometric patterns, abstract shapes, and complex lines, giving them a highly stylized and non-representational look. **2. High Contrast and Bold Colors:** There's a frequent use of bold, vibrant colors set against stark backgrounds or high-contrast combinations, creating a visually striking and energetic feel. **3. Digital and Glitchy:** Many images incorporate digital aesthetics, such as pixelation, glitch effects, and digital ...

Figure 2: **Global self-verbalizations of <REAL> and <AIGEN>.** Greedy completions of *"*<TOK> ⟨*modality*⟩ *are characterized by…"* for text (left) and image (right), decoded from a frozen MLLM whose only learnable parameters are the two embedding rows.

Gemini vs. GPT-4o

<AIGEN_GEMINI>

1. **Raw or unformatted**: The text is likely to be unedited, without proper punctuation, grammar, or spelling corrections. It may also be written in a plain or simple font without any formatting, such as headings, paragraphs, or bullet points.
2. **Unedited or uncensored**: The text may contain profanity, slang, or other forms of language that would not be suitable for general audiences...

Claude Sonnet vs. GPT-4o

<AIGEN_CLAUDE>

written in a clear, concise, and straightforward manner.

It is implied that the text in question is unadorned, possibly lacking in flair or eloquence, and may not be written in an overly complex or ornate style. The connotation is of a straightforward and factual communication, perhaps with a sense of simplicity and directness.

GPT-4o vs. Gemini

<AIGEN_GPT4o>

- Is likely to be informal or conversational
- Is possibly written for a younger audience
- May be written in a casual tone
- Possibly includes colloquial expressions or slang
- Could be used in a chat or messaging application

Figure 3: **Per-generator self-verbalizations on EvoBench.** Greedy completions of <AIGEN$_g$> under three pairwise contrastive splits, with a separate token trained to discriminate between generators

4.7 What the neologisms describe

Finally, we qualitatively examine three forms of self-verbalization: *global* (concept-level, no input), *per-generator* (one neologism trained per generator), and *instance* (datapoint-conditioned).

Global verbalizations identify modality-appropriate regularities. Fig. 2 shows representative completions of *"*<AIGEN> ⟨*modality*⟩ *are characterized by …"* across all four modalities. Despite never being trained on explanations, the frozen MLLM points to modality-specific patterns, such as conciseness and detached tone for AI text, or limited color palettes and uniform composition for AI images.

Per-generator neologisms surface generator-specific tells. On EvoBench [Yu et al., 2025], we train a separate <AIGEN$_g$> embedding for each generator g using the same protocol as Sec. 4.2. Fig. 3 shows greedy verbalizations elicited by contrasting each <AIGEN$_g$> against samples from another generator; the resulting descriptions capture known stylistic differences across generators.

Instance verbalizations point at features in the specific input. Because $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ live inside an MLLM, the same tokens can be conditioned on a specific input $\boldsymbol{x}$ to explain why it is (or

is not) <AIGEN> (Sec. 3.2). Fig. 4 (App. C.4) pairs representative inputs with their predicted label $\hat{y}(\boldsymbol{x})$, score $s_m(\boldsymbol{x})$, and instance verbalization.

Even though the backbone is frozen and only two embedding rows are updated, Figs. 2-4 show that the trained tokens capture interpretable, content-level features. We include additional examples, failure modes, and per-generator comparisons on EvoBench in App. C.4.

5 Discussion

To our knowledge, our method is the first AI-generated content detector to span text, image, audio, and video through a single shared representation. Prior cross-modality methods either fuse synchronous audio-visual streams [Khalid et al., 2021, Ilyas et al., 2023, Bohacek and Farid, 2024b] or stack modality-specific heads on a shared encoder [Huang et al., 2025a], and neither approach transfers to held-out modalities. Our method also enables interpretability out of the box, whereas prior explainable detectors require human-annotated rationales [Zhang et al., 2024, Liu et al., 2024a, Huang et al., 2025b, Hondru et al., 2026]. We can query the trained tokens in natural language, whether globally, per generator, or conditioned on a specific input, to read out what the detector encodes and validate it via plug-in evaluation. Because only $2\,d$ parameters get updated and verbalization is built into the backbone, our method extends directly to new modalities (e.g., 3D, motion capture), per-generator diagnostics, and other binary classification tasks whose decision boundary can be described in natural language.

Limitations. Though the proposed detection method retains strong performance across adversarial attacks, detection success varies with the generator model. As a result, there may be generator models trained in the future that degrade detection performance. More broadly, any AI-generated content detector may produce both false positives and false negatives, particularly under distribution shift. We therefore do not recommend deployment in high-stakes settings without human oversight and additional testing in the target domain. Misusing AI-detection tools can degrade trust in reliable sources including media organizations, students, and institutions, while incorrectly granting authenticity to deceptive or synthetic content that evades detection. With regard to natural-language descriptions, we make a best-effort attempt to evaluate faithfulness. However, true counterfactual tests are not possible in our setting as counterfactuals would themselves be either AI- or human-written, and thus confound evaluation. Our hope is that verbalizations provide insight into the mechanisms behind AI-detection, but they should not be interpreted as holistic causal explanations of the model.

Acknowledgments and Disclosure of Funding

The authors wish to thank John Hewitt for helpful suggestions. This project did not receive external funding. The authors report no competing interests.

References

S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

E. Artemova, J. S. Lucas, S. Venkatraman, J.-Y. Lee, S. Tilga, A. Uchendu, and V. Mikhailov. Beemo: Benchmark of expert-edited machine-generated outputs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6992–7018, 2025.

S. Barrington, M. Bohacek, and H. Farid. The DeepSpeak dataset. *arXiv preprint arXiv:2408.05366*, 2024.

S. Barrington, E. A. Cooper, and H. Farid. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15(1):11004, 2025.

A. R. Basani and P.-Y. Chen. Diversity boosts AI-generated text detection. *arXiv preprint arXiv:2509.18880*, 2025.

J. J. Bird and A. Lotfi. CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024.

M. Bohacek and H. Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022.

M. Bohacek and H. Farid. Human action CLIPS: Detecting AI-generated human motion. *arXiv preprint arXiv:2412.00526*, 2024a.

M. Bohacek and H. Farid. Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4323, 2024b.

L. Cao. A practical synthesis of detecting AI-generated textual, visual, and audio content. *arXiv preprint arXiv:2504.02898*, 2025.

N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, et al. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

D. Cooke, A. Edwards, S. Barkoff, and K. Kelly. As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760*, 2024.

D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024.

J. Dolezal, S. Alam, M. Graham, and M. Bohacek. The impact of AI-generated text on the internet. *arXiv preprint arXiv:2604.26965*, 2026.

L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, and C. Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, 2024.

X. Fu, S. Liu, Y. Xu, P. Lu, G. Hu, T. Yang, T. Anantasagar, C. Shen, Y. Mao, Y. Liu, et al. Learning human-perceived fakeness in AI-generated videos via multimodal LLMs. *arXiv preprint arXiv:2509.22646*, 2025.

R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

Google DeepMind. Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`, 2026.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.

J. Hewitt, R. Geirhos, and B. Kim. We can't understand AI using our existing vocabulary. *arXiv preprint arXiv:2502.07586*, 2025a.

J. Hewitt, O. Tafjord, R. Geirhos, and B. Kim. Neologism learning for controllability and self-verbalization. *arXiv preprint arXiv:2510.08506*, 2025b.

V. Hondru, E. Hogea, D. Onchis, and R. T. Ionescu. ExDDV: A new dataset for explainable deepfake detection in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4273–4284, 2026.

L. Huang, Z. Zhang, Y. Zhang, X. Zhou, and S. Wang. RU-AI: A large multimodal dataset for machine-generated content detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 733–736, 2025a.

Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, and G. Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025b.

H. Ilyas, A. Javed, and K. M. Malik. AVFakeNet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023.

S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, and S. Lyu. Can ChatGPT detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024.

J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. 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 (ICASSP)*, pages 6367–6371. IEEE, 2022.

H. Khalid, S. Tariq, M. Kim, and S. S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

S. A. Khan and D.-T. Dang-Nguyen. CLIPping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

M. Li, J. He, Y. Yu, J. Xu, S. Lei, Y. Teng, Y. Wang, and X. Wang. UniMark: Artificial intelligence generated content identification toolkit. *arXiv preprint arXiv:2512.12324*, 2025a.

Y. Li, X. Liu, X. Wang, B. S. Lee, S. Wang, A. Rocha, and W. Lin. FakeBench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security*, 2025b.

Y. Lin, W. Song, B. Li, Y. Li, J. Ni, H. Chen, and Q. Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European conference on computer vision*, pages 104–122. Springer, 2024.

J. Liu, F. Zhang, J. Zhu, E. Sun, Q. Zhang, and Z.-J. Zha. ForgeryGPT: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024a.

P. Liu, Q. Tao, and J. T. Zhou. Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges. *arXiv preprint arXiv:2406.06965*, 2024b.

V. Livernoche, A. Arodi, A. Musulan, Z. Yang, A. Salvail, G. M. Caron, J.-F. Godbout, and R. Rabbany. OpenFake: An open dataset and platform toward real-world deepfake detection. *arXiv preprint arXiv:2509.09495*, 2025.

N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24480–24489, 2023.

Oxidane. Tmr: Target mining roberta for ai text detection, 2025. URL `https://huggingface.co/Oxidane/tmr-ai-text-detector`.

S. Park, V. Ramamurthi, and O. Terry. Neologism learning as a parameter-efficient alternative to fine-tuning for model steering. *arXiv preprint arXiv:2512.18551*, 2025.

L. Pellegrini, D. Cozzolino, S. Pandolfini, D. Maltoni, M. Ferrara, L. Verdoliva, M. Prati, and M. Ramilli. AI-GenBench: A new ongoing benchmark for AI-generated image detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2025.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Y. Shi, Y. Gao, Y. Lai, H. Wang, J. Feng, L. He, J. Wan, C. Chen, Z. Yu, and X. Cao. SHIELD: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Visual Intelligence*, 3(1):9, 2025.

K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher. End-to-end anti-spoofing with RawNet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

S. Tariq, S. S. Woo, P. Singh, I. Irmalasari, S. Gupta, and D. Gupta. From prediction to explanation: Multimodal, explainable, and interactive deepfake detection framework for non-expert users. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11716–11725, 2025.

K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos, and V. Mezaris. Towards quantitative evaluation of explainable AI methods for deepfake detection. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, pages 37–45, 2024.

L. Wang, J. Zhao, X. Zhang, X. Guo, Y. Yuan, T. Zhang, J. Chu, Y. Jiang, X. Yang, L. Jin, et al. ERF-BA-TFD+: a multimodal model for audio-visual deepfake detection. *Vicinagearth*, 2(1): 1–12, 2025a.

Q. Wang, H. Huang, G. Pang, S. Erfani, and C. Leckie. AUDETER: A large-scale dataset for deepfake audio detection in open worlds. *arXiv preprint arXiv:2509.04345*, 2025b.

H. Wen, T. Li, Z. Huang, Y. He, and G. Cheng. BusterX++: Towards unified cross-modal ai-generated content detection and explanation with MLLM. *arXiv preprint arXiv:2507.14632*, 2025.

Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.

S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie. A sanity check for AI-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

M. Yang, S. Deng, W. T. Doan, K. Wang, T. Yang, H. Singh, and Y. Tian. Explainable AI-generated image detection RewardBench. *arXiv preprint arXiv:2511.12363*, 2025.

W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren. AVoiD-DF: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

X. Yu, Y. Yu, D. Liu, K. Chen, W. Zhang, N. Yu, and J. Shao. EvoBench: Towards real-world llm-generated text detection benchmarking for evolving large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14605–14620, 2025.

Y. Zhang, B. Colman, X. Guo, A. Shahriyari, and G. Bharaj. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pages 399–415. Springer, 2024.

W. Zhong, Y. Xiao, M. Xu, and X. Cheng. VMID: A multimodal fusion LLM framework for detecting and identifying misinformation of short videos. *arXiv preprint arXiv:2411.10032*, 2024.

B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al. Language-Bind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023a.

M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. GenImage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023b.

A Implementation Details

A.1 Hyperparameter Selection

Hyperparameters for each backbone are selected using a held-out split of $1,000$ examples from RAID. Llama-based models are trained with a learning rate of 10^{-4} and Gemma-4-E4B-it is trained with a learning rate of 10^{-3}. The batch size applied to both is 32 and the optimizer is AdamW with a 20-step warmup followed by cosine decay to 0.01x the original learning rate. We train $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ using the APO-up objective from Hewitt et al. [2025b]:

$$\mathcal{L}(y_c, y_r) = -\log \sigma\left(\beta \log \frac{p_\theta(y_c)}{p_\theta(y_r)} + \beta \log \frac{p_{\theta_0}(y_c)}{p_{\theta_0}(y_r)}\right) - \alpha_{\text{APO}} \log \sigma\left(\beta \log \frac{p_\theta(y_c)}{p_{\theta_0}(y_c)}\right), \quad (4)$$

where p_θ is the model with the trainable embeddings, p_{θ_0} is the frozen reference, y_c is an AI-generated response, and y_r is a human-written response (or vice versa for the <REAL> token). Unlike the original implementation, which uses per-response instructions, we condition on the fixed prompts defined in the main text. We set $\beta = 0.1$ and $\alpha_{\text{APO}} = 1.0$.

A.2 Gemma 4 Adaptations

The auxiliary per-layer embedding table. Most decoder-only LMs use a single learnable embedding matrix $\boldsymbol{E} \in \mathbb{R}^{d \times |\mathcal{V}|}$ shared between the input and the output head. Gemma 4 E-series adds a second token-indexed table of per-layer embeddings $\boldsymbol{E}^{\text{PLE}} \in \mathbb{R}^{h \times |\mathcal{V}|}$, which are added to each transformer layer alongside the residual stream. Vocabulary expansion (Sec. 2), hence, requires resizing both tables. We extend $\boldsymbol{E}$ with two trainable rows $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ as described in the main text, and we extend $\boldsymbol{E}^{\text{PLE}}$ with two rows initialized to the vocabulary mean $\frac{1}{|\mathcal{V}|}\sum_v \boldsymbol{E}^{\text{PLE}}_v$.

Why $\boldsymbol{E}^{\text{PLE}}$ is held frozen. Only $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ get updated, while the new rows in $\boldsymbol{E}^{\text{PLE}}$ remain frozen. Updating both tables splits the learned representation across $\boldsymbol{E}$ and $\boldsymbol{E}^{\text{PLE}}$, which leaves $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ close to their initialization. Because the LM head is tied to $\boldsymbol{E}$, self-verbalization (Sec. 3.2) then only outputs near-synonyms of the initialization word. Freezing $\boldsymbol{E}^{\text{PLE}}$ keeps the number of trainable parameters at $2\,d_{\text{model}}$ (Sec. 3.1).

Gradient masking. With tied input and output embeddings, gradients from the LM head flow into all rows of $\boldsymbol{E}$. Masking the gradient on $\boldsymbol{E}$ so that all other rows are zeroed out, alongside freezing $\boldsymbol{E}^{\text{PLE}}$, ensures that only $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ get updated.

Per-modality prompt templates. The image and text templates in Sec. 3.1 place τ at two positions. Audio and video use single-position templates: $\Pi_{\text{audio}}(\tau, \boldsymbol{x}) = \langle \text{audio}(\boldsymbol{x}) \rangle$ "This is a τ" and $\Pi_{\text{video}}(\tau, \boldsymbol{x}) = \langle \text{frames}(\boldsymbol{x}) \rangle$ "These frames are from a τ video". We sample eight uniformly spaced frames per video clip and pass them to the multimodal processor as image placeholders. Audio waveforms are resampled to 16 kHz and clipped to 25 s before processing.

Warm-start tokens. Both $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ are initialized from a single, modality-appropriate token (Sec. 3.1): `photograph` (image), `audio` (audio), `writing` (text), and `video` (video). For video, the initialization token matches the continuation token (Sec. 3.1), which did not affect training in our experiments, though a non-overlapping token (e.g., `footage`) can also be used.

A.3 Continuation-Token Selection

Evaluating $s_m(\boldsymbol{x})$ in Eq. 1 at a fixed continuation token $w_m \in \mathcal{V}$ ensures that the prediction comes down to the difference between $\boldsymbol{E}_{\texttt{<REAL>}}$ and $\boldsymbol{E}_{\texttt{<AIGEN>}}$. Any single token $w_m \in \mathcal{V}$ that forms a valid completion of the prompt template works, whereas multi-token continuations would require aggregating log-probabilities across positions.

At runtime, we select w_m as the first candidate from an ordered per-modality list that maps to a single non-unknown token ID:

$$w_{\text{image}} : \texttt{Image},\ \texttt{image},\ \texttt{photo},\ \texttt{photograph}$$
$$w_{\text{audio}} : \texttt{Audio},\ \texttt{audio},\ \texttt{recording},\ \texttt{voice},\ \texttt{sound},\ \texttt{clip}$$
$$w_{\text{video}} : \texttt{video},\ \texttt{Video},\ \texttt{recording},\ \texttt{footage},\ \texttt{clip}$$
$$w_{\text{text}} : \texttt{text},\ \texttt{Text},\ \texttt{writing},\ \texttt{passage},\ \texttt{content}$$

For the Gemma 4 tokenizer, the first candidate in each list resolves to a single token ID and is used across all experiments. For tokenizers that split it into multiple tokens, the other candidates are included as backup.

A.4 Baselines

We compare our method against modality-specific and multimodal baselines across all four modalities.

Text. We evaluate DivEye[1] [Basani and Chen, 2025], Binoculars[2] [Hans et al., 2024], TMR-Detector [Oxidane, 2025], and linear classifiers on Llama-8B and Gemma-4-E4B-it. DivEye extracts 10 surprisal features from GPT-2 and trains an XGBoost classifier with default hyperparameters. These features consist of the mean, standard deviation, variance, skewness, and kurtosis of per-token surprisal, the mean and standard deviation of first-order surprisal differences, and the variance, entropy, and auto-correlation of second-order log-likelihood differences. Binoculars is used out of the box to compute the cross-perplexity ratio between Falcon-7B and Falcon-7B-Instruct, where lower scores indicate AI-generated text. TMR-Detector is a RoBERTa classifier trained on RAID with hard-example mining and achieves the highest AUROC on the private RAID test set. We also train linear classifiers directly on the last-layer activations of Llama-8B and Gemma-4-E4B-it.

Audio. We compare against two end-to-end audio anti-spoofing models, AASIST [Jung et al., 2022] and RawNet2 [Tak et al., 2021], using the official implementation[3]. AASIST uses spectro-temporal graph attention networks. RawNet2 operates directly on raw waveforms.

Image. We compare against UnivFD[4] [Ojha et al., 2023] and AIDE[5] [Yan et al., 2024]. UnivFD uses a frozen `ViT-L/14` CLIP feature extractor and a trainable classification head, which we train on GenImage. For AIDE, which combines `ResNet-50` and `ConvNeXt` backbones for feature extraction, we use the defaults from the official repository.

Video. We compare against SBI[6] [Shiohara and Yamasaki, 2022] and CDFA [Lin et al., 2024]. SBI trains a frame-level `EfficientNet-B4` on pseudo-fakes, focused exclusively on videos with faces. Because our training and evaluation distributions go beyond faces, we replace blending based on facial landmarks with a landmark-free version. Specifically, we blend a rectangular region using a soft elliptical alpha mask of a randomly selected size, and optionally apply JPEG compression. CDFA trains a frame-level Xception backbone using JPEG re-encoding and random DCT-band masking. For both methods, the video-level score is the average across eight frames sampled uniformly from the video.

Multimodal. We evaluate pretrained checkpoints of SIDA[7] (`SIDA-13B`) [Huang et al., 2025b] and FakeShield[8] (`FakeShield-v1-22B`) [Xu et al., 2024] out of the box. SIDA combines LLaVA, LISA, and SAM, and is trained in two stages on SID_Set (real, tampered, and synthetic images). The first stage classifies inputs via a `[CLS]` token, and the second stage segments tampered regions while generating natural-language explanations. FakeShield uses two modules: a detection module

[1] https://github.com/IBM/diveye/tree/main
[2] https://github.com/ahans30/Binoculars
[3] https://github.com/clovaai/aasist
[4] https://github.com/WisconsinAIVision/UniversalFakeDetect
[5] https://github.com/shilinyan99/AIDE
[6] https://github.com/mapooon/SelfBlendedImages
[7] https://huggingface.co/saberzl/SIDA-13B
[8] https://huggingface.co/zhipeixu/fakeshield-v1-22b

that classifies forgery types using domain tags, and a localization module that predicts pixel-level tampering masks alongside natural-language explanations.

A.5 Datasets

We evaluate our method across text, audio, image, and video using an in-distribution (ID) dataset (split into training and evaluation sets) and a held-out out-of-distribution (OOD) dataset for each modality.

Text. We use RAID[9] [Dugan et al., 2024] as the ID dataset, training on $20,000$ samples ($10,000$ real/AI pairs) and evaluating on $2,000$ held-out samples. For OOD evaluation, we use $2,000$ samples ($1,000$ human/LLM response pairs to shared prompts) from Beemo[10] [Artemova et al., 2025].

Audio. We construct the ID training set of $2,400$ samples ($1,200$ real/AI pairs) by pairing real speech from In-The-Wild (ITW)[11] [Müller et al., 2022] (37.9 hours across 58 speakers) with synthetic speech from the `audiobook-tts-cosyvoice` subset of AUDETER[12] [Wang et al., 2025b]. For the OOD evaluation set (300 samples, 150 pairs), we pair the remaining held-out real clips from ITW with ITW's native deepfakes.

Image. We use a 224×224 JPEG mirror of GenImage[13] [Zhu et al., 2023b] (70/15/15 train/val/test split) as the ID dataset, sampling $1,000$ images (500 pairs) for training and 100 (50 pairs) for evaluation. For OOD evaluation, we sample 100 images (50 pairs) from CIFAKE[14] [Bird and Lotfi, 2024], which pairs CIFAR-10 images with Stable Diffusion v1.4 generations to test generalization across both resolution and generator shifts.

Video. We use Deepfake-Eval-2024 (DFE-2024)[15] [Chandra et al., 2025] as the ID dataset, which contains $1,221$ training and 815 test videos of in-the-wild social media deepfakes across 88 websites and 52 languages. We sample $1,000$ videos (500 pairs) for training and 100 (50 pairs) for evaluation. For OOD evaluation, we sample 100 videos (50 pairs) from DeeptraceReward[16] [Fu et al., 2025], a dataset of $6,636$ clips ($3,318$ pairs) spanning seven text-to-video generators.

A.6 Compute Resources

All experiments were run on two NVIDIA GH200 GPUs. Because only $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ get updated while the backbone remains frozen, individual training runs took 2-10 hours depending on the modality, backbone, and dataset size. Total compute across all training, cross-modal transfer, and perturbation experiments was approximately four wall-clock days.

[9] https://huggingface.co/datasets/liamdugan/raid
[10] https://huggingface.co/datasets/toloka/beemo
[11] https://huggingface.co/datasets/mueller91/In-The-Wild
[12] https://huggingface.co/datasets/wqz995/AUDETER
[13] https://huggingface.co/datasets/Qwerty0193/genimage-processed
[14] https://huggingface.co/datasets/dragonintelligence/CIFAKE-image-dataset
[15] https://huggingface.co/datasets/nuriachandra/Deepfake-Eval-2024
[16] https://huggingface.co/datasets/DeeptraceReward/RewardData

B Self-Verbalization Probes

This appendix documents the prompts and decoding protocol used to elicit self-verbalizations of $\boldsymbol{E}_{\texttt{<REAL>}}$ and $\boldsymbol{E}_{\texttt{<AIGEN>}}$ in Sec. 3.2 and Sec. 4.7.

B.1 Global Verbalization

Image.

> **User:** <AIGEN> *is a concept used to classify images. Before you answer, describe what* <AIGEN> *images look like. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok,* <AIGEN> *images are visually characterized by . . .*
>
> **User:** *List 5 visual properties or descriptive phrases for the concept* <AIGEN>*. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok, here are 5 visual properties of* <AIGEN> *images: . . .*
>
> **User:** *Describe an image that perfectly embodies the concept* <AIGEN>*. What does it look like?*
> **Assistant (forced start):** *An image that perfectly embodies* <AIGEN> *would . . .*

Video.

> **User:** <AIGEN> *is a concept used to classify videos. Before you answer, describe what* <AIGEN> *videos look like. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok,* <AIGEN> *videos are visually characterized by . . .*
>
> **User:** *List 5 visual or temporal properties or descriptive phrases for the concept* <AIGEN>*. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok, here are 5 visual/temporal properties of* <AIGEN> *videos: . . .*
>
> **User:** *Describe a video that perfectly embodies the concept* <AIGEN>*. What does it look like?*
> **Assistant (forced start):** *A video that perfectly embodies* <AIGEN> *would . . .*

Audio.

> **User:** <AIGEN> *is a concept used to classify audio. Before you answer, describe what* <AIGEN> *audio sounds like. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok,* <AIGEN> *audio recordings are acoustically characterized by . . .*
>
> **User:** *List 5 acoustic properties or descriptive phrases for the concept* <AIGEN>*. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok, here are 5 acoustic properties of* <AIGEN> *audio: . . .*
>
> **User:** *Describe an audio recording that perfectly embodies the concept* <AIGEN>*. What does it sound like?*
> **Assistant (forced start):** *An audio recording that perfectly embodies* <AIGEN> *would . . .*

Text.

> **User:** <AIGEN> *is a concept used to classify written text. Before you answer, describe what* <AIGEN> *text reads like. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok,* <AIGEN> *text is linguistically characterized by . . .*
>
> **User:** *List 5 linguistic or stylistic properties or descriptive phrases for the concept* <AIGEN>*. I am trying to understand what* <AIGEN> *means to you.*
> **Assistant (forced start):** *Ok, here are 5 linguistic properties of* <AIGEN> *text: . . .*
>
> **User:** *Describe a piece of text that perfectly embodies the concept* <AIGEN>*. What does it read like?*
> **Assistant (forced start):** *A piece of text that perfectly embodies* <AIGEN> *would . . .*

B.2 Likely-Words Verbalization

User: `Write words that are more common in <AIGEN> text as opposed to <HUMAN> text. One per line, no explanation.`

Assistant prefill: `Words more common in <AIGEN> text as opposed to <HUMAN> text:\n1.`

For the human direction, we swap `<AIGEN>` and `<HUMAN>`. For the natural-language baseline, we replace them with "`AI-generated`" and "`human-written`".

B.3 Instance Verbalization

Image.

User: ⟨image⟩ *The concept* `<AIGEN>` *is being used to classify this specific image. Describe precisely what you observe in this image — textures, edges, lighting, noise patterns, sharpness, skin detail, background coherence, or any subtle visual artifacts — that are most relevant to the concept* `<AIGEN>`*. Be concrete and specific to what you actually see here, not generalities.*
Assistant (forced start): *In this specific image, the visual characteristics most relevant to* `<AIGEN>` *are:* . . .

User: ⟨image⟩ *You are evaluating this image under the concept* `<AIGEN>`*. List 3 precise visual cues you observe in this image — such as sharpness, color distribution, facial texture, reflection consistency, background blending, edge aliasing, or unusual smoothness — that align with or contradict* `<AIGEN>`*.*
Assistant (forced start): *Here are 3 specific visual cues I observe in this image that relate to* `<AIGEN>`*:* . . .

User: ⟨image⟩ *Imagine explaining to a skeptical human why THIS specific image should be labeled* `<AIGEN>` *versus the alternative. What do you see in this image that supports or undermines the concept* `<AIGEN>`*? Be honest about any evidence on both sides.*
Assistant (forced start): *Looking at this specific image,* `<AIGEN>` *applies (or does not apply) because* . . .

Video.

User: ⟨frames⟩ *The concept* `<AIGEN>` *is being used to classify this specific video. Describe precisely what you observe across these frames — textures, edges, lighting, motion consistency, facial detail, background stability, temporal coherence, or any subtle visual or temporal artifacts — that are most relevant to the concept* `<AIGEN>`*. Be concrete and specific to what you actually see here, not generalities.*
Assistant (forced start): *In these specific frames, the visual and temporal characteristics most relevant to* `<AIGEN>` *are:* . . .

User: ⟨frames⟩ *You are evaluating this video under the concept* `<AIGEN>`*. List 3 precise visual or temporal cues you observe across these frames — such as sharpness, facial texture, motion smoothness, edge aliasing, frame-to-frame consistency, background coherence, or unusual artifacts — that align with or contradict* `<AIGEN>`*.*
Assistant (forced start): *Here are 3 specific cues I observe in these frames that relate to* `<AIGEN>`*:* . . .

User: ⟨frames⟩ *Imagine explaining to a skeptical human why THIS specific video should be labeled* `<AIGEN>` *versus the alternative. What do you see across these frames that supports or undermines the concept* `<AIGEN>`*? Be honest about any evidence on both sides.*
Assistant (forced start): *Looking at these specific frames,* `<AIGEN>` *applies (or does not apply) because* . . .

Audio.

User: ⟨audio⟩ *The concept* `<AIGEN>` *is being used to classify this specific audio clip. Describe precisely what you hear in this clip — prosody, intonation, breath sounds, formant quality, naturalness, emotional consistency, background noise, or any subtle acoustic or*

spectral artifacts — that are most relevant to the concept <AIGEN>*. Be concrete and specific to what you actually hear here, not generalities.*
Assistant (forced start): *In this specific audio clip, the acoustic characteristics most relevant to* <AIGEN> *are: . . .*

User: ⟨audio⟩ *You are evaluating this audio clip under the concept* <AIGEN>*. List 3 precise acoustic cues you hear in this clip — such as prosody, formant clarity, breath naturalness, spectral artifacts, background ambience, or unusual smoothness — that align with or contradict* <AIGEN>*.*
Assistant (forced start): *Here are 3 specific acoustic cues I hear in this clip that relate to* <AIGEN>*: . . .*

User: ⟨audio⟩ *Imagine explaining to a skeptical human why THIS specific audio clip should be labeled* <AIGEN> *versus the alternative. What do you hear in this clip that supports or undermines the concept* <AIGEN>*? Be honest about any evidence on both sides.*
Assistant (forced start): *Listening to this specific audio clip,* <AIGEN> *applies (or does not apply) because . . .*

Text.

User: *The concept* <AIGEN> *is being used to classify this specific piece of text: "⟨excerpt⟩". Describe precisely what you notice in this text — sentence structure, vocabulary diversity, coherence, repetitiveness, register, idiom use, or any subtle stylistic markers — that are most relevant to the concept* <AIGEN>*. Be concrete and specific to what you actually read here, not generalities.*
Assistant (forced start): *In this specific text, the linguistic characteristics most relevant to* <AIGEN> *are: . . .*

User: *You are evaluating this text under the concept* <AIGEN>*: "⟨excerpt⟩". List 3 precise linguistic cues you observe in this text — such as sentence rhythm, lexical variety, transitions, repetition patterns, hedging, or unusual fluency — that align with or contradict* <AIGEN>*.*
Assistant (forced start): *Here are 3 specific linguistic cues I observe in this text that relate to* <AIGEN>*: . . .*

User: *Imagine explaining to a skeptical human why THIS specific text should be labeled* <AIGEN> *versus the alternative: "⟨excerpt⟩". What do you read in this text that supports or undermines the concept* <AIGEN>*? Be honest about any evidence on both sides.*
Assistant (forced start): *Looking at this specific text,* <AIGEN> *applies (or does not apply) because . . .*

User: *The learned marker* <AIGEN> *should be referred to in your answer only as "Type-A text". Imagine explaining to a skeptical human why THIS specific text should be labeled Type-A text rather than not Type-A text: "⟨excerpt⟩". Treat "Type-A" as an opaque label; do not translate it into another concept name. What do you read in this text that supports or undermines the Type-A label? Be honest about any evidence on both sides.*
Assistant (forced start): *Looking at this specific text, the Type-A label applies (or does not apply) because . . .*

C Additional Results

C.1 Robustness to Prompting Variations

We evaluate whether neologism training is sensitive to the prompt used during training. Using the same RAID split as reported in Table 1 (10k train examples, 3k test examples, randomly selected), we evaluate six prompts. We find that the resultant token embeddings are stable across variants (cosine similarity > 0.9), and that detection performance is similar across variants. See Table 8 for full results.

Table 8: Prompt robustness and embedding alignment for learned Llama <AIGEN> tokens. AUROC is evaluated using the same prompt used for training on the same RAID split used in Table 1. Cosine is to the mean of all listed <AIGEN> embeddings.

Training/evaluation prompt	**AUROC**	**Cosine to mean**
`Write {token} text.`	0.99	0.951
`Generate a passage of {token} text.`	0.98	0.961
`Write text that is characteristic of {token}.`	0.99	0.951
`Produce the following as {token} text.`	0.98	0.949
`Continue with a passage in {token} style.`	0.99	0.967
`Give a {token} response.`	0.97	0.941

C.2 Embedding geometry mirrors verbalization similarity

A natural reading of the cross-modal transfer in Sec. 4.3 is that $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ pairs trained independently on different modalities should occupy related directions in the embedding space. We test this directly and connect the geometric finding back to verbalization.

We train one neologism pair per modality, holding the backbone (Gemma-4-E4B-it) and training protocol fixed, and analyze the resulting embeddings. For each pair of modalities (m, m') we compute the cosine similarity $\cos\big(\boldsymbol{E}^{(m)}_{\texttt{<AIGEN>}}, \boldsymbol{E}^{(m')}_{\texttt{<AIGEN>}}\big)$, and analogously for <REAL>. We additionally compare each learned vector to the backbone's vocabulary distribution, reporting its norm percentile relative to ordinary word tokens, its cosine to the vocabulary mean, and its top-k nearest vocabulary tokens.

Table 9: **Geometry of Learned $(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}})$ Pairs.** Left: per-modality norm and cosine to the vocabulary mean (all norms fall at the 100th percentile of the backbone vocabulary). Right: cross-modality cosines between independently trained embeddings, compared to a random-vocabulary baseline.

Modality	**Norm**	**Cos. mean**
Learned <REAL> *embedding*		
Text	2.26	+0.17
Audio	2.50	-0.03
Image	2.01	+0.16
Video	3.08	+0.14
Learned <AIGEN> *embedding*		
Text	2.35	+0.15
Audio	2.39	+0.05
Image	2.07	+0.05
Video	3.17	+0.15

	Cos. between <AIGEN>				**Cos. between** <REAL>			
	Txt	Aud	Img	Vid	Txt	Aud	Img	Vid
Text	1.00	–	–	–	1.00	–	–	–
Audio	+0.06	1.00	–	–	-0.00	1.00	–	–
Image	+0.04	-0.01	1.00	–	+0.09	+0.04	1.00	–
Video	+0.04	+0.07	+0.05	1.00	+0.08	+0.07	+0.05	1.00

Within-run $\cos\big(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}}\big)$:
text = +0.15, audio = +0.16, image = +0.16, video = +0.01.

Random vocabulary baseline (i.e., the mean pairwise cosine between two regular vocab tokens): +0.12.

Table 9 summarizes the geometry. Each learned token, in every modality, sits at the 100th percentile of the backbone's vocabulary norm distribution—i.e., longer than any ordinary word embedding—with norms ranging from 2.00 (image <REAL>) to 3.17 (video <AIGEN>). Cross-modality cosines for <AIGEN> range from -0.012 (audio↔image) to $+0.072$ (audio↔video); for <REAL> the range is -0.002 (text↔audio) to $+0.090$ (text↔image). Both ranges are at or below the random-vocabulary

baseline of $+0.118$, indicating that independently trained per-modality embeddings find largely *distinct* directions in $\mathbb{R}^d$ rather than converging on a shared one. This is consistent with the modality-specific gap observed for image under joint training in Sec. 4.3. Within a single run, $\cos\big(\boldsymbol{E}_{\texttt{<REAL>}}, \boldsymbol{E}_{\texttt{<AIGEN>}}\big)$ takes values 0.148 (text), 0.156 (audio), 0.157 (image), and 0.012 (video), i.e., the two trained tokens occupy approximately as related directions as two random vocabulary tokens. This is consistent with their role as opposite poles of a binary decision rather than two highly correlated concepts.

C.3 Self-Verbalization

Table 10: **Verbalized words score consistently with the detector.** We elicit word lists from Llama-3.1-8B-Instruct using the contrastive prompt above, then score each generated word in isolation with the learned detector. Scores are raw detector log-probability margins $s_{\text{text}}(w)$; larger values indicate more AI-like words. $W_{\langle\cdot\rangle}$ are elicited using *trained tokens*; $W_{(.)}$ are elicited using natural-language prompts. Learned AI verbalizations score above learned human verbalizations, learned AI verbalizations score above natural-language AI baseline verbalizations, and learned human verbalizations score below natural-language human baseline verbalizations ($p \approx 0$, Welch one-sided tests).

Elicitation output	Mean $s_{\text{text}}(w)$	Frequent verbalized words / phrases
$W_{\langle\texttt{AIGEN}\rangle}$	$+0.27$	*introduction, abstract, conclusion, analysis, article, research*
$W_{\langle\texttt{REAL}\rangle}$	-0.64	*however, furthermore, consequently, meanwhile, nevertheless*
$W_{\text{(AI-generated)}}$	-0.48	*furthermore, consequently, additionally, notwithstanding*
$W_{\text{(Human written)}}$	-0.40	*anyway, hmm, um, honestly, you know, actually*

C.4 Per-Instance Verbalization

Label: AI **Predicted:** AI **Score:** 2.02

"I don't know if it has been mentioned before, but I have recently been having trouble with Turnitin. It is an on-line service for checking your work for plagiarism and other errors. The problem is that it seems to be unable to tell me what the plagiarized pages are about. Here is one from my senior year that has been plagiarized by another student. Here is one from my junior year that has also been plagiarized. And here is a [. . .]"

Supports AI. (1) Formal tone: The text is written in a formal tone, which is typical of AI text. The language used is straightforward and objective, without any emotional or personal undertones. **(2) Specific details:** The author provides specific examples and details about their experience with Turnitin, which is a characteristic of AI text. This level of detail suggests that the author is trying to convey information in a clear and concise manner. **(3) Request for help:** The author ends the text with a polite request for help, which is a common feature of AI text. This indicates that the author is seeking a solution to their problem and is willing to collaborate with others. **Undermines AI. (1) Informal language:** Although the tone is generally formal, there are a few instances of informal language, such as the use of "I don't know if it has been mentioned before" and "Thank you!". This suggests that the author may not be adhering strictly to AI conventions. **(2) Personal experience:** The text is centered around the author's personal experience with Turnitin, which is not typical of AI text. AI text often focuses on abstract concepts or general information, rather than personal anecdotes. **(3) Emotional undertones:** Although the tone is generally formal, there is a hint of frustration and confusion in the text, particularly when the author mentions the problem with Turnitin. This emotional undertone is not typically found in AI text. **Overall.** While the text exhibits some characteristics of AI text, such as a formal tone and specific details, it also shows some deviations from the typical AI style, such as informal language and personal experience.

Label: Human **Predicted:** Human **Score:** -18.28

"Writing a research paper for a business accounting class and just turned in the rough draft for the paper and Turnitin returned a 45% plagiarism score on it. I didn't plagiarize any of it. The Turnitin system literally flagged the terms "Public Company Accounting Oversight Board", "Sarbanes Oxley Act", and "Securities and Exchange Commission" as all plagiarized. These are all the official names of either a government firm or [. . .]"

Supports AI. (1) Formal tone: The text is written in a formal tone, which is typical of AI text. The language used is straightforward, and the structure is clear and concise. **(2) Use of technical terms:** The text includes technical terms like "Public Company Accounting Oversight Board," "Sarbanes Oxley Act," and "Securities and Exchange Commission," which suggests a formal and specialized context, consistent with AI text. **(3) Academic context:** The text mentions a research paper for a business accounting class, which is a typical context for AI text. **Undermines AI. (1) Informal language and structure:** Despite the formal tone, the text has a somewhat informal structure and language, which is not typical of AI text. The use of phrases like "Turnitin returned a 45% plagiarism score" and "I know, under penalty of death, that I did not plagiarize anything in my paper" suggests a more conversational tone. **(2) Emotional language:** The text includes emotional language, such as "terrible part" and "I know, under penalty of death," which is not typical of AI text. This language suggests a more personal and emotional tone, which is not characteristic of AI text. **(3) Questionable grammar and punctuation:** The text has some grammatical errors, such as the use of a comma splice ("I didn't plagiarize any of it. The Turnitin system. . . ") and missing articles ("a government firm" instead of "a government firm or a law"). **Overall.** While the text has some characteristics that are typical of AI text, such as a formal tone and use of technical terms, it also has some features that are not typical, such as informal language and structure, emotional language, and questionable grammar and punctuation.

Label: AI **Predicted:** AI **Score:** 9.30

"Ingredients: 1 cup all-purpose flour 1 cup cold butter, cut into cubes 1/2 cup granulated sugar 1 1/2 teaspoons baking powder 1/2 teaspoon baking soda 1/2 teaspoon salt 2 eggs 1 teaspoon vanilla extract 2 cups fruit, such as Raspberries or blueberries (fresh or frozen) Instructions: 1. Preheat oven to 375 degrees F (190 degrees C). 2. In a Large bowl, combine flour, butter, sugar, baking powder, baking soda, and salt. Use a [. . .]"

Supports AI. (1) Step-by-step instructions: The text is divided into numbered steps, which is a hallmark of AI text. Each step is concise and clearly describes what to do. **(2) Specificity:** The text provides specific measurements for ingredients (e.g., "1 cup all-purpose flour") and temperatures (e.g., "375 degrees F"). This attention to detail is typical of AI text. **(3) Use of technical terms:** The text uses technical terms like "pastry blender" and "baking powder," which suggests a level of expertise and attention to detail. **(4) Emphasis on precision:** The text includes specific instructions for mixing (e.g., "Use a pastry blender or Fork to mix until mixture resembles coarse crumbs") and baking (e.g., "Bake for 35–40 minutes, or until top is golden brown and a toothpick inserted into the center comes out Clean"). **Undermines AI. (1) Informal tone:** The text uses a conversational tone, which is not typically associated with AI text. The use of phrases like "such as Raspberries or blueberries" and "before Serving" suggests a more casual approach. **(2) Lack of formal language:** The text does not use formal language or technical jargon, which is often characteristic of AI text. The use of simple sentences and everyday vocabulary undermines the AI label. **(3) Some ambiguity:** The text includes some ambiguous instructions, such as "mix until just combined" and "comes out Clean." While these instructions are still clear, they are not as precise as those typically found in AI text. **Overall.** While the text exhibits some characteristics typical of AI text, such as specificity and attention to detail, it also has some features that undermine the label, such as an informal tone and some ambiguity.

Label: Human **Predicted:** Human **Score:** -9.11

"2 cups flour 2 teaspoons baking powder 3/4 cup sugar 1/8 teaspoon salt 1/2 cup melted butter 2 large eggs at room temperature 2 8-ounce packages of cream cheese, softened 3/4 cup sugar 1 tablespoon flour 12 ounces evaporated milk 2 teaspoons vanilla extract 3 large eggs, at room temperature Ground cinnamon, for the top (optional) 1/2 cup nuts such as walnuts or pecans, for the top (optional). Preheat the oven to 325 °F. Have a [. . .]"

Supports AI. (1) Step-by-step instructions: The text is divided into clear steps, starting with preheating the oven and preparing the pan, followed by combining ingredients, and then pressing the dough into the pan. This structure suggests a meticulous and organized approach, which is characteristic of AI text. **(2) Attention to detail:** The text includes specific measurements, temperatures, and instructions for mixing and preparing ingredients. This level of detail is typical of AI text, which often prioritizes precision and thoroughness. **(3) Use of technical terms:** The text employs technical terms like "stand mixer," "hand-held electric mixer," and "bowl of a stand mixer or hand-held electric mixer." This use of specialized vocabulary suggests a focus on accuracy and clarity, which is consistent with the AI label. **Undermines AI. (1) Optional ingredients:** The text includes optional ingredients like ground cinnamon and nuts, which suggests a more relaxed approach to the recipe. AI text often strives for precision and consistency, which might be compromised by the inclusion of optional ingredients. **(2) Lack of explicit timing:** The text does not provide explicit timing for each step or the overall preparation process. While this might be due to the complexity of the recipe, it could also suggest a more flexible or adaptable approach, which is not typically characteristic of AI text. **(3) Informal tone:** The text has a conversational tone, with phrases like "at room temperature" and "for the top (optional)." While this tone is not necessarily inconsistent with AI text, it does suggest a more approachable and less formal approach, which might undermine the AI label. **Overall.** The text exhibits both characteristics that support and undermine the AI label. While it includes detailed instructions and technical terms, it also includes optional ingredients and lacks explicit timing, which might suggest a more flexible or adaptable approach.

Figure 4: Instance verbalizations for held-out RAID examples. Each card shows the input excerpt, gold label, detector prediction $\hat{y}(x)$, detector score $s_m(x)$, and the elicited local verbalization. "AI" is substituted for "Type-A" in verbalizations. See prompts in Section B.3.

Code, weights, and data are available at `https://neologism-detection.github.io/`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract discusses the core contribution: a parameter efficient, multi-modal method for AI text detection that admits validated natural language descriptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our methods and results in a dedicated subsection (see Sec. 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: We present our method along with motivating theoretical infrastructure, but do not present theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, the paper outlines the specific models, loss functions, metrics, prompts, and hyper-parameters used to reproduce our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release code and trained token embeddings that can be used reproduce our main results. All data used is publicly available and specified in the body/appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the datasets, baselines, AUROC evaluation, held-out-generator protocol, cross-dataset OOD protocol, and perturbation settings. The remaining hyperparameter details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where applicable, we present the significance of our results along with the statistical test performed. Where this is not feasible, we provide robustness checks across training configurations (leave-one-out generator results, different adversarial configurations, etc.).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section A.6 in the Appendix specifies the hardware requirements to reproduce our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the Code of Ethics and are confident that our work conforms with the requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Sections 1 and 5 describe the positive and negative societal consequences of AI detection.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release a scraped dataset or generative model capable of producing harmful outputs that would require access safeguards. Though there is some potential for misuse of AI-detector technology, this misuse arises from the deployment context. We describe responsible use of released artifacts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (RAID, AUDETER, GenImage, DFE-2024, Beemo, In-the-Wild, CIFAKE, DeeptraceReward, EvoBench) and baseline implementations are cited in App. A.4–A.5; we use each under its original license and respect its stated terms of use. We release no derivative datasets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced by this work are documented in the paper (Sec. 3, App. A) and will be released with full documentation at camera-ready time on de-anonymized hosting.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not include any human experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not use any human participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the subject of this work (Llama-3.1-8B and Gemma-4-E4B), not a tool used in producing it. No LLMs were used as a non-standard component of paper preparation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.