Visual Deception Detection for Financial Security: Homoglyph and Deepfake Identification

by

Aleksandar Jovanovic-Hacon

B.S. Artificial Intelligence and Decision Making, MIT, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2026

Authored by: Aleksandar Jovanovic-Hacon
Department of Electrical Engineering and Computer Science
December 10, 2025

Certified by: Amar Gupta
Research Scientist, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair
Master of Engineering Thesis Committee

Visual Deception Detection for Financial Security: Homoglyph and Deepfake Identification

by

Aleksandar Jovanovic-Hacon

Submitted to the Department of Electrical Engineering and Computer Science
on December 10, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Visually-deceptive text, manifesting through homoglyph attacks, poses a growing threat across cybersecurity, financial systems, and identity verification. While existing string detection methods are easily scalable, they fail to capture the nuanced perceptual similarities between textual inputs. This paper introduces Visually-Aligned Text Embeddings (VA-TE), a contrastive learning framework that uses curriculum learning to train a lightweight projection head on top of a pretrained vision-language model (VLM) text encoder, translating semantic embeddings into representations aligned with human visual perception. Crucially, this method operates directly on text, enabling faster, lighter-weight detection than prior image-based approaches that require rendering and training on millions of images. Evaluated on a large-scale homoglyph benchmark, VA-TE attains competitive performance (AUC = 0.95) using a SigLIP encoder backbone, approaching the performance of state-of-the-art (SOTA) image-based methods while offering substantial gains in scalability and deployment efficiency. Further, fusing these visually-aligned embeddings with complementary string-similarity features yields SOTA results (AUC = 0.98), underscoring the value of multi-modal signals for robust spoof detection. Taken together, VA-TE establishes a self-supervised procedure for converting semantic text features into visually grounded representations applicable across perceptual string-matching tasks.

Thesis supervisor: Amar Gupta
Title: Research Scientist

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Dr. Amar Gupta, for his unwavering guidance, encouragement, and support throughout the course of this work. From the initial stages of formulating ideas to the final steps of refining this thesis, Dr. Gupta's insight and mentorship were essential. His leadership has shaped not only the quality of this research but also my development as a researcher and thinker.

I would also like to thank the other members of our research team who contributed to various aspects of this project. I am especially grateful to Sean-Winston Luo, who worked closely with me on Section 1 of this thesis (VA-TE) and whose collaboration, feedback, and persistence were invaluable. I also extend my thanks to Zhurui Sheng and Ricardo Carrillo, whose partnership was central to building and refining the deepfake detection system, both technically and conceptually. Their dedication and teamwork made the process both productive and rewarding. I am similarly grateful to Dr. Rafael Palacios for his thoughtful input throughout the past year. His guidance in drafting manuscripts and communicating our findings across data and science contexts greatly strengthened the clarity and impact of this work.

I would like to thank our sponsors at Itaú Unibanco—Miguel Domingos Wanderley, Fernando Beserra, Jair Carvalho, and Lucas Orosco—for their support, engagement, and collaboration. Their helpful domain insights throughout this work greatly informed and contextualized the research.

Finally, I would like to thank my family for their steadfast support. To my sister, Ana, whose companionship and encouragement during our time living together in Boston sustained me through the challenges of graduate school. Without her, I would not have been able to gain admission to the program, let alone complete this thesis. I am also deeply grateful to my parents, Christopher and Aleksandra, for inspiring my curiosity, nurturing my love for learning, and serving as the guiding light throughout my academic journey. Their belief in me has been my greatest source of strength.

Portions of the thesis text were revised using large language models for grammar correction, tense consistency, and stylistic clarity. All technical content, analyses, experiments, and original ideas are my own.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The proliferation of digital content has brought with it an escalating challenge: visual deception. This threat manifests critically in financial systems, where adversaries exploit visual similarities to bypass security measures. Two problems exemplify this challenge. First, in homoglyph attacks, adversaries create visually similar account names to deceive users into trusting decptive accounts. These spoofed names look nearly identical to legitimate companies, fooling human observers, while evading detection systems that rely on text-based string matching rather than visual appearance. Second, AI-generated synthetic media, including deepfakes, threaten biometric verification and identity authentication systems that rely on facial recognition. These deception techniques span multiple areas and types of impropriety, yet share a common strategy: exploiting the gap between visual appearance and underlying representation to deceive both humans and automated systems. Given the scale and reach of these threats, robust and accurate detection systems are essential to maintain the integrity of modern financial infrastructure, identity verification platforms, and information ecosystems that depend on content authenticity. Traditional detection methods have approached these challenges by identifying specific artifacts or patterns characteristic of deceptive content. In the text domain, string matching algorithms rely on character-level metrics such as edit distance and token overlap, treating homoglyphs as a text problem when the deception operates visually. In the audio-video domain, deepfake detectors learn to recognize generation artifacts left by specific models. However, this artifact-driven paradigm suffers from a critical limitation: adversaries continuously evolve their techniques, rendering artifact-based detectors brittle. A detector trained to recognize artifacts from one generative model often fails catastrophically when confronted with outputs from newer, unseen models. As generative capabilities advance, the cat-and-mouse game of artifact detection becomes increasingly unsustainable. This thesis proposes a paradigm shift: rather than chasing the ever-changing artifacts of deceptive content, the approach instead learns the stable, intrinsic properties that characterize authentic content. This reframing transforms detection from a reactive pursuit of adversarial signatures to a proactive learning of what genuine content looks like. By focusing on the commonalities shared by real images or the natural alignment between text semantics and visual appearance, the trained detectors are able to generalize

across deception methods.

1.2 Thesis Overview

This thesis addresses visual deception for two distinct tasks, developing specialized approaches for each. These two distinct tasks are delineated in the following chapters.

1.2.1 Part I: Textual Visual Deception (Chapter 2)

Chapter 2 tackles homoglyph attacks, where adversaries substitute visually similar characters to evade text-based spoof detection. A naive visual solution might treat text as images, rendering strings and applying computer vision models, but this proves impractical at scale due to the computational overhead and memory requirements of image processing for every account registration or transaction. Instead, the approach introduces Visually-Aligned Text Embeddings (VA-TE), which exploits a key insight: Vision-Language Models (VLMs) pretrained on image-text pairs learn text encoders that implicitly capture visual characteristics of text appearance. By fine-tuning the VLM text encoder directly, VA-TE encodes visual properties from text itself, without rendering images, enabling fast, memory-efficient inference at deployment scale. Through contrastive learning with curriculum-based hard negative mining, VA-TE learns to embed visually similar strings close together in representation space, even when they are distant according to character-based string metrics. Combined with complementary string-based features in an ensemble, VA-TE achieves 98% average precision, matching or exceeding vision-based methods while offering substantial advantages in speed, memory footprint, and deployability for real-world systems.

1.2.2 Part II: Image-Level Visual Deception (Chapter 3)

In an era of rapidly evolving generative models, existing detectors overfit to artifacts specific to training-set generative models and fail on unseen model families. A disentangled representation learning framework is proposed that aims to separate authenticity-relevant features from identity and content information within pretrained audio-visual embeddings. Specifically, dual projection heads with orthogonality constraints learn complementary subspaces: an authenticity head trained via variance minimization to capture properties shared across real samples, and an identity head trained via prototypical contrastive learning to encode content-specific information. A temporal transformer classifier then operates on the authenticity embeddings to produce frame-level predictions. The approach is evaluated on AVDeepfake1M++ [1] and ShareVeo3 [2], with out-of-distribution generalization assessed on a novel test set of 150 videos (11,000+ segments) collected from OpenAI's Sora 2. Preliminary results reveal challenges with representation collapse during multi-objective optimization, providing diagnostic insights into the difficulties of disentangled learning for this task and motivating specific directions for future work.

While these two problems differ substantially in their technical domains, from text embeddings to image forensics, they share a unifying methodological philosophy: learning stable

representations of authentic content, enabling more robust detection systems.

1.3 Contributions

This thesis presents a unified framework for visual deception detection, addressing two complementary challenges: homoglyph-based textual deception and AI-generated media. The primary contributions are as follows:

1. **Visually-Aligned Text Embeddings (VA-TE) for Homoglyph Detection.** A novel approach is introduced that leverages vision-language model text encoders to capture visual properties of Unicode characters without requiring explicit image rendering. This method achieves competitive performance with traditional visual approaches while offering superior scalability.

2. **Disentangled Representation Learning for Deepfake Detection.** A dual-projection architecture with orthogonality constraints is proposed to learn semantically meaningful and generalizable features for distinguishing authentic from synthetic video content. The approach separates authenticity-relevant features from generator-specific artifacts, with the goal of improving out-of-distribution generalization to unseen generation methods.

3. **Novel Evaluation Dataset.** A new out-of-distribution test set is collected and annotated, comprising 150 videos (11,000+ segments) generated by OpenAI's Sora 2, providing a challenging benchmark for evaluating generalization to state-of-the-art generation methods not contained in the training datasets.

4. **Gradient-Balanced Multi-Objective Optimization.** Gradient normalization techniques [3] are adapted to balance competing objectives in representation learning, demonstrating effectiveness in preventing loss collapse in disentangled learning scenarios.

Together, these contributions advance the broader goal of detecting visual deception across modalities, from character-level substitution attacks to sophisticated AI-generated media-based deceptions.

Chapter 2

Visually-Aligned Text Embeddings for Homoglyph Detection

2.1 Introduction

Research exploring visual similarity of text has remained underexplored, despite vast applications within cybersecurity [4], website infrastructure [5,6], bank check processing [7], and other growing digital fields. Organizations operating online platforms can benefit from systems that protect against visually deceptive text. Businesses centered around user accounts must create software designed to prevent username impersonations [8], while financial technology companies require systems to ensure precise digital check processing. Each of these cases demands high accuracy, minimizing false negatives that prevent malicious visually-similar text from scraping through, while avoiding false positives that cause significant customer inconvenience.

A common spoofing tactic is the use of *homoglyphs*, characters from different writing systems that look identical but have distinct Unicode code points. For instance, the Latin letter "a" (U+0061) and the Cyrillic "a" (U+0430) appear the same to readers but not to computers. Similarly, attackers may substitute the letter "O" with the digit "0," or replace "e" with accented forms such as "é." Such substitutions exploit human perception while bypassing naive string matching, making homoglyph detection essential for spoof prevention.

Despite this problem's relevance, existing text similarity research focuses on semantic relationships [9–12] —aligning words and phrases based on meaning rather than visual appearance. Methods for detecting visually-similar text have lagged behind industry needs, with most companies using string matching techniques to fulfill various mission-critical detection tasks. These methods include string similarity metrics like Levenshtein distance [13], which define decision boundaries based on the number of edits required to match two names. Fuzzy matching techniques are also used to assess relative similarity, where names with high similarity scores are flagged as potentially anomalous. Typically, newly registered text, including account registrations and usernames, is iteratively checked against a reference dataset, with higher similarity indicating visually-similar text.

String-based similarity detection methods face a major challenge: they fail to capture the perceptual nuances of string modifications, as these are intentionally designed to be

visually subtle to evade downstream detection. Methods in current use can detect minor textual alterations, but often create false positives when string lengths become smaller [14]. This necessitates the development of more sophisticated methods that align with human-like perception of visual string similarity, specifically a scalable and practical implementation suitable for research and prototyping contexts. In practice, attackers blend visual similarity with textual manipulation patterns. A robust detection system must therefore capture both dimensions of deception—perceptual similarity and structural variation—to defend against a spectrum of spoofing attacks.

Visually-Aligned Text Embeddings (VA-TE) are proposed, a contrastive learning [15] framework that expands upon pre-trained vision-language models [16] capable of encoding the visual characteristics of text. Through building on models that generate semantically rich embeddings directly from text, this approach captures subtle, visually deceptive name variations, achieving significantly higher separability than string-based methods, improving ROC-AUC from 0.81 [17] to 0.95, efficiency and scalability advantages over existing approaches [17,18]. To fulfill this task, a lightweight projector [19] is fine-tuned on top of these pre-trained embeddings to enhance the system's ability to distinguish between legitimate names and visually-similar spoofs.

Contrastive learning [15] is used, experimenting with various contrastive objectives [20–22] to optimize similarity in the target embedding space. Additionally, curriculum learning [23] is employed to progressively introduce more challenging negative examples during training, enabling the model to learn nuanced decision boundaries that improve sensitivity to string variations. To evaluate the effectiveness of VA-TE, spoof detection is framed as the primary evaluation task, measuring the quality of the learned embeddings through their ability to distinguish between legitimate company names and visually deceptive spoof counterparts.

The paper is structured as follows: Section 2.2 reviews relevant literature, spanning traditional string similarity methods, contrastive learning techniques, and vision–language models. Section 2.3 delineates the proposed method, including the architectural design and the training objectives. Section 2.4 describes the datasets, evaluation protocols, and baseline comparisons, while Section 3.5 presents experimental results. Section 2.6.3 discusses avenues for future work, and Section 2.6 concludes with a summary of the findings.

2.2 Related Works

2.2.1 String Matching Methods

Despite challenges with accuracy and implementation [14], string-based similarity methods act as fast, computationally efficient proxies for evaluating visual similarity between strings, demonstrating value as components in similarity detection systems [24]. These methods fall into two main categories: edit distance metrics and fuzzy matching methods, each capturing different aspects of textual manipulation patterns commonly exploited in spoofing attacks. While individually limited, their distinct strengths and failure modes make them well-suited for ensemble-based approaches.

Levenshtein Distance Metrics

Levenshtein distance [13] measures the minimum number of single-character edits (insertions, deletions, and substitutions) required to transform one string into another. This metric excels at detecting minor manipulations in longer strings. However, the edit distance frequently exhibits systematic failures that limit its standalone effectiveness. For short strings, legitimate variations and unrelated names can trigger false positives due to the metric's insensitivity to string length. A single character difference between genuinely distinct entities can produce a low edit distance that incorrectly suggests spoofing. Conversely, sophisticated spoofing attacks on longer strings can involve many character edits while maintaining visual similarity, resulting in high edit distances that evade detection. This can be seen in Table 2.1.

Table 2.1: Levenshtein distance misclassifications (threshold of ≤ 2 edits).

Levenshtein Distance Failure #1	**Levenshtein Distance Failure #2**
'3M' vs 'H&M'	'Microsoft' vs 'Micrsofts.'
Distance = 2 $\rightarrow$ Spoof **(Wrong)**	Distance = 3 $\rightarrow$ Non-Spoof **(Wrong)**

Despite these limitations, Levenshtein distance achieves reasonable performance as a standalone classifier, with reported ROC-AUC of 0.81 (see Section 2.4.6 for metric selection rationale) on domain name spoofing tasks [17].

Fuzzy Matching Methods

Percentage-based similarity metrics, including Token Sort Ratio, Partial Ratio, and Token Set Ratio[1], calculate similarity based on token overlap and string length normalization. These methods address edit distance's length sensitivity by producing normalized similarity scores, making them particularly effective for longer strings where proportional similarity is more meaningful than absolute character differences.

While commonly used by financial institutions, fuzzy matching methods exhibit failure patterns that limit their standalone effectiveness. When strings share common words or substantial token overlap, percentage metrics yield high similarity scores that create false positives for genuinely distinct entities. Conversely, well-designed spoofs that appear visually similar but are textually dissimilar, particularly in short strings, often evade detection entirely, resulting in false negatives. Examples are shown below in Table 2.2.

Fuzzy matching achieves superior standalone performance compared to editing distance, with a reported ROC-AUC of 0.86 on domain spoofing tasks [17], reflecting its improved handling of string length variations.

[1]Fuzzy Matching library, including Token Set Ratio, can be found here: https://rapidfuzz.github.io/RapidFuzz/Usage/fuzz.html

Table 2.2: Token set ratio misclassifications (threshold of ≥ 80).

Token Set Ratio Failure #1	**Token Set Ratio Failure #2**
'Invest Co' vs 'Square Invest Co'	'Capital One' vs 'Capito1 0ne'
Ratio = 100 $\rightarrow$ Spoof **(Wrong)**	Ratio = 55 $\rightarrow$ Non-Spoof **(Wrong)**

Complementary Strengths and Ensemble Potential

An analysis of failure conditions reveals that edit distance and fuzzy matching offer complementary strengths for spoof detection. Levenshtein distance performs well on short strings with character-level manipulations, where fuzzy matching often fails. Conversely, for longer strings with proportionally similar tokens, fuzzy matching excels while edit distance struggles. These patterns suggest that combining both metrics in an ensemble approach could capture a wider range of spoofing behaviors than either method alone.

Despite these strengths, both techniques fundamentally fail to account for the perceptual subtleties of visual similarity, an essential aspect of effective spoofing as illustrated by the misclassifications in Tables 2.1 and 2.2. This limitation motivates the development of more advanced methods that embed visual characteristics into their representations to improve similarity analysis and spoof detection.

2.2.2 A Visual Non-ML Approach: HitZone Mapping

HitZone Mapping [25] assesses visual similarity by dividing characters into grids and measuring the degree of overlap between corresponding squares. Each character is mapped onto a grid with varying degrees of granularity, allowing for detection of minor visual differences and providing an interpretable framework for spoof detection. While HitZone maps have been successfully applied at the character level, they have not been widely explored in the context of word-level similarity detection.

When comparing different length strings, words must be carefully aligned so hitboxes match-up on a character-to-character basis. This task would require a dynamic programming approach that has not yet been formally developed.

2.2.3 Embeddings for Similarity Measurement

Unlike the previously mentioned methods, embeddings provide an efficient and continuous framework for similarity assessment, most commonly cosine similarity [9,26]. Learning a task-specific embedding space enables effective thresholding and ranking, both of which are essential for real-time retrieval and large-scale systems [27].

Accordingly, embeddings have been used for similarity assessment across many different tasks. Within computer vision, they power systems for facial recognition [20] and image classification [28], both of which require precise ranking of visually-similar features and items.

Extending even to symbols, embeddings have enabled tasks like sign language recognition [29] and handwritten text retrieval [30,31], where visual patterns rather than strict symbolic sequences define similarity. In text-based tasks, embeddings have achieved widespread usage in word semantic settings that measure meaning-based closeness [9,11]. The overwhelming literature outlining the success of embeddings in similarity analysis [32] motivates the usage of embeddings for visual text similarity analysis.

2.2.4 Contrastive Learning

Contrastive learning [15] has been effectively used to learn task-specific embeddings requiring fine-grained distinction[33], including face forgery detection[34], text-image alignment[10], and self-supervised representation learning[19].

In particular, SimCLR [19] presents a simple yet powerful framework for contrastive learning of visual representations, using data augmentation to generate positive pairs, a neural network encoder to compute embeddings, and a projection head to map these embeddings into a space where contrastive loss is applied. This framework motivates the methodological approach presented in this thesis.

Additionally, prior work has applied contrastive learning frameworks, specifically Siamese convolutional neural networks [35], to the problem of homoglyph detection by learning visually-aligned representations of text strings [17,18]. These studies demonstrate the ability of contrastive loss for distinguishing visually similar strings, underscoring the relevance of contrastive learning for spoof detection tasks.

In contrast to classification-based training, which learns embeddings incidentally while optimizing for label prediction [36], contrastive learning explicitly shapes the embedding space to align with meaningful notions of similarity by bringing similar (positive) samples closer and pushing dissimilar (negative) samples apart [15,37]. The goal is to learn a feature space in which similar inputs are mapped to nearby points, while dissimilar inputs are separated by a significant margin, using distance metrics like cosine distance.

In the literature, a range of loss functions have been developed to implement this learning objective:

- **Contrastive Loss:** [15] This formulation penalizes positive pairs that are far apart and negative pairs that are too close, typically including a margin hyperparameter to encourage strong separation without collapsing the embedding space.

$$\mathcal{L}_{\text{contrastive}} = y \, \|z_i - z_j\|^2 + (1 - y) \, [\, m - \|z_i - z_j\| \,]_+^2 \tag{2.1}$$

  where $y = 1$ for positive pairs and $y = 0$ for negative pairs, and m is the margin.

- **Triplet Loss:** [20] Extends the contrastive formulation to anchor-positive-negative triplets, ensuring the anchor is closer to the positive than the negative by a margin.

$$\mathcal{L}_{\text{triplet}} = \max\left(0, \|\mathbf{z}_a - \mathbf{z}_p\|^2 - \|\mathbf{z}_a - \mathbf{z}_n\|^2 + \alpha\right) \tag{2.2}$$

  where $\mathbf{z}_a$, $\mathbf{z}_p$, and $\mathbf{z}_n$ are the anchor, positive, and negative embeddings, and α is the margin.

- **InfoNCE:** [21] Commonly used in self-supervised learning frameworks, InfoNCE generalizes the contrastive objective to a multi-negative classification task, maximizing the similarity of the anchor-positive pair relative to a set of negatives and encouraging discriminative representation learning in large batches.

$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{a\in A(i)}\exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}\right) \tag{2.3}$$

  where τ is the temperature and $A(i)$ includes one positive and multiple negatives.

- **Supervised Contrastive Loss (SupCon):** [22] Extends InfoNCE to leverage batches of positive samples. This is particularly beneficial in spoof detection where multiple spoof variants may exist for a single legitimate entity.

$$\mathcal{L}_{\text{sup}} = -\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i^\top z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i^\top z_a/\tau)} \tag{2.4}$$

  where $P(i)$ is the set of positives for anchor i, excluding itself.

A critical component of effective contrastive learning is the use of hard negatives [37,38] - negative samples that are close to positives in the initial embedding space before applying contrastive learning. By challenging the model to distinguish between such near-positives, hard negatives refine decision boundaries and encourage the learning of robust, discriminative representations. This consideration naturally motivates the use of curriculum learning strategies, described in the following section.

2.2.5 Curriculum Learning

Curriculum learning [23] is a training paradigm that structures the learning process by gradually increasing the difficulty of training examples over time, mimicking the way humans learn progressively from simpler to more complex tasks. Originally introduced in natural language processing and computer vision tasks, curriculum learning has since been applied to deep metric learning [39] and contrastive representation learning [40], where it has been shown to improve convergence, generalization, and robustness of the learned embedding spaces.

In contrastive learning specifically, where models are trained to bring positive pairs closer and push negative pairs apart in the embedding space, the selection and ordering of training examples, particularly negative samples, plays a critical role in shaping model performance. A naïve strategy that samples negatives uniformly at random may lead to uninformative or overly trivial contrastive signals [41], which in turn limit the model's ability to develop nuanced decision boundaries. This is especially problematic in tasks like homoglyph detection, where many adversarial examples are intentionally designed to be visually deceptive, requiring the model to capture subtle typographic cues.

To address this, recent studies in curriculum-based contrastive learning propose incorporating progressively harder negative samples over the course of training [38]. This structured

training helps the model refine its embedding space to capture more fine distinctions. This concept is closely related to hard negative mining [20,37], a well-established technique in deep metric learning, where negatives that are difficult to distinguish from positives are prioritized to accelerate learning and improve discriminative capacity.

Several curriculum learning strategies have been explored in prior work:

- **Curriculum Learning** [23] introduces samples incrementally based on a difficulty metric (e.g., loss, similarity score), allowing the model to focus first on easier instances and later adapt to harder ones.
- **Self-Paced Curriculum Learning (SPL)** [42] ranks training samples based on their contribution to model loss, ensuring continued focus on examples where the model struggles most.
- **Bandit-Based Adaptive Sampling** [43] applies multi-armed bandit frameworks to dynamically allocate sampling effort across difficulty tiers, selecting the most informative samples based on learning progress.

These strategies have demonstrated improvements across various domains, including image retrieval [44], face verification [39], and few-shot classification [45]. However, their use in vision-language pretraining and visually-aligned text representation learning remains unexplored. In this work, the approach builds on these principles and extends curriculum learning into the realm of textual visual-similarity analysis by integrating it with contrastive learning over pre-trained embeddings.

By leveraging curriculum learning to shape the training trajectory of hard negative examples, a pretrained text encoder—originally optimized for semantic alignment—is repurposed to instead encode visual similarity between text strings.

2.2.6 Vision-Language Models

Vision-Language Models (VLMs) [16] have recently emerged as powerful tools for learning joint representations of images and text, finding applicability for image-text matching, image-to-text generation, and zero-shot classification [46]. Trained on large datasets, typically of image-caption pairs [10], these models align semantic information in a shared embedding space, making them effective at recognizing meaning across both modalities. While most VLMs are trained for semantic alignment, prior work has shown that visual features of text can emerge in their internal representations [47], and that semantic similarity in text often correlates with visual similarity [48]. This motivates the investigation into whether such models, when enhanced with projection layers [19], can be effectively adapted for VA-TE. The following examples illustrate cases where semantic similarity aligns with visual similarity (Table 2.3).

Alongside visual-semantic correlation, many VLMs are trained on images that contain naturally occurring text [10], like street signs, logos, and item labels, that are paired with text that describe them. This indirect exposure to textual appearance in visual contexts allows the model to learn association between word meaning and visual appearance [47]. These two factors motivate the usage of VLMs for detecting *perceptual* similarity of text.

Table 2.3: Examples of strong visual-semantic alignment.

Word 1	**Word 2**	**Visual-Semantic Relationship Explanation**
Doggy	Doggie	Visually similar spellings with identical semantic meaning
Google	Go0gle	Visually similar company names with high semantic overlap

To determine the model-specific approach, VLMs are categorized into four architectural types: contrastive-based, VLMs with masking objectives, generative-based, and VLMs from pretrained backbones[49].

- **Contrastive-based VLMs** (CLIP [10], OpenCLIP [**Cherti2023**], SigLip [50], ALIGN [51]) use separate encoders for image and text modalities, aligning their outputs through a contrastive loss function.
- **VLMs with masking objectives** (FLAVA [52]) process image and text together through a single transformer with early fusion and cross-modal attention, making them especially effective for tasks that require understanding detailed interactions between different modalities.
- **Generative-based VLMs** (BLIP [53], BLIP-2 [54], CoCa [55]) combine a vision encoder with a generative text decoder. CoCa, in particular, fuses together dual-encoder retrieval with captioning capabilities.
- **VLMs from Pretrained Backbones** (LLaVA [56] and GPT-4V [57]) integrate visual inputs through adaptors or embedded vision tokens into a pre-trained LLM. They are typically more suited for semantic understanding tasks.

For identifying visually-similar text, VLMs capable of encoding text directly are most desirable. Leveraging text encoders avoids the usage of images for visual embeddings, reducing memory overhead, training complexity. A baseline test is conducted to confirm the efficacy of contrastive-based models and select a pretrained VLM.

2.3 Methods

2.3.1 Model Architecture and Learning Objectives

The proposed framework, **VA-TE**, builds upon the power of pretrained VLM embeddings to encode visual characteristics of text [47], rather than just semantic meaning. However, using a pretrained VLM without additional fine-tuning is insufficient, as the model faces problems when semantic-visual correlation fails. See Table 2.4 for reference.

To refine the input embedding space, training is performed using four contrastive objectives: pairwise loss, triplet loss, InfoNCE, and Supervised Contrastive Loss (SupCon), each implemented based on original formulations in the literature. The only notable deviation is

Table 2.4: Examples of weak visual-semantic alignment.

Word 1	**Word 2**	**Visual-Semantic Relationship Explanation**
Doggy	Puppy	Visually different words with identical semantic meaning
Google	Goggle	Visually similar company names with different semantics

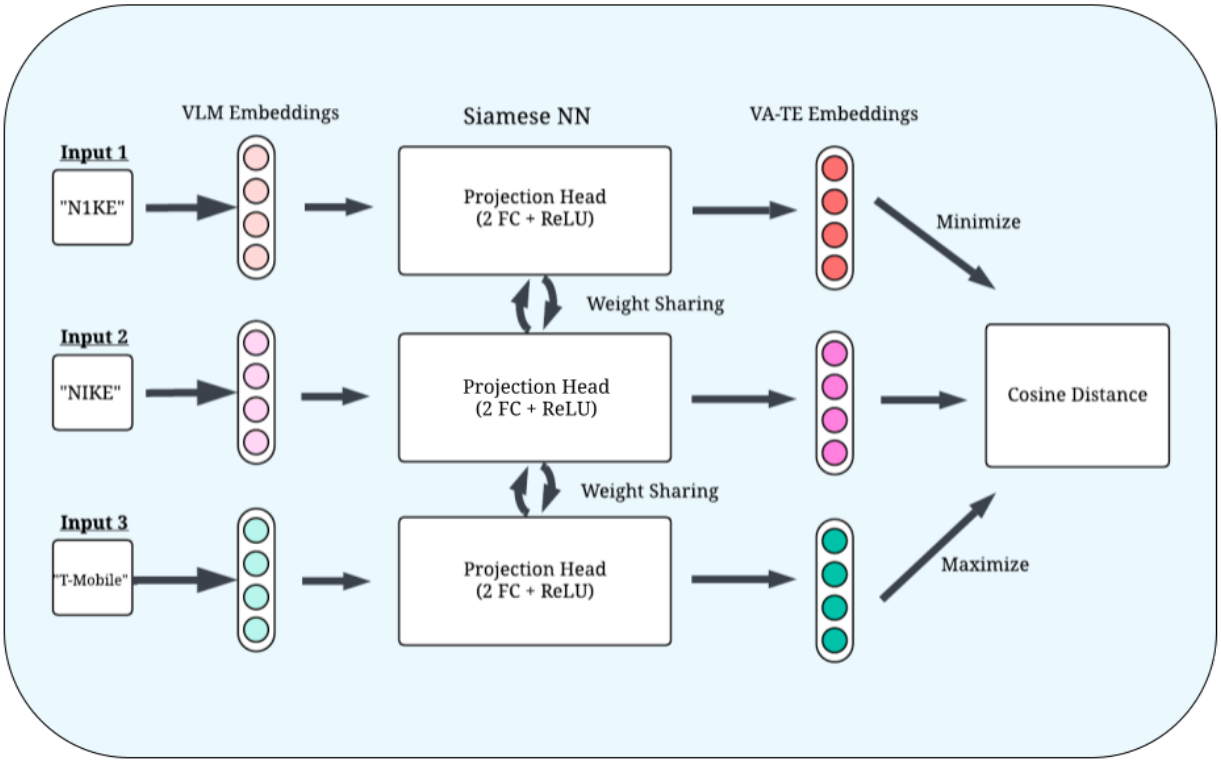


Figure 2.1: Overview of the VA-TE training architecture with light linear projector head.

the use of cosine similarity in the triplet loss instead of the traditional Euclidean distance—a change that yields equivalent results in practice due to prior embedding normalization.

Supporting these contrastive objectives, VA-TE follows a simple yet effective architecture (Fig. 2.1). Each input string is first passed through a frozen text encoder from a pretrained vision-language model, which produces a high-dimensional semantic embedding. These embeddings are then fed into a lightweight projection head, consisting of two shared linear layers with ReLU activation, designed to map semantic representations into a space that better reflects visual similarity. During training, the architecture operates as a Siamese network, receiving inputs and optimizing distances based on label relationships, essentially pulling positive pairs closer together, while negative pairs are pushed apart. A prediction is made by computing the cosine similarity score between 2 input names, and classifying as a spoof if the score exceeds a predefined threshold.

2.3.2 Curriculum Learning

To strengthen the discriminative capability of the model, three curriculum learning strategies are employed. Datasets of three difficulties: easy, medium, and hard, are used, based on the separability of positives and negatives in the input (semantic) embedding space.

- **Manual Approach:** a fixed-schedule curriculum where the training process is divided into three equal stages by epoch, each stage increasing in example difficulty. This structured progression allows the model to first reinforce easily separable relationships in the input embedding space, before being challenged with hard negatives that closely resemble positives—forcing the model to learn fine-grained distinctions between visually-similar text.

- **Automated Approach:** instead of set stages, each epoch uses a training set with a mix of easy, medium, and hard samples. The ratio of samples from each difficulty level is determined by a scheduling function which dynamically adjusts these ratios as training progresses. The curriculum starts with a high proportion of easy samples and gradually shifts toward more hard negatives. This method follows a predefined curriculum, but provides smoother transitions than the manual approach, enabling more stable learning and potentially more effective adaptations to challenging examples. This scheduling strategy is defined by (2.5) and (2.6) and illustrated in Fig. 2.2.

$$t = \frac{\text{current epoch}}{\text{total \# of epochs}}, \quad c(t) = \cos(\pi t) \tag{2.5}$$

$$\begin{bmatrix} \text{easy}(t) \\ \text{medium}(t) \\ \text{hard}(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(1 + c(t))(1 - t) \\ \frac{1}{2}\left[1 - (1 - 2t)c(t)\right] \\ \frac{1}{2}(1 - c(t))t \end{bmatrix} \tag{2.6}$$

- **Adaptive Bandit-Based Approach:** frames curriculum learning as a multi-armed bandit problem, where each difficulty level represents an arm of the bandit. At each epoch, the trainer balances exploration and exploitation by following an epsilon-greedy policy, selecting which dataset to use either at random or by estimated reduction in loss. This adaptive strategy allows the training process to dynamically focus on the most beneficial dataset, rather than adhering to a predefined schedule.

2.3.3 Ensemble with Textual Features

Hard negatives pose a major challenge in spoof detection because they appear close to genuine company names in the VLM's input (semantic) embedding space. Although contrastive objectives aim to separate these examples, hard negatives often remain near the anchor, especially when they share structural traits the model deems semantically relevant. Consequently, non-spoof names are frequently misclassified, hard negatives exhibit high cosine similarity despite differing in string-based metrics (Table 2.5), exposing a key limitation of

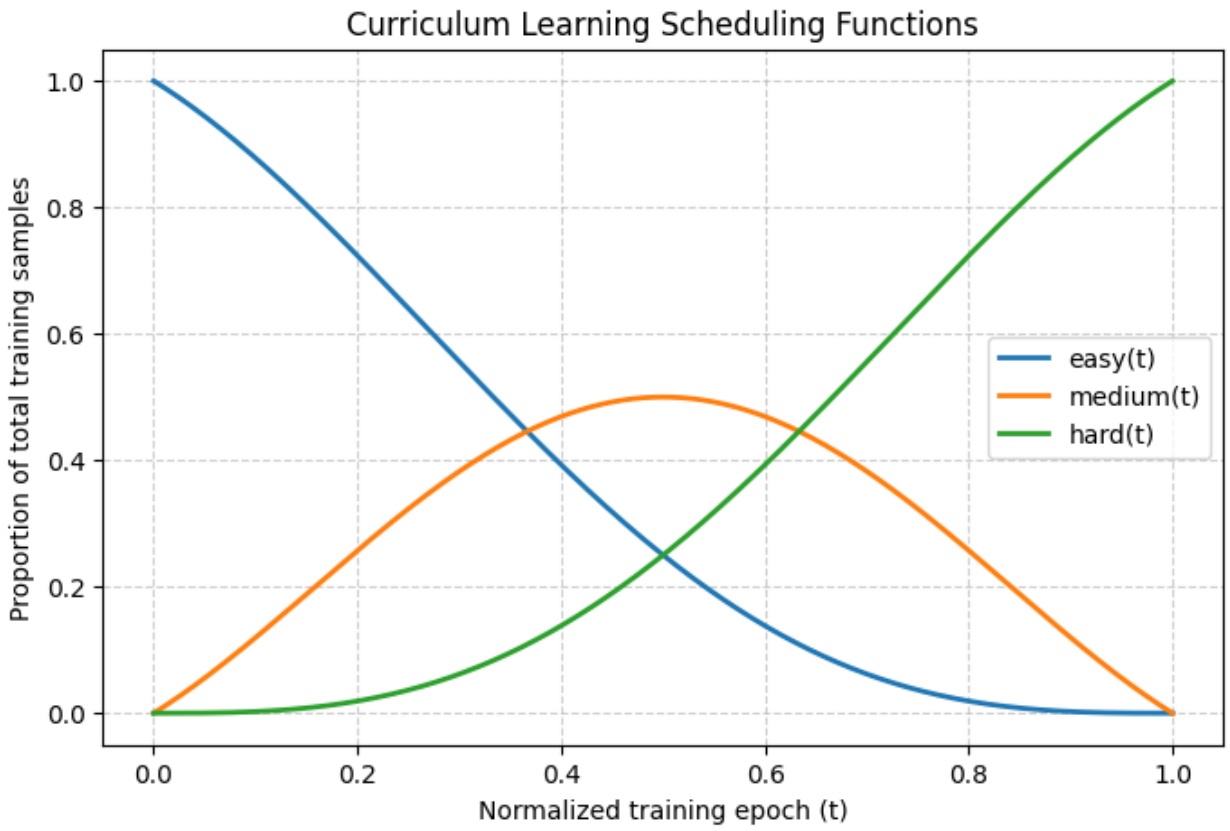


Figure 2.2: Automated curriculum scheduling function. Easy samples dominate early epochs, medium samples peak midway, and hard samples increase toward training end

utilizing embedding-based methods alone: their reliance on transforming semantic similarity into visual-alignment can limit homoglyph detection accuracy.

Table 2.4 presents two representative examples where VLMs struggle due to conflicting visual and semantic cues. The first example, *Google* vs. *Goggle*, illustrates a **hard positive**. Although the two names are visually similar, they lie far apart in semantic space. As a result, even with fine-tuning, the VLM may fail to bring them sufficiently close in the embedding space to correctly classify the pair as a spoof, resulting in a false negative. In contrast, the second example, *Doggy* vs. *Puppy*, represents a **hard negative**. Here, the names are semantically similar but visually distinct. Because they are already close in semantic space, the model may fail to separate them adequately, leading to a false positive—incorrectly identifying a non-spoof as a spoof. These cases underscore the difficulty of spoof detection when visual and semantic signals diverge, motivating the addition of non-embedding string methods that integrate strictly textual cues. This divergence becomes especially apparent in Table 2.5.

Table 2.6 shows that spoof and non-spoof pairs differ consistently across all difficulty levels: spoofs have lower Levenshtein distances and higher token set ratios. This distinction is especially pronounced in the hard subset, where embedding models struggle. The spoofs average 1.78 in Levenshtein distance and 82.07 in token set ratio, while non-spoofs average

Table 2.5: Metric profiles of spoof-related examples. Spoofs and hard negatives are visually similar, but differ in string-based metrics.

Example Type	**Cosine Similarity**	**Token Set Ratio**	**Levenshtein Distance**
Spoof	High	High	Low
Hard Negative	High	Low	High
Hard Positive	Low	High	Low

Table 2.6: Average Levenshtein distance and token set ratio across spoof difficulty levels. Spoofs show lower distances and higher ratios.

Dataset	**Spoof**	**Levenshtein Distance**	**Token Set Ratio**
Easy	No	9.57	21.15
Easy	Yes	1.77	82.06
Medium	No	3.54	68.12
Medium	Yes	2.13	77.36
Hard	No	6.18	48.57
Hard	Yes	1.78	82.07

6.18 and 48.57, respectively. These results suggest that string-based metrics remain effective at separating spoof types that evade visual-semantic detection. Thus, to address the limitations of embedding models on visually deceptive cases, an ensemble approach is adopted that integrates string similarity with embedding-based representations, leveraging complementary strengths to improve robustness.

Given a pair of strings (x_1, x_2), the following are computed: (i) $s_{\text{VA-TE}} = \cos(\hat{z}(x_1), \hat{z}(x_2))$ where $\hat{z}(\cdot)$ are L2-normalized VA-TE embeddings; (ii) $r_{\text{TSR}} \in [0, 1]$, the Token Set Ratio (normalized to $[0, 1]$) capturing token overlap similarity; and (iii) $d_{\text{Lev}} \in \mathbb{N}$, the Levenshtein edit distance.[2]

A Gradient Boosting Classifier $g(\cdot)$ is trained on the feature vector $\phi(x_1, x_2) = [s_{\text{VA-TE}},\ r_{\text{TSR}},\ d_{\text{Lev}}]$ to predict $p(\text{spoof} \mid x_1, x_2) = g(\phi(x_1, x_2))$. Unless noted otherwise, $n_{\text{estimators}} = 100$, max_depth $= 6$, learning_rate $= 0.1$, random_state $= 42$ are used. Features are left unscaled (tree-based models are scale-invariant). g is fit on the training split only, the operating threshold τ is selected on the validation split via Youden's J, and metrics are reported on the held-out test set, ensuring data leakage does not occur.

Through this approach, $s_{\text{VA-TE}}$ encodes perceptual similarity learned from contrastive fine-tuning, while r_{TSR} and d_{Lev} are complementary, length-aware and edit-aware textual cues. A learned fusion lets the model weight each cue differently across regimes (string length,

[2]The python-Levenshtein and RapidFuzz's TokenSetRatio libraries are used; distances are computed on raw strings.

token overlap, character-level attacks), yielding a more robust homoglyph detector.

2.3.4 Parameter Tuning and Implementation Details

To effectively train the VA-TE model, several hyperparameters that influence overall model performance and convergence stability are fine-tuned. These include learning rate, batch size, and internal layer size (for linear projection layers), as well as contrastive loss-specific temperature or margin parameters:

- **Margin**, used specifically for pairwise and triplet losses, defines the minimum required separation between negative and positive samples. This way, the model learns separation between difficult data points.
- **Temperature**, relevant only for InfoNCE and SupCon loss, scales similarity scores before softmax normalization. Lower temperatures produce sharper distributions with high contrast between positives and negatives, while higher temperatures produce less separation between contrasting data.

To determine optimal hyperparameters, **Optuna** [3], an optimization library that explores different regions of the hyperparameter space through probabilistic modeling, is used.

Aside from accurate and efficient training, the implementation was designed specifically for scalability and use within prototyping systems. Other systems [17,18] require training from scratch on artificially-created images of text, running into dataset creation issues with inconsistent text sizing, along with scalability problems. These issues are bypassed by building on pretrained VLM text encoders that don't require image inputs. These advantages make the model more deployable in applications that involve text-based visual-similarity analysis. Through use of a lightweight MLP projector and curriculum-based contrastive learning, the implementation lowers costs while simultaneously improving performance beyond other existing methods.

2.4 Experiments

2.4.1 Evaluation Task

To evaluate the effectiveness of VA-TE, homoglyph detection of visually deceptive website domains [17] is framed as the core downstream task. Within this setting, attackers register domains that closely resemble those of legitimate sites, swapping, deleting, or inserting characters to create visually similar and deceptive alternatives. At its core, this evaluation allows determination of whether the model's learned embeddings are effective for textual visual similarity analysis.

2.4.2 Dataset

The publicly available dataset introduced by Woodbridge et al [17] is used, having been constructed to support visual similarity learning for homoglyph detection in both domain

[3]Documenation found at https://optuna.org/

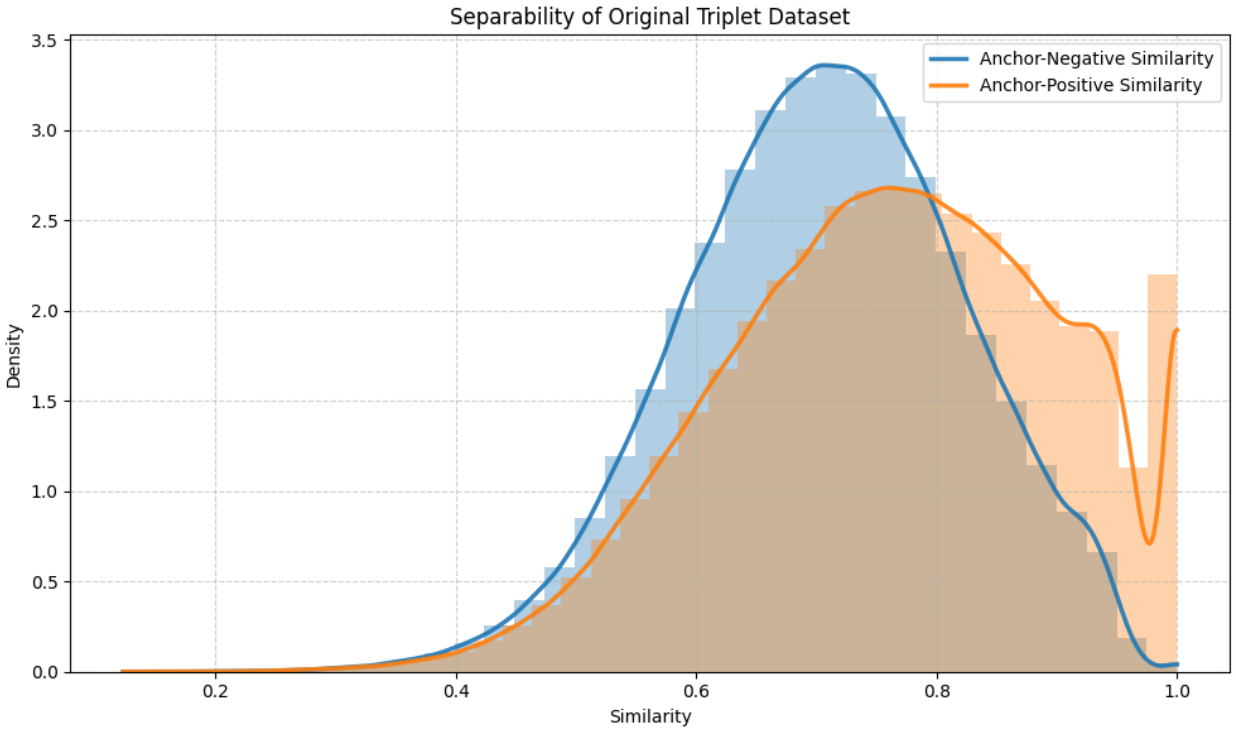


Figure 2.3: Separability of the triplet dataset without any modifications.

and process names. The focus is on the domain name dataset, and training and evaluation are performed using this dataset to benchmark the ensemble approach against prior works.

This original dataset was designed for contrastive learning using a Siamese CNN, where string pairs are rendered as grayscale images and labeled according to whether they represent a spoofing attack. The authors construct spoofed domains using predetermined homoglyph substitutions—targeted character swaps involving visually similar ASCII and Unicode characters. These manipulations yield domain pairs that are lexically distinct yet visually deceptive, reflecting the adversarial nature of homoglyph attacks.

The dataset is divided into three disjoint subsets—training (976,122 examples), validation (51,380 examples), and testing (256,886 examples)—for a total of 1,284,388 labeled pairs. Each example is a tuple containing a real domain name, a candidate spoof or unrelated domain, and a binary spoof label. This dataset is adapted to later generate triplets, along with InfoNCE- and SupCon-compatible examples.

At inference time, the trained model takes in two domain names, generates their embeddings using the projector head, and computes their cosine similarity, which is then used to decide whether the input is a homoglyph attack.

Dataset Manipulations

Woodbridge et al. [17] express domain names in the format `example-domain.com`, introducing a constant ".com" suffix. This fixed token acts as a background signal in the input, possibly reducing the discriminative capabilities of the VLM's pooled embedding and degrading

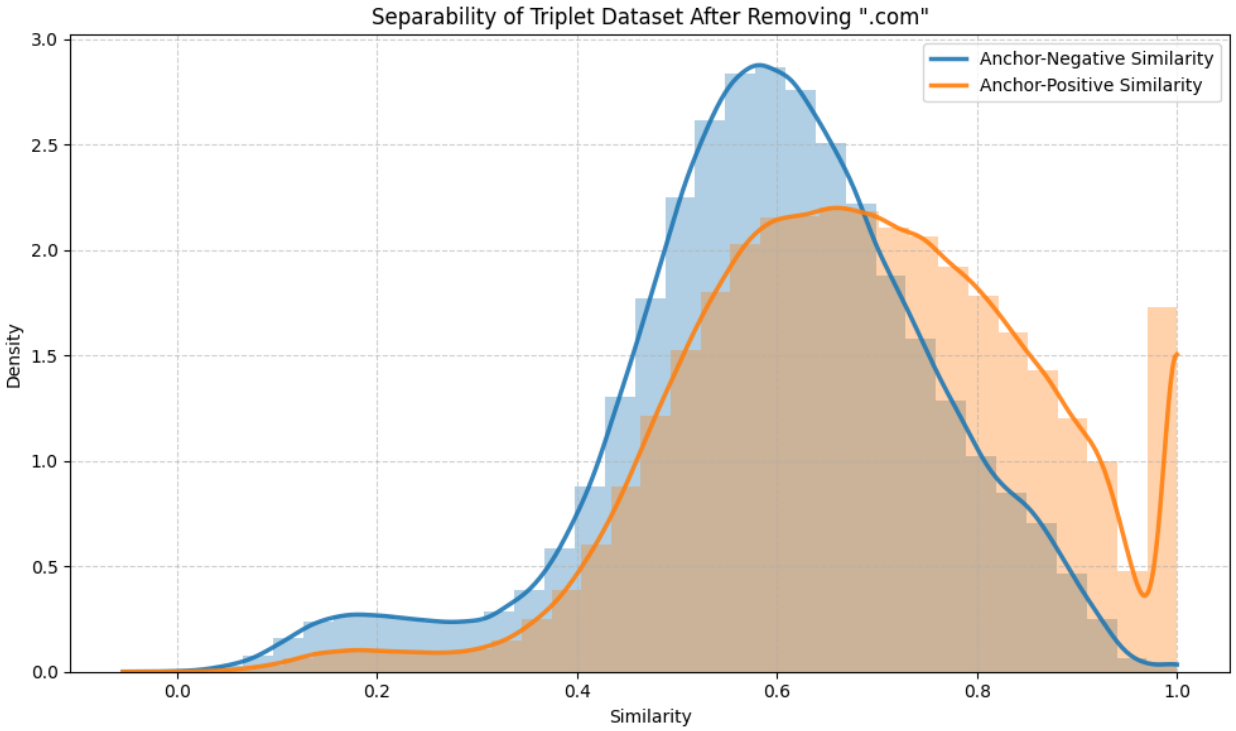


Figure 2.4: Separability of the triplet dataset after ".com" modification.

performance. Based on insights from separability analysis, ".com" is removed from all domain names to mitigate this effect (Table 2.7).

Table 2.7: Example domain transformation after removing `.com`.

Original	`.com` **Removed**
`example-domain.com`	`example-domain`

Analysis indicates greater separability between the positive and negative pairs after removing `.com`.

Visually, Figure 2.3 and Figure 2.4 reveal a shift in the cosine similarity distribution of anchor-negative and anchor-positive examples after removing ".com," with negative embeddings spreading further from the positive set. This shift is quantitatively supported by all three distributional metrics in Table 2.8. Wasserstein distance increased from 0.0568 to 0.0893 (+57%), indicating greater separation between anchor–positive and anchor–negative similarity distributions. KL divergence rose from 0.1205 to 0.1442 (+19.7%), while Jensen–Shannon distance grew from 0.0347 to 0.0394 (+13.5%). Together, these increases suggest improved contrast between positives and negatives post-removal, although the drop in overall similarity values may necessitate adjustment of decision thresholds.

For the curriculum learning experiments, datasets of varying difficulty are created by altering the negative sampling strategy. Difficulty is defined based on the prevalence of hard negatives, samples that are highly similar to the anchor in the input (semantic) embedding

Table 2.8: Removing `.com` on separability of anchor–positive and anchor–negative cosine similarities.

Dataset	KL Divergence	JS Distance	Wasserstein
Original Dataset	0.1205	0.0347	0.0568
`.com` Removed	0.1442	0.0394	0.0893

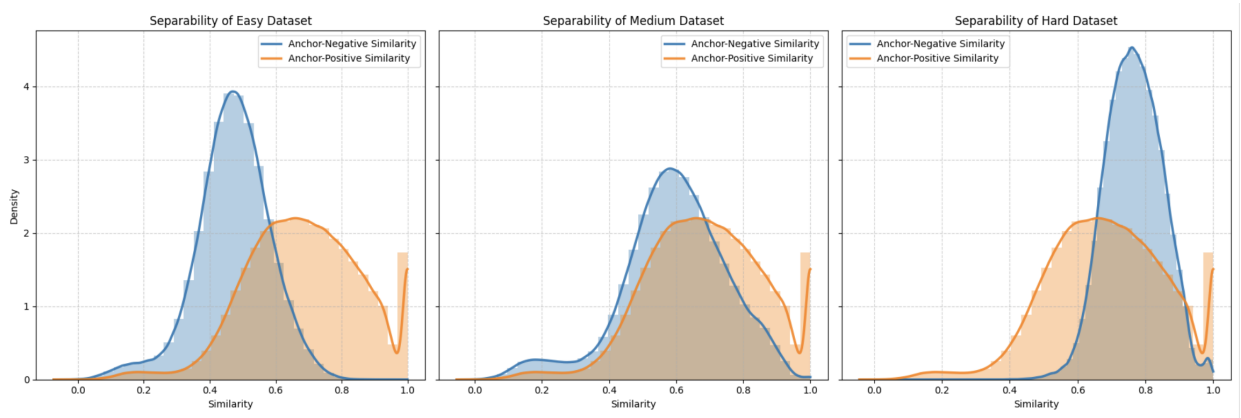


Figure 2.5: Similarity distribution between anchor–negative and anchor–positive pairs across easy, medium, and hard datasets.

space. To quantify this, the cosine similarity distributions of anchor-negative and anchor-positive pairs are compared (see Fig. 2.5), using the latter as a reference to assess the prevalence of hard negatives.

To construct an easy dataset, random negative examples are sampled, which tend to be more distinct from their respective anchors and thus yield a higher baseline separability than the original (medium difficulty) training dataset. For a hard dataset, hard negatives are mined, resulting in a much higher average anchor-negative cosine similarity. Training on these hard negatives forces the model to learn more subtle perceptual distinctions, improving robustness against highly deceptive homoglyph attacks.

After removing ".com" from all domain names, mean negative similarity increases with dataset difficulty—0.4701 for the easy set, 0.5932 for the medium set, and 0.7658 for the hard set (Table 2.9). This trend reflects the growing challenge of distinguishing anchor–negative pairs as difficulty rises, with negatives in the hard set exhibiting similarity scores closer to those of positives—even surpassing them in some cases.

2.4.3 Baseline Experiments

Vision-Language Models

Before applying training methods, several pre-trained VLMs are evaluated to determine which could serve most effectively as the backbone for VA-TE. Though these models were initially trained to capture semantic similarity across image-text pairs, the hypothesis is that their internal representation may already encode relevant perceptual information, particularly

Table 2.9: Mean similarity values for anchor–negative and anchor–positive pairs after removing `.com`, across dataset difficulty levels.

Difficulty	Mean Negative	Mean Positive
Easy	0.4701	0.6825
Medium	0.5932	0.6825
Hard	0.7658	0.6825

regarding the visual appearance of text.

Four prominent VLMs are tested: CLIP, SigLIP, CoCa, and FLAVA. These models were selected to represent a range of architectures and initial training datasets. It is hypothesized that dual-encoder designs, because they encode text and images separately, would be best suited to create VA-TE. For this reason, two different dual-encoder VLMs are tested: CLIP and SigLIP.

- **CLIP (Contrastive Language-Image Pretraining)** [10] uses contrastive loss to train on 400 million image-text pairs, minimizing the distance between matched pairs, and vice versa. Trained with softmax-normalized dot-product similarity and cross-entropy loss, CLIP excels at high-level semantic consistency. It uses a standard Transformer for text encoding purposes.

- **SigLIP (Sigmoid Loss of Language-Image Pretraining)**[50] is an adaptation of the CLIP framework that replaces softmax-based contrastive loss with a sigmoid-based pairwise loss, training the model using a binary cross-entropy over all pairwise matches within a batch. In contrast to CLIP's softmax-based objective, which emphasizes relative similarity across a batch, SigLIP evaluates the relatedness of each image-text pair independently using a sigmoid-based loss. This pairwise formulation allows the model to learn finer-grained distinctions, which may be particularly helpful for preserving perceptual detail relevant to visual-similarity analysis.

Contrastive-based models are intuitively best-positioned for visual-similarity analysis. We verify this by testing select models from other architectures, namely CoCa and FLAVA.

- **CoCa (Contrastive Captioners)**[55] is a hybrid encoder-decoder VLM that blend contrastive pretraining with image captioning objectives, introducing a captioning loss where the model learns to generate the associated text from the visual input.

- **FLAVA (Foundational Language and Vision Alignment)**[58] , a fusion-encoder model, is designed to jointly learn representations from both visual and textual modalities using a transformer backbone. Performing early fusion, concatenating visual and language tokens then processing them together, FLAVA excels at understanding modality interactions.

Each model was evaluated in its frozen, zero-shot form, using its text encoder to generate embeddings for domain names, before being compared to each other using cosine similarity. With this baseline test, the evaluation determines how well each model's native embeddings

map visually-aligned text, without any additional fine-tuning. Selection criteria include ROC AUC, as well as accuracy, recall, and precision at Youden's J.

SigLIP yielded the best baseline results as shown in Table 2.10, indicating its effectiveness for creating VA-TE. Accordingly, the remaining experiments have been conducted using SigLIP embeddings as input.

Table 2.10: *Baseline VLM test results.*

Model Name	ROC AUC	Accuracy	Precision	Recall
CLIP	0.6223	0.5447	0.7517	0.4357
CoCa	0.6798	0.5947	0.8721	0.4330
FLAVA	0.6092	0.5276	0.8806	0.3067
SigLIP	**0.6874**	**0.6083**	**0.7735**	**0.5525**

2.4.4 Fine-Tuning Experiments: Contrastive Loss and Curriculum Learning

The SigLIP text encoder is built upon by training two lightweight linear projection layers to map inputs into the target embedding space. To evaluate the effectiveness of different learning strategies, all pairwise combinations of four contrastive loss functions with each curriculum learning approach are tested. As a control, each loss function is also fine-tuned without curriculum learning using only original training data from [17], isolating the contribution of curriculum design.

2.4.5 Ensemble

After identifying the best VA-TE configuration (contrastive objective + curriculum strategy) using the validation set, the model is frozen and similarity scores $s_{\text{VA-TE}}$ are computed on the training, validation, and test splits. A Gradient Boosting Classifier is then trained using three input features: $s_{\text{VA-TE}}$, Token Set Ratio, and Levenshtein distance. The decision threshold is tuned on the validation set, and test performance is evaluated via ROC-AUC. To assess the benefit of ensembling, ablation results are also reported for using VA-TE alone and string-based features alone.

2.4.6 Evaluation Metrics

The primary evaluation metric is ROC-AUC, a threshold independent value which enables direct comparison with previous works [17,18] and provides a robust summary of the model's ability to distinguish visually similar text. Accuracy, Precision, and Recall are additionally reported to offer a more detailed view of performance under practical performance metrics.

2.5 Results

2.5.1 Baseline Performance

To assess the effectiveness of traditional string matching approaches, Levenshtein Distance and Token Set Ratio were evaluated on the spoof detection task. As shown in Table 2.11, Token Set Ratio slightly outperforms Levenshtein Distance in ROC-AUC (0.8350 vs. 0.8137), indicating the capture of a slightly larger range of visually-similar text. However, these results are not viable compared to existing methods, despite scalability advantages. This motivates the usage of VLMs for better learned representations of visual similarity.

Table 2.11: *Baseline String Matching Results*[4]

Metric	ROC AUC	Accuracy	Precision	Recall
Levenshtein Distance	0.8137	0.8170	0.9046	0.5450
Token Set Ratio	0.8350	0.7616	0.9110	0.6973

2.5.2 VA-TE Performance

Across 20 different VA-TE training configurations, the interaction between contrastive loss functions and curriculum learning strategies was explored. As shown in Table 2.12, the pairwise contrastive loss consistently achieves the highest ROC-AUC scores, reaching 0.95 across nearly all curriculum variations. For triplet loss, scores hovered around 0.91–0.92 regardless of curriculum. InfoNCE and SupCon performed competitively (0.92–0.94), with Automated Curriculum slightly boosting SupCon to 0.94. These results indicate that pairwise loss is robust to sampling difficulty, triplet remains weaker overall, while InfoNCE and SupCon show modest gains from curriculum learning.

Table 2.12: *Validation and test ROC-AUC across VA-TE training strategies.*

Strategy	Pair	Triplet	InfoNCE	SupCon
Medium Negatives	0.95	0.92	0.92	0.92
Hard Negatives	0.94	0.91	0.93	0.93
Manual Curriculum	0.95	0.92	0.93	0.92
Automated Curriculum	0.95	0.93	0.93	0.94
Bandit Curriculum	0.95	0.92	0.92	0.92

Taken together, these findings underscore VA-TE's competitive performance as demonstrated on public benchmarks. Nevertheless, vision-language models alone are limited in

[4]Accuracy, precision, and recall are computed at the Youden's J threshold.

capturing the textual perturbations common in homoglyph attacks. This motivates an ensemble approach, which corroborates VA-TE similarity scores with Levenshtein distance and fuzzy string matching.

2.5.3 Ensembles

To evaluate the benefits of combining vision-language and string-based features, an ensemble model was constructed using the best VA-TE configuration alongside traditional string similarity metrics. As detailed in Table 2.13, the ensemble achieved a ROC-AUC of 0.98, outperforming both the VA-TE model alone (0.95) and the string-based metrics alone (0.89). This significant improvement highlights the complementary strengths of visual-semantic embeddings and token-level string comparisons, especially in detecting visually deceptive text where either modality alone may be insufficient.

Table 2.13: *ROC-AUC for Ensemble Strategies*. "String Metrics" refers to the combination of Edit Distance and Token Set Ratio.

Ensemble Strategy	**ROC AUC**
String Metrics	0.89
VA-TE + String Metrics	0.98

2.5.4 Comparison to State-of-the-Art

Compared to prior SOTA methods, the VA-TE framework alone achieves highly competitive results while maintaining a lightweight and scalable architecture. As shown in Table 2.14, the VA-TE ensemble attains strong performance on the evaluated dataset with a ROC-AUC of 0.98, reflecting promising research outcomes rather than production-level guarantees. Furthermore, while other deep learning approaches require substantial training time and preprocessing, VA-TE benefits from compact linear projection heads over pretrained encoders, supporting efficient experimentation without implying readiness for deployment. These results demonstrate that a semantically grounded representation can be effectively translated into a visually aligned embedding space, enabling high performance on downstream tasks.

Table 2.14: *Comparison of proposed methods to state-of-the-art models on the ROC-AUC metric.*

Method	**ROC AUC**
VA-TE	0.95
VA-TE + String Metrics	0.98
Siamese-GRU [18]	0.98
Siamese-LSTM [18]	0.97
Siamese-CNN [17]	0.97
Siamese-CNN [18]	0.93

2.6 Discussion

2.6.1 Performance Analysis

Experiments highlight important insights into the efficacy of VA-TE and the role of different training configurations.

Across all settings, pairwise contrastive loss consistently achieved the strongest performance, reaching a ROC-AUC of 0.95. This result is consistent with the SigLIP encoder's original pairwise training objective [50] and avoids the margin sensitivity issues that hinder alternative loss formulations. In particular, triplet loss plateaued around 0.91-0.92, underperforming pairwise. This is likely due to difficulties in margin tuning and limited batch sizes restricting effective negative sampling. InfoNCE and SupCon achieved competitive results, but showed greater dependence on curriculum learning to achieve optimal performance.

The impact of curriculum learning varied according to the loss function used. For pairwise and triplet losses, curriculum learning yielded negligible improvements, suggesting that these objectives are relatively robust to changes in negative sampling difficulty. In contrast, InfoNCE and SupCon gained modest but consistent benefits from Automated Curriculum, while manual and bandit-based approaches showed little added value. Across the four loss functions, the Automated Curriculum was consistently the strongest performer, matching the best result or achieving the top score outright, providing smoother transitions and more stable convergence than alternative curricula (as seen in Table 2.12).

When compared against prior benchmarks, VA-TE achieved its best standalone performance with pairwise loss and automated curriculum (ROC-AUC = 0.95). More importantly, when fused with string-based similarity metrics such as Levenshtein distance and Token Set Ratio, the ensemble attained strong performance with a ROC-AUC of 0.98. This surpasses or matches the strongest image-based approaches, including Siamese-CNN and Siamese-GRU models [17,18]. Unlike those methods, VA-TE operates directly on text rather than rendered images, providing substantial advantages in scalability, memory efficiency, and operational simplicity for research workflows. These findings underscore both the robustness of pairwise contrastive learning in visually-aligned text embeddings and the complementary value of

combining embedding-based and string-based features.

2.6.2 Advantages of VA-TE

Scalability Benefits over Image-Based Approaches

VA-TE offers substantial advantages over existing image-based homoglyph detection methods in terms of computational efficiency and memory requirements. Previous state-of-the-art approaches [17,18] require rendering each string as a binary image before training and inference. Specifically, Woodbridge et al. [17] render strings as 150×12 pixel grayscale images. The associated datasets contain 1,284,388 labeled pairs, requiring the generation and storage of over 2.5 million images. In contrast, VA-TE operates directly on text strings, eliminating image rendering entirely and reducing memory overhead by orders of magnitude.

Beyond storage considerations, image-based methods introduce additional computational overhead during both training and inference workloads. Each inference requires: 1) String-to-image rendering using specific fonts and sizing parameters, 2) Image preprocessing and normalization, and 3) Forward pass through a convolutional neural network designed for image input. VA-TE eliminates steps 1 and 2 entirely, requiring only a lightweight forward pass through the frozen VLM text encoder followed by the projection head. This streamlined pipeline supports low-latency processing in settings where computational efficiency is important, without implying deployment in any specific operational environment.

Deployment Simplicity and Robustness

VA-TE's text-only design significantly simplifies potential integration into systems by avoiding the rendering used by image-based methods, which depend on parameters like font choice, size, anti-aliasing, and background color. Inconsistencies in these parameters can degrade model performance and require pipeline maintenance. By relying on pretrained VLM text encoders, VA-TE avoids these issues entirely, enabling consistent performance across environments in research or prototype contexts without suggesting production deployment.

Additionally, image-based approaches impose artificial constraints on input string length due to fixed image dimensions. CNN-based models often require truncating or padding strings to fit a predetermined image size (e.g., the 25-character limit in [17]), introducing a source of information loss or distortion. By contrast, VA-TE can process strings of arbitrary length limited only by the underlying VLM's context window, typically accommodating hundreds of characters.

Furthermore, VA-TE leverages the robustness of large-scale pretrained models. SigLIP and similar VLMs are trained on massive, diverse datasets and demonstrate strong generalization capabilities. By building on these established representations rather than training from scratch on synthetically generated images, VA-TE inherits the robustness and stability of the underlying foundation model, models, positioning it as a flexible approach suitable for exploratory, analytic, or prototype-oriented use cases.

2.6.3 Future Work

Application to Different Text Domains

Extending beyond English, visually deceptive text persists in other, more complex writing systems. Detecting homoglyphs in languages such as Chinese, Arabic, or Cyrillic would require more fine-grained character distinctions and introduces additional challenges, including font-based homographs and diacritic variations.

Alternatively, VA-TE can serve as a verification layer for Optical Character Recognition (OCR) pipelines, particularly in document digitization. Recent literature has shown that OCR systems are susceptible to misclassifying visually similar characters—particularly in difficult-to-distinguish languages like Arabic [59]. Embedding-based homoglyph detection could identify and automatically correct these OCR-induced misclassifications, providing vast applications in security-critical contexts, including bank check processing, passport scanning, and historical document transcription.

Enhancing Text Embeddings via Perceptual Decoding

While the current approach leverages the joint text-image embedding space of pretrained VLMs, it relies exclusively on the *text encoder* to generate representations of input strings. These embeddings, though semantically rich and convenient for scalable prototyping, are aligned with the final-layer outputs of the VLM's text encoder and therefore may not capture perceptual features necessary for fine-grained homoglyph detection.

To address this limitation, an architectural extension is proposed that grounds text embeddings in low- and mid-level visual features extracted from rendered images of the input text. Specifically, a lightweight decoder $f_d : \mathbb{R}^{d_{text}} \to \mathbb{R}^{d_{vis}}$ could be introduced, trained to map text embeddings to multi-scale image features, extracted from shallow, intermediate, and deeps layers of the VLM's vision encoder. During training, this decoder learns to reconstruct visual features from text embeddings alone, encouraging the text representations to encode subtle typographic details. Crucially, once trained, the decoder becomes part of the permanent inference pipeline, transforming text embeddings into visually-grounded representations without requiring image inputs at inference. Training can be approached in two ways:

1. **Preprocessing module:** Pre-train the encoder to enhance the text embedding space prior to standard contrastive training.
2. **Joint Optimization:** Simultaneously train the decoder and contrastive objectives, allowing the text encoder to adapt its representations for both visual reconstruction and similarity learning.

This formulation offers a promising path for improving visual discrimination in spoof detection while preserving the efficiency and simplicity of text-only inference.

Enhanced Model Architectures and Training Strategies

Future architectures could benefit from delayed weight update strategies, where gradients are accumulated over multiple batches before parameter updates [60]. Given the relatively

lightweight projection head (two linear layers) and the noisy nature of contrastive learning objectives [19], gradient accumulation could provide more stable learning signals and improved global optimization trajectories [61]. This approach would be particularly valuable when computational constraints limit batch sizes, allowing the model to approximate the benefits of larger-batch training without additional memory overhead.

Chapter 3

Learning What Is Real: Intrinsic Authenticity Detection for Generalization Across Deepfake Methods

3.1 Introduction

Deepfake generation technology has been advancing at an unprecedented pace, posing significant threats to systems dependent on identity verification and public trust. Early deepfakes relied on simple face-swapping techniques using Generative Adversarial Networks (GANs), but modern diffusion models are now able to produce synthetic media with photorealistic quality that is increasingly difficult to distinguish from authentic content. This rapid evolution in generation capabilities and methods has exposed critical vulnerabilities in existing detection approaches.

Current deepfake detection methods face two fundamental limitations. First, many approaches train exclusively on synthetic data generated by a single model, learning to identify artifacts specific to that generation method [62]. While this strategy achieves reasonable performance on data from within the training distribution, it drastically underperforms when confronted with novel generation techniques. The accelerating pace of advancement in this field poses a critical weakness for identifying authentic media. Second, alternative approaches attempt to leverage embeddings from various modalities, namely audio and video features, to distinguish real from synthetic media [62]. However, these learned representations predominantly encode content-related information rather than the subtle artifacts and inconsistencies that reliably indicate synthetic generation.

These limitations motivate the approach: disentangling authenticity-related features from those encoding identity and content within pretrained representations, isolating the signals most relevant for deepfake detection. By explicitly separating these feature types, the goal is to reduce the influence of extraneous information, which may be similar across both real and fake media, while amplifying detection-relevant signals that generalize across generation methods. The central hypothesis is that disentanglement will enable learned representations to cluster authentic and synthetic media into distinct regions of a latent space, achieving robust generalization to previously unseen generation techniques. This approach addresses both

the overfitting problem of technique-specific detectors and the content-encoding limitation of existing representation-based methods, offering a more principled path toward reliable, generalizable deepfake detection.

3.2 Related Works

3.2.1 Deepfake Generation

Generative Adversarial Networks

Along with other developments in AI, synthetic media generation has rapidly evolved, with each subsequent generative model family producing increasingly realistic samples. Early deepfake methods relied on Generative Adversarial Networks (GANs), which introduced an adversarial generator-discriminator framework that quickly became the leading paradigm for synthetic media creation [63]. GAN models were soon applied across modalities, including audio, where early works like WaveNet demonstrated that adversarially trained architectures could generate highly realistic audio samples [64]. Significant progress in GANs was driven by Karras and colleagues through a sequence of architectural innovations. ProGAN [65] iteratively increases the size of both the generator and discriminator, which stabilizes training and enables high-resolution face syntheses with improved diversity. StyleGAN [66] reorganized the generator around a style-based architecture that first maps the input latent embedding into an improved latent space and then injects this representation into each layer of the generator. This allows the generator model to naturally disentangle coarse structure, mid-level semantic features, and fine-grained texture details. Parallel to these advances in generative modeling, face-swap models such as Face2Face [67] and SimSwap [68] focused on identity transfer rather than full-frame synthesis, enabling real-time and high-quality manipulations by conditioning on target facial expressions and landmarks. Despite these advances, GAN-generated media often contains recognizable artifacts that remain detectable by modern detection systems [69]. These limitations, combined with the training instability of GANs, helped motivate the shift toward diffusion-based models that offer greater stability and consistently produce more realistic synthetic media.

Diffusion Models

Diffusion models, first introduced in 2015 [70] based on ideas from statistical physics, function by gradually adding noise to data and then learning a reverse denoising process to recover the underlying signal. This procedure allows diffusion models to more accurately model the full data distribution and avoid the mode-collapse issues of GANs. By 2020-2021, diffusion models emerged as the state-of-the-art class of generative models with the invention of foundational architectures such as Denoising Diffusion Probabilistic Models (DDPMs), which showed dramatic improvements in image quality and diversity [71]. Compared to GANs, diffusion models produce fewer artifacts, and instead suffer from issues like temporal inconsistency in video outputs [69]. Current diffusion-based video generation models, including OpenAI's Sora and Google's Veo3, now produce photorealistic video and audio, motivating the development of detection techniques that remain robust to rapidly advancing generative models.

3.2.2 Deepfake Detection Approaches

Existing deepfake detection methods can be broadly categorized into forensic and machine learning based approaches, each with distinct strengths and limitations [69].

Forensic Methods

Forensic techniques use signal processing, hand-crafted features, and domain knowledge to detect manipulation artifacts. Forensic techniques typically analyze frequency domain statistics and resolution inconsistencies [72], along with surface warping [73], inconsistent lighting, and other errors resembling those found in related video-processing tasks such as face tracking and editing [74]. Biometric-based forensic methods exploit the inability of generators to accurately model biological signals, such as heart-rate patterns [75–77], lip-sync mismatches [78], and joint jaw-ear motion dynamics [79]. While effective when visible artifacts are present, forensic methods fail on modern generators that produce high-quality, high-fidelity outputs with minimal artifacts. Forensic methods are also vulnerable to video compression and low-resolution inputs, which disrupt or erase the subtle forensic signals these methods rely on [69].

Machine Learning Methods

Machine learning–based detectors move beyond forensic cues by learning discriminative features directly from data. By operating in latent spaces that capture richer semantic and structural information, these models can leverage large datasets, reduce information loss, and capture subtle inconsistencies that may not be visible in pixel or frequency space features. Early CNN-based detectors such as XceptionNet [80] achieved strong performance by capturing the spatial features of manipulation artifacts. Subsequent works introduced attention mechanisms to emphasize both localized and global cues [81], as well as temporal modeling via 3D CNNs, ConvLSTMs, and transformer-based encoders to detect temporal inconsistencies across frames [82]. More recently, attention-based multimodal fusion approaches have leveraged contextual cross-modal information to improve both detection accuracy and temporal localization of manipulated segments [83].

As deepfakes have improved in quality, unimodal visual analysis has proven insufficient, leading to multimodal methods that jointly model audio and video correlations. Mittal et al. [84] pioneered the use of affective cues, demonstrating that inconsistencies between perceived emotions in audio and video modalities provide robust detection signals. More recent work has combined audio-visual attention with contrastive learning to improve generalization across manipulation types [85]. Hybrid approaches have also emerged that combine learned representations with forensic priors; for instance, NoiseDF [86] leverages noise trace analysis alongside multi-head attention to capture discrepancies between manipulated facial regions and authentic backgrounds. However, systematic evaluation on multimodal deepfake datasets has revealed that purely multimodal baselines often underperform ensemble-based approaches, suggesting that naively fusing heterogeneous modalities does not guarantee improved detection [62].

Despite these advances, ML-based detectors still struggle to generalize to unseen generative models [87]. Recent work has explored training-free approaches that leverage fact-checking

principles to detect zero-day attacks without requiring exposure to specific deepfake types [88], while unsupervised methods using contrastive learning with pseudo-labels have shown promise for detection without ground-truth supervision [89]. A key reason for poor generalization is representation entanglement: identity, semantic content, temporal patterns, and authenticity cues are all mixed together within the embedding space. This entanglement encourages detection models to rely on identity-specific or dataset-specific shortcuts, degrading performance under distribution shift and novel generation methods. This limitation motivates methods that explicitly disentangle authenticity-related information from identity and other confounding factors.

3.2.3 Disentangled Representation Learning

While disentanglement has been widely studied in generative modeling [90–92], its application to deepfake detection remains limited. Classical formulations such as $\beta-$VAE [91] and InfoGAN [92] demonstrate that enforcing independence across latent dimensions encourages models to recover underlying generative factors such as pose, identity, and illumination. More recent representation learning methods, including variance–invariance-covariance regularization [93], further highlight the importance of decorrelation, invariance, and non-collapse properties in producing robust latent spaces.

In discriminative settings, separating task-relevant signals from confounding factors improves robustness and reduces reliance on dataset-specific or identity-specific correlations. This is especially applicable to deepfake detection, where standard models often entangle who appears in the video with whether the video is authentic and simultaneously overfit to the specific generative models present in their training set. As a result, detectors rely on identity shortcuts or generator-specific artifacts rather than manipulation-invariant cues, leading to significant failures under distribution shift. Recent anomaly detection approaches address this by decomposing latent feature spaces into homogeneous (consistent across all real images) and heterogeneous (image-specific) components [87]. These methods generalize better to unseen generative models because these methods learn intrinsic properties of real images rather than generator-dependent artifacts. However, these methods do not disentangle identity information from authenticity information, leaving a critical source of shortcut learning unaddressed.

This approach directly targets this gap through disentangled, real-video–driven pre-training. The approach leverages the assumption that all real videos share a common authenticity signal that is independent of content, identity, or scene variations. To isolate this signal, pre-training is conducted exclusively on real videos and two complementary latent subspaces are learned:

1. A projection head maps embeddings of real videos into a low-variance manifold capturing the statistical commonalities shared across all authentic footage. This manifold represents the intrinsic "realness" structure of authentic spatiotemporal data.

2. A second projection head, constrained to be orthogonal to the authenticity-preserving one, captures all remaining variation (identity, semantics, motion style, context). This subspace is trained contrastively on augmentations of the same video to ensure it captures expressive, person-specific variation.

The orthogonality constraint minimizes mutual information between the two latent codes, encouraging the authenticity branch to encode only manipulation-relevant signals, and the identity branch to absorb variation unrelated to authenticity. At inference time, the authenticity-preserving projection head serves as an anomaly detector. Real videos project close to the learned low-variance manifold, while synthetic videos, which lack the intrinsic real-video commonalities, deviate sharply from it. This combines principles from disentangled representation learning, anomaly detection, and commonality learning to model authenticity directly, without needing examples of every manipulation type. Finally, to evaluate the learned spaces, classifier-free embedding metrics are adopted to assess the quality of disentanglement and robustness without relying solely on downstream classifier tuning. These metrics allow for fair comparison across representation spaces and directly reflect the geometric structure induced by the method.

3.3 Methods

3.3.1 Overview: Disentangled Representation Learning Framework

Deepfake detection is approached through unsupervised representation learning that disentangles authenticity-related signals from content and identity features within pretrained audio-visual embeddings. Traditional end-to-end detection systems couple representation learning with classifier optimization, requiring sequential tuning of two interdependent components: first learning an embedding space, then training a classifier on top of it. If the learned representations fail to capture meaningful structure, no amount of classifier tuning can recover performance. Moreover, evaluating such systems requires downstream classification, conflating representation quality with classifier capacity.

Instead, a self-supervised evaluation framework is adopted that directly assesses the geometric properties and intrinsic separability of the learned embedding space. The hypothesis is that if authentic and synthetic media are fundamentally different in their intrinsic properties, this difference should manifest as natural clustering in a well-designed representation space, observable through unsupervised metrics (cluster coherence, distributional divergence, separation gaps) without requiring classifier training. This approach isolates representation learning from classification, enabling principled diagnosis of where detection systems succeed or fail.

The framework consists of a frozen pretrained encoder f_{enc} followed by two trainable projection heads that map embeddings into complementary subspaces: an authenticity projection f_{auth} trained via variance minimization to capture properties shared across all real samples, and an identity projection f_{id} trained via prototypical contrastive learning to encode content-specific information. An orthogonality constraint enforces independence between these subspaces, preventing the authenticity head from encoding identity shortcuts that would harm generalization to unseen subjects.

The key challenge in this approach is preventing representation collapse during multi-objective optimization. Naive variance minimization can produce degenerate solutions where all embeddings (real and fake) collapse to a single point, satisfying the variance objective while destroying class separability. This is addressed through a variance floor regularization

strategy that maintains minimum embedding spread while still encouraging tight clustering of authentic samples.

3.3.2 Disentangled Representation Learning

Dual Projection Heads

The representation learning architecture consists of a shared encoder f_{enc} followed by two projection heads, f_{auth} and f_{id}. Given an input frame x, the encoder produces an embedding $z = f_{\text{enc}}(x)$, which is projected into two complementary subspaces:

$$z^{\text{auth}} = f_{\text{auth}}(z) \quad : \text{captures authenticity-related signals.} \tag{3.1}$$

$$z^{\text{id}} = f_{\text{id}}(z) \quad : \text{captures identity and content information.} \tag{3.2}$$

The encoder f_{enc} may be any pretrained visual or audio-visual model (Section 3.4.2). Both projection heads are implemented as two-layer MLPs with ReLU activations and output dimension $d = 128$.

This dual-head structure enables explicit supervision of each subspace using tailored objectives, while the shared encoder ensures computational efficiency and a unified representation from which both identity and authenticity can be decoded. The design is motivated in part by recent work showing that separating complementary factors in the representation space can improve generalization in media forensics [87].

Orthogonality Constraint

To enforce independence between identity and authenticity representations, an orthogonality constraint is introduced that penalizes correlation between z^{id} and z^{auth}. Formally, for a batch of N samples, minimization occurs for:

$$\mathcal{L}_{\text{orth}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \text{sim}\left(z_i^{\text{id}}, z_j^{\text{auth}}\right) \right| \tag{3.3}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This constraint draws on the information bottleneck principle [94] and decorrelated representation learning methods. Specifically, Barlow Twins [95] introduced cross-correlation regularization to minimize redundancy between embedding dimensions, while VICReg [93] extended this with explicit variance and covariance terms. The orthogonality loss adapts these principles to enforce decorrelation between *separate* embedding subspaces rather than within a single representation, encouraging the two projection heads to extract complementary, non-redundant information from the shared encoder.

This decorrelation is critical for generalization: if identity and authenticity are orthogonal in the learned embedding space, then knowing someone's identity provides no information about whether their video is real or fake. A model that disentangles these factors should not memorize identity-specific deepfake patterns (e.g., "person A is always fake in the training set") but instead learn identity-agnostic authenticity cues.

Training Objectives

The dual projection heads are trained with two complementary objectives: (1) a prototypical contrastive loss for identity learning, and (2) a variance minimization loss for authenticity learning.

Identity Learning via Prototypical Contrastive Loss The identity head f_{id} is trained to cluster embeddings based on content similarity, regardless of authenticity. The approach leverages the augmentation structure of the datasets: in AVDeepfake-1M++, multiple augmented versions of the same source video share identical content but differ in noise/perturbations, while in ShareVeo3, consecutive frames within a video share semantic content.

Prototypical contrastive learning is adopted to make training computationally efficient ($O(n)$ rather than $O(n^2)$ pairwise comparisons in standard contrastive learning). For each "content group" (augmentations of the same source video or frames from the same video), a prototype is computed as the mean embedding:

$$c_k = \frac{1}{|G_k|} \sum_{i \in G_k} z_i^{\text{id}} \tag{3.4}$$

where G_k is the set of samples belonging to content group k. The prototypical loss encourages embeddings to be close to their assigned prototype and far from other prototypes:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{\exp\left(-d(z_i^{\text{id}}, c_k)/\tau\right)}{\sum_j \exp\left(-d(z_i^{\text{id}}, c_j)/\tau\right)} \right] \tag{3.5}$$

where $d(a, b)$ is Euclidean distance, τ is a temperature hyperparameter, and the sum is over all prototypes in the batch.

Authenticity Learning via Variance Minimization The authenticity head f_{auth} is trained using an anomaly detection approach: training occurs only on real videos, encouraging their embeddings to form a compact cluster in the z^{auth} space. The intuition is that real videos share common authenticity characteristics (e.g., natural sensor noise, consistent photometric properties), whereas fake videos exhibit diverse artifacts depending on the generation method. By minimizing variance among real embeddings, the aim is to create a compact "real manifold" that authentic samples occupy, while fake samples—lacking these shared intrinsic properties—deviate from this manifold.

However, naive variance minimization presents a critical challenge: the objective cannot distinguish between the intended outcome (a compact real cluster with room for fakes to scatter) and a degenerate collapsed state (all embeddings mapping to a single point). When combined with the prototypical loss, which pulls content groups together regardless of authenticity, the optimization can find a trivial solution where both real and fake embeddings collapse to a small region, satisfying the variance objective while destroying class separability.

To address this, a variance floor regularization is introduced that maintains minimum embedding spread while still encouraging tight clustering. For a batch of real samples

(authenticity label = 1), minimization occurs for:

$$\mathcal{L}_{\text{var}} = \underbrace{\frac{1}{N_{\text{real}}} \sum_{i:y_i=1} \left\| z_i^{\text{auth}} - \mu_{\text{real}} \right\|^2}_{\text{variance minimization}} + \lambda_{\text{reg}} \underbrace{\left[\max(0, \tau - \sigma^2)^2 + 5 \cdot \max(0, \tau - \sigma^2) \right]}_{\text{variance floor regularization}} \tag{3.6}$$

where μ_{real} is the mean real embedding in the batch, σ^2 is the empirical variance, τ is a minimum variance threshold, and λ_{reg} is the regularization weight. The regularization term combines quadratic and linear penalties: the quadratic term $\max(0, \tau - \sigma^2)^2$ provides smooth gradients when variance approaches the floor, while the linear term $5 \cdot \max(0, \tau - \sigma^2)$ provides stronger gradients when variance falls significantly below the threshold, preventing complete collapse.

This formulation allows the model to learn tight real clusters (minimizing the first term) while maintaining sufficient embedding space for fakes to occupy distinct regions (enforced by the second term). The variance floor τ acts as a hyperparameter controlling the minimum allowable spread of real embeddings, with larger values encouraging more conservative clustering and smaller values allowing tighter compression.

Joint Optimization The total loss for the representation learning stage combines three objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{var}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} \tag{3.7}$$

where λ_{orth} weights the orthogonality constraint. Each loss is normalized by its initial value to ensure similar magnitudes across objectives, then equal unit weights are applied (1.0 for prototype, 1.0 for variance, $\lambda_{\text{orth}} = 0.1$ for orthogonality). This normalization strategy proved more stable than gradient-based balancing methods [3] in preliminary experiments, as it avoids the interaction between adaptive weight scheduling and variance floor regularization.

3.4 Experiments

3.4.1 Datasets

The method is trained and evaluated using three complementary datasets: AVDeepfake-1M++, ShareVeo3, and Sora2, which together provide diverse coverage of deepfake generation methods and manipulation scenarios. AVDeepfake-1M++ focuses on face-swap and voice-cloning manipulations applied to real source footage, while ShareVeo3 and OpenAI Sora2 represent end-to-end text-to-video generation models that synthesize entire videos from scratch. These generative models pose a particularly severe threat to identity verification systems, as videos can be generated targeting specific faces and voices from minimal training data, often just a few seconds of footage or publicly available social media content. ShareVeo3 is used for training to expose the model to diffusion generation, and out-of-distribution generalization is evaluated on Sora2, a state-of-the-art video generation model not represented in the training data [96].

For all datasets, videos and audio are segmented into 0.15-second segments, and audio and video embeddings are generated for each segment using pretrained encoders. These

embeddings are stored in a serverless PostgreSQL database for efficient retrieval during training and evaluation. The rationale for this temporal segmentation granularity is discussed in Section 3.4.2, where the trade-off between temporal resolution and computational efficiency is analyzed. All datasets used are public or publicly accessible; no proprietary, customer, or confidential corporate data was used

AVDeepfake-1M++

AVDeepfake-1M++ [1] is a large-scale audio-visual deepfake dataset containing over 2 million videos spanning 7,109 unique subjects across training, validation, and test splits. The dataset employs diverse generation methods for both visual and audio manipulation, including:

- **Visual deepfakes:** WAV2LIP (100% of visual manipulations in base dataset), TalkLip (100% in base), and an expanded set in AVDeepfake-1M++ including TalkLip (65.9%), Diff2Lip (22.9%), and LatentSync (11.2%)
- **Audio deepfakes:** SV2TTS (100% in base dataset), VITS (66.6% in base), YourTTS (33.4% in base), and an expanded set in AVDeepfake-1M++ including VITS (38.8%), YourTTS (38.0%), F5TTS (9.9%), and XTTSv2 (13.2%)

From this large-scale dataset, the training subset is constructed by selecting 241 source videos from the training split, each with approximately 10 augmented versions created through synthetic perturbations. The preprocessing pipeline extracts embeddings at 0.15 second intervals, yielding 231,889 total embeddings. This subset has two key structural properties that are leveraged for representation learning:

- **Multiple augmentations per source:** Each source video is augmented with diverse perturbations including Gaussian blur, salt-and-pepper noise, compression artifacts, color quantization, motion blur, and others. Crucially, all augmentations of the same source share identical underlying content (same subject and audio-visual semantics), differing only in low-level signal properties induced by perturbations and authenticity signals. This provides strong supervision for identity learning: augmented versions of the same source should cluster tightly in z^{id} space regardless of their authenticity label or applied perturbations.
- **Partial fake injections:** Fake videos in AVDeepfake-1M++ are created by injecting synthesized audio or video segments into real source material, with temporal injection boundaries annotated in the video metadata. This structure provides fine-grained authenticity labels at the segment level.

After preprocessing and embedding extraction, the training subset contains: 225,843 fully real embeddings (covering original source videos and real augmentations with perturbations but no deepfake manipulation), 2,394 fully fake embeddings (videos with complete audio and/or visual synthesis throughout), and 3,652 partial fake embeddings (videos with temporal injection, where only segments are manipulated). Class imbalance is addressed through balanced batching during training: each batch contains equal numbers of real and fake

samples, sampled uniformly across content groups (where a content group is defined as all augmentations derived from the same source video).

For prototypical learning, content groups are defined by source video identity. All augmentations of the same source video, regardless of perturbation type or authenticity, belong to the same content group and should share similar z^{id} representations. All available augmentations are used rather than subsampling to maximize the diversity of perturbations during training, which improves robustness.

ShareVeo3

ShareVeo3 [2] is a fully synthetic video dataset containing 1,460 videos generated by Google Veo3 [97]. Unlike AVDeepfake-1M, ShareVeo3 videos are entirely fake and lack explicit augmentations from a single source video. However, temporal continuity within each video is leveraged: consecutive frames share semantic content (same scene, same subject) and can be treated as a content group for identity learning. ShareVeo3 serves two purposes: (1) it provides additional fake training data from a different generation method, reducing overfitting to AVDeepfake-1M's specific artifacts, and (2) it enables evaluation on fully synthetic videos, which differ from partial injection attacks. A total of 63,184 embeddings are generated from 1,460 fully fake videos (no real samples, no partial fakes). During training, ShareVeo3 fakes are paired with real samples from AVDeepfake-1M++ in balanced batches.

Sora2 (Out-of-Distribution Evaluation)

To evaluate the model's ability to generalize to future generative methods, an out-of-distribution test set is constructed from OpenAI's Sora2 [96], widely considered the current state-of-the-art in video generation. The top 150 videos are collected from the Sora2 app's search feed, yielding 11,317 embeddings after preprocessing at 0.15-second intervals.

This dataset serves as the critical test of the method's robustness to novel generation methods. The hypothesis is that by training the authenticity projection head exclusively on real videos to learn a low-variance manifold of intrinsic real-footage properties (Section 3.3.2), the model naturally rejects Sora2 videos as anomalies because they deviate from the shared statistical regularities that define authentic video data. Crucially, the model has not seen examples from this generation method during training, making this a pure test of generalization to unseen deepfake techniques.

To characterize the distribution of the Sora2 test set, each video is manually annotated across multiple dimensions. Table 3.1 summarizes the metadata distribution.

This metadata distribution reveals that the Sora2 test set predominantly contains human-centric, naturalistic content with visible faces and speech, characteristics aligned with the most critical deepfake detection scenarios. However, the substantial inclusion of non-human content (12.7%), stylized and surreal videos (15.3% combined), and diverse cinematographic conditions (varying camera angles, motion, and lighting) is crucial for validating the approach. Strong performance across this heterogeneous content distribution indicates that the method learns general intrinsic properties of authentic video data, rather than face-specific forensic artifacts. This distinction is critical: traditional deepfake detectors perform latent forensics, identifying traces left by specific manipulation algorithms. By contrast, the approach performs

Table 3.1: Sora2 test set metadata distribution across 150 videos.

Category	**Attribute**	**Count (%)**
Human Presence	Contains humans	115 (76.7%)
	Visible faces	110 (73.3%)
Number of People	0 people	39 (26.0%)
	1 person	75 (50.0%)
	2 people	17 (11.3%)
	3 people	4 (2.7%)
	4 people	4 (2.7%)
	5+ people	19 (12.7%)
Speech	Talking subjects	106 (70.7%)
Camera Shot	Close-up	34 (22.7%)
	Medium	87 (58.0%)
	Wide	29 (19.3%)
Camera Motion	Static	41 (27.3%)
	Slight movement	71 (47.3%)
	Heavy motion	38 (25.3%)
Lighting	Good lighting	119 (79.3%)
	Low-light	31 (20.7%)
Content Realism	Natural/Photorealistic	108 (72.0%)
	Stylized	19 (12.7%)
	Surreal	23 (15.3%)

a deeper *latent analysis*, characterizing the shared structure of real videos across arbitrary content.

Importantly, Sora2 videos contain no real samples or partial injections; all 11,517 embeddings are fully synthetic. During evaluation, these fake embeddings are paired with real samples from AVDeepfake-1M++ to compute detection metrics.

3.4.2 Embedding Selection

Before training the disentanglement model, a systematic evaluation is conducted to identify which pretrained audio and video embeddings provide the strongest baseline separability between real and fake samples. Candidate embeddings are evaluated using 5-fold cross-validation on a balanced subset of AVDeepfake1M++, training simple logistic regression classifiers on each embedding type independently. This analysis informs the choice of input representations for the full pipeline.

Audio Embeddings

Four audio embedding approaches are evaluated: OpenL3 [98], HuBERT [99], Wav2Vec2 [100], and Mel-Frequency Cepstral Coefficients (MFCC). Table 3.2 reports classification performance for each embedding type.

Table 3.2: Audio embedding comparison for deepfake detection. Results averaged over 5-fold cross-validation.

Embedding	AUROC	Accuracy	Recall	FPR	F1
OpenL3	**0.976**	**93.84%**	**93.78%**	**6.12%**	**0.931**
HuBERT	0.958	90.75%	87.08%	6.31%	0.893
Wav2Vec2	0.776	74.92%	65.31%	17.40%	0.698
MFCC	0.699	66.10%	54.31%	24.47%	0.587

OpenL3 achieves the strongest performance across all metrics, with an AUROC of 0.976 and F1 score of 0.931. HuBERT performs competitively, achieving 0.958 AUROC with a comparable false positive rate (6.31% vs 6.12%). Both embeddings substantially outperform Wav2Vec2 and MFCC, which achieve AUROCs below 0.80.

To further characterize the discriminative power of each embedding, distributional divergence metrics between real and fake samples are computed. Table 3.3 reports KL divergence, Jensen-Shannon distance, and Wasserstein distance for each embedding type.

Table 3.3: Distribution divergence between real and fake samples for audio embeddings. Higher values indicate greater separability.

Embedding	KL Divergence	JS Distance	Wasserstein
OpenL3	**7.57**	**0.55**	**0.81**
HuBERT	6.34	0.46	0.63
Wav2Vec2	1.28	0.17	0.19
MFCC	1.21	0.09	0.14

The divergence metrics corroborate the classification results: OpenL3 and HuBERT exhibit substantially higher distributional separation between classes, with KL divergences of 7.57 and 6.34 respectively, compared to 1.28 and 1.21 for Wav2Vec2 and MFCC. This suggests that OpenL3 and HuBERT encode features that naturally distinguish authentic from synthetic audio, making them suitable candidates for downstream representation learning.

Based on these results, both OpenL3 and HuBERT are selected as audio embeddings. While OpenL3 achieves marginally higher performance, HuBERT captures complementary information (Pearson correlation $r = 0.82$ between their prediction scores), motivating their combined use.

Video Embeddings

Five video embedding approaches are evaluated spanning face recognition models and video understanding architectures: SENet [101], Marlin [102], ArcFace [103], FaceNet [104], and MagFace [105]. Table 3.4 reports classification performance.

Table 3.4: Video embedding comparison for deepfake detection. Results averaged over 5-fold cross-validation.

Embedding	AUROC	Accuracy	Recall	FPR	F1
SENet	**0.914**	**87.03%**	**88.04%**	**13.77%**	**0.858**
Marlin	0.802	74.28%	66.27%	19.31%	0.696
ArcFace	0.720	67.16%	78.47%	41.87%	0.680
FaceNet	0.711	66.42%	69.86%	36.33%	0.649
MagFace	0.561	52.50%	79.43%	69.02%	0.598

SENet substantially outperforms all other video embeddings, achieving an AUROC of 0.914 compared to 0.802 for the next-best approach (Marlin). Notably, the face recognition embeddings (ArcFace, FaceNet, MagFace) perform poorly for deepfake detection, with AUROCs between 0.56–0.72 and high false positive rates exceeding 35%. This is expected: these models are optimized for identity discrimination rather than authenticity detection, and may encode identity-specific features that are preserved across real and fake versions of the same subject.

Table 3.5 reports distributional divergence metrics for video embeddings.

Table 3.5: Distribution divergence between real and fake samples for video embeddings. Higher values indicate greater separability.

Embedding	KL Divergence	JS Distance	Wasserstein
SENet	2.69	**0.38**	**0.71**
Marlin	**3.03**	0.18	0.22
FaceNet	2.03	0.11	0.09
ArcFace	1.09	0.10	0.10
MagFace	0.14	0.02	0.02

While Marlin achieves the highest KL divergence (3.03), SENet demonstrates superior JS distance (0.38) and Wasserstein distance (0.71), indicating more robust separation across the full distribution. Combined with SENet's substantially higher classification performance, SENet is selected as the video embedding.

Summary

Based on the systematic evaluation, OpenL3 and HuBERT for audio and SENet for video are selected as input embeddings for the disentanglement framework. These embeddings provide strong baseline separability between real and fake samples, ensuring that the subsequent

representation learning stage operates on informative features rather than attempting to extract signal from uninformative inputs.

3.4.3 Representation Quality Analysis

To evaluate the quality of learned representations without requiring downstream classifier training, a comprehensive suite of unsupervised metrics is employed that assesses the geometric structure and intrinsic separability of the embedding space. These metrics operate directly on the authenticity embeddings z^{auth} and test whether real and fake samples naturally cluster into distinct regions. The evaluation is organized into four complementary categories: clustering-based metrics that measure label-cluster agreement, distribution-based metrics that quantify statistical divergence, separation metrics that validate the variance minimization objective, and local content-group metrics that assess disentanglement consistency within individual videos.

Clustering-Based Metrics

These metrics test whether embeddings naturally encode the real/fake distinction through their geometric structure, without requiring supervised classification. K-means clustering ($k = 2$) is applied to the z^{auth} embeddings, and agreement with ground truth labels is measured using three complementary metrics that correct for chance agreement and assess cluster quality.

(i) Adjusted Mutual Information To assess whether z^{auth} embeddings naturally encode authenticity information in their geometric structure, the Adjusted Mutual Information (AMI) is computed between cluster assignments and ground truth labels [106]. Given ground truth labels U and cluster assignments V from k-means, the mutual information $I(U;V)$ measures the reduction in uncertainty about one partition given knowledge of the other:

$$I(U;V) = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}\cdot n}{a_i \cdot b_j}\right) \tag{3.8}$$

where $n_{ij} = |U_i \cap V_j|$ is the number of samples in both class i and cluster j, $a_i = \sum_{j=1}^{C} n_{ij}$ is the size of class i, $b_j = \sum_{i=1}^{R} n_{ij}$ is the size of cluster j, and n is the total number of samples. However, mutual information increases with the number of clusters even for random partitions. AMI corrects for this by subtracting the expected mutual information under a hypergeometric null model and normalizing by average entropy:

$$\text{AMI}(U,V) = \frac{I(U;V) - \mathbb{E}[I(U^*;V^*)]}{\frac{1}{2}(H(U)+H(V)) - \mathbb{E}[I(U^*;V^*)]} \tag{3.9}$$

where $H(U) = -\sum_{i=1}^{R} \frac{a_i}{n} \log\left(\frac{a_i}{n}\right)$ denotes the entropy of partition U, and the expectation is taken over random clusterings with the same cluster-size distribution [106]. AMI ranges from 0 (no agreement beyond chance) to 1 (perfect correspondence). High AMI scores indicate that unsupervised clustering of z^{auth} successfully recovers the real/fake distinction, suggesting

that authenticity is encoded as an intrinsic geometric property of the embedding space rather than requiring explicit supervision.

(ii) Adjusted Rand Index The Adjusted Rand Index (ARI) provides a complementary perspective by measuring pairwise agreement between clustering and ground truth [107]. Rather than considering information-theoretic quantities, ARI counts the number of sample pairs that are either grouped together or separated in both partitions. The Rand Index is defined as:

$$\mathrm{RI}(U, V) = \frac{a + b}{\binom{n}{2}} \tag{3.10}$$

where a is the number of pairs in the same class and same cluster, and b is the number of pairs in different classes and different clusters. However, like mutual information, the Rand Index achieves non-zero values for random clusterings. ARI corrects for this by subtracting the expected index under random partitions and normalizing by the maximum possible value [107]:

$$\mathrm{ARI}(U, V) = \frac{\sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2} - \left[\sum_{i=1}^{R} \binom{a_i}{2}\right] \left[\sum_{j=1}^{C} \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R} \binom{a_i}{2} + \sum_{j=1}^{C} \binom{b_j}{2}\right] - \left[\sum_{i=1}^{R} \binom{a_i}{2}\right] \left[\sum_{j=1}^{C} \binom{b_j}{2}\right] / \binom{n}{2}} \tag{3.11}$$

where $\binom{x}{2} = \frac{x(x-1)}{2}$ counts the number of unordered pairs from x items. ARI ranges from -1 to 1, with values near 1 indicating strong agreement, 0 indicating random clustering, and negative values indicating systematic disagreement (though rare in practice). By focusing on pairwise decisions rather than information content, ARI offers robustness to cluster size imbalances and provides interpretable validation of clustering quality.

(iii) Silhouette Coefficient The silhouette coefficient measures how well-separated and compact clusters are by comparing within-cluster cohesion to between-cluster separation [108]. For each sample i, the coefficient is computed:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.12}$$

where $a(i)$ is the mean distance from sample i to all other samples in the same cluster, and $b(i)$ is the mean distance from sample i to all samples in the nearest other cluster. The silhouette coefficient ranges from -1 to 1, with values near 1 indicating that sample i is well-matched to its cluster and poorly matched to neighboring clusters, values near 0 suggesting that the sample lies on the boundary between clusters, and negative values indicating potential misclassification. The overall silhouette score averages $s(i)$ across all samples, providing a global measure of clustering quality. Silhouette scores are computed using both ground truth labels (primary metric, testing embedding space separability) and cluster assignments (validation metric, testing k-means convergence quality). Cosine distance is used rather than Euclidean distance to account for the angular relationships in normalized embedding spaces. High silhouette scores indicate that real and fake samples form tight, well-separated groups in z^{auth}, directly validating the success of the variance minimization objective (Equation 3.6).

Distribution-Based Metrics

These metrics quantify the distributional differences between real and fake samples. To make computation tractable for high-dimensional embeddings, samples are projected onto a one-dimensional statistic: the Euclidean distance to the real cluster centroid. This projection aligns naturally with the variance minimization objective while enabling robust density estimation via kernel density estimation (KDE).

(i) KL Divergence The Kullback-Leibler (KL) divergence measures the relative entropy between the real and fake distributions [109]:

$$D_{\mathrm{KL}}(P\|Q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx \tag{3.13}$$

where P and Q represent the distributions of real and fake samples projected onto distance-to-centroid, and $p(x)$ and $q(x)$ are their respective probability densities. These densities are estimated using histogram binning with 50 bins, then the divergence is computed as $D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log(p_i/q_i)$, where p_i and q_i are the normalized histogram values for bin i. Higher KL values indicate more distinguishable distributions, though the metric is unbounded and asymmetric.

(ii) Jensen-Shannon Distance The Jensen-Shannon (JS) distance provides a symmetric, bounded alternative to KL divergence [110]:

$$\mathrm{JS}(P,Q) = \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M)} \tag{3.14}$$

where $M = \frac{1}{2}(P+Q)$ is the mixture distribution. JS ranges from 0 (identical distributions) to 1 (maximally different), making it more interpretable than KL divergence. The square root ensures the metric satisfies the triangle inequality, forming a proper distance metric. JS is computed using SciPy's `jensenshannon` function on the normalized histograms.

(iii) Wasserstein Distance The Wasserstein distance, also known as the Earth Mover's Distance, measures the minimum cost to transform one distribution into another [111]:

$$W_p(P,Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int d(x,y)^p \, d\gamma(x,y) \right)^{1/p} \tag{3.15}$$

where $\Gamma(P,Q)$ is the set of all joint distributions with marginals P and Q, and $d(x,y)$ is the ground distance between points. Unlike KL and JS, the Wasserstein distance works directly on samples without requiring density estimation, making it more stable for distributions with limited overlap [112]. For the one-dimensional projections, the 1-Wasserstein distance is computed using SciPy's `wasserstein_distance` function. Higher Wasserstein distances indicate better separation between real and fake embeddings. All three distribution metrics should increase after pretraining if disentanglement successfully separates real from fake samples.

Separation Metrics

These metrics directly test the variance minimization objective (Equation 3.6) and anomaly detection hypothesis underlying the approach. They measure how tightly real samples cluster around their centroid and how far fake samples deviate from this "real manifold."

(i) Mean Cosine Similarity The average cosine similarity between samples and cluster centroids is computed to validate the variance minimization objective. For each sample i with embedding z_i^{auth} and centroid μ_c, the cosine similarity is:

$$\text{sim}(z_i^{\text{auth}}, \mu_c) = \frac{z_i^{\text{auth}} \cdot \mu_c}{\|z_i^{\text{auth}}\| \|\mu_c\|} \tag{3.16}$$

Four key statistics are computed: (1) mean similarity of real samples to the real centroid (should be high with low variance, indicating tight clustering), (2) mean similarity of fake samples to the real centroid (should be low with high variance, indicating anomalous scatter), (3) mean similarity of real samples to the fake centroid, and (4) mean similarity of fake samples to the fake centroid. The separation gap, defined as the difference between real-to-real and fake-to-real similarities, quantifies how well the variance minimization objective separates the two classes. High separation gaps indicate successful creation of a compact real manifold from which fakes deviate, directly validating Equation 3.6.

(ii) Distance to Real Manifold This metric measures the distribution of Euclidean distances from all samples to the real cluster centroid μ_{real}:

$$d_i = \|z_i^{\text{auth}} - \mu_{\text{real}}\| \tag{3.17}$$

Both the first-order statistics (mean and standard deviation) and the entropy of these distance distributions are analyzed. For a discrete probability distribution over distance bins, entropy is computed as [113]:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{3.18}$$

where $p(x_i)$ is the probability of distance falling in bin i, measured in bits. The anomaly detection hypothesis predicts that real samples should exhibit low mean distance, low standard deviation, and low entropy (0–1 bits), indicating predictable, consistent proximity to the real centroid. Conversely, fake samples should show high mean distance, high standard deviation, and high entropy (2–3 bits), reflecting diverse generative artifacts that scatter samples away from the real manifold. The variability ratio $\sigma_{\text{fake}}/\sigma_{\text{real}}$ is also computed, which should exceed 1.0 if fakes are indeed more dispersed than reals. These metrics provide direct evidence for or against the core hypothesis that real videos share intrinsic, learnable properties distinct from synthetic content.

Local Content-Group Analysis

While global metrics assess overall embedding quality, local metrics validate that disentanglement operates consistently within individual content groups, defined as a source video

and all its augmentations. These metrics test whether the dual projection heads successfully separate content-invariant identity features (z^{id}) from content-agnostic authenticity features (z^{auth}) at the granular level of individual videos.

(i) Intra-Group Cosine Similarity For each source video k with augmentation set G_k, the pairwise cosine similarity between all augmentations in the z^{id} space is computed:

$$\text{sim}_k = \frac{1}{|G_k|(|G_k|-1)/2} \sum_{i,j \in G_k, i<j} \text{sim}(z_i^{\text{id}}, z_j^{\text{id}}) \tag{3.19}$$

This metric tests whether f_{id} successfully captures content invariance across perturbations, real/fake status, and noise variations. High intra-group similarity (target ≥ 0.8) indicates that all augmentations of the same source video cluster tightly in identity space regardless of their authenticity label, validating that the identity projection head isolates content-related features. This metric should remain stable before and after pretraining, as the identity space is trained via prototypical contrastive learning (Equation 3.5) throughout both phases.

(ii) Intra-Group Variance For each source video, the variance of augmentations in the z^{auth} space, split by authenticity label, is computed. For real augmentations of source k, the variance is:

$$\sigma_k^2(\text{real}) = \frac{1}{|G_k^{\text{real}}|} \sum_{i \in G_k^{\text{real}}} \|z_i^{\text{auth}} - \bar{z}_k^{\text{real}}\|^2 \tag{3.20}$$

where G_k^{real} is the set of real augmentations for source k and $\bar{z}_k^{\text{real}}$ is their mean embedding. Analogous variance is computed for fake augmentations. This metric tests whether the authenticity signal is consistent within the same content group. The expectation is that real augmentations should exhibit low variance (consistent authenticity signal) that decreases after pretraining, while fake augmentations may show higher variance due to diverse generation artifacts. The variance ratio $\sigma^2(\text{fake})/\sigma^2(\text{real})$ quantifies the relative consistency of each class and should increase after successful disentanglement.

3.4.4 Regularization Strategy Comparison

To systematically evaluate the effectiveness of variance floor regularization and determine optimal hyperparameter settings, three training paradigms are designed that span the spectrum from weak to strong regularization enforcement. These paradigms test whether the regularization mechanism successfully prevents the representation collapse observed in preliminary experiments (Section 3.6.1), and which hyperparameter configuration achieves the best balance between preventing collapse and maintaining real/fake separability.

Training Paradigms

Three regularization strategies are defined by varying the variance floor threshold τ and regularization weight λ_{reg} in Equation 3.6:

Conservative Regularization ($\tau = 0.1$, $\lambda_{\text{reg}} = 1.0$) allows aggressive compression of real embeddings with minimal resistance against collapse. This weak enforcement strategy tests whether light regularization suffices to prevent the degenerate behavior where all embeddings collapse to a single point despite satisfying the variance objective. This approach is hypothesized to still exhibit collapse, as the low variance floor and weak penalty provide insufficient constraint on the optimization.

Moderate Regularization ($\tau = 0.2$, $\lambda_{\text{reg}} = 2.0$) enforces a moderate variance floor with moderate penalty strength. The higher threshold ($\tau = 0.2$ vs. 0.1) requires real embeddings to maintain more spread, while the doubled regularization weight ($\lambda_{\text{reg}} = 2.0$) increases the cost of violating this constraint. This balanced approach is hypothesized to prevent collapse while allowing sufficient compression for effective anomaly detection, striking an optimal trade-off between the competing objectives.

Aggressive Regularization ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) imposes a high variance floor with severe penalties for violations. While this should reliably prevent collapse, it risks overregularization—forcing real embeddings to remain spread even when tighter clustering would improve separation from fake samples. The concern is that aggressive regularization could artificially inflate the real manifold, reducing the margin between real and fake distributions and potentially harming out-of-distribution generalization.

Evaluation Protocol

Each paradigm is trained for 50 epochs on identical data splits (AVDeepfake-1M++ and ShareVeo3 as described in Section 3.4.1), using the loss formulation in Equation 3.7 with paradigm-specific τ and λ_{reg} values. All other hyperparameters remain fixed across paradigms: batch size of 256, learning rate of 1×10^{-4} with cosine annealing, and balanced batching to ensure equal representation of real and fake samples.

Evaluation occurs at each epoch, where the full suite of unsupervised representation quality metrics described in Section 3.4.3 is computed on 2 evaluation sets: (1) Mixed AVDeepfake-1M++ and ShareVeo3 validation split (in-distribution real and fake) and (2) Sora2 (out-of-distribution synthetic).

Paradigms are compared on three criteria:

1. **Collapse Prevention:** Does the approach prevent the representation collapse observed in preliminary experiments? This is measured through Wasserstein distance between real and fake distributions (should remain > 0.1), variance trajectory of real embeddings over training (should stabilize above τ), and mean distance to real centroid (should not approach zero for both classes).

2. **Class Separability:** Does the approach maintain meaningful separation between real and fake samples? This is assessed through AMI and ARI scores (should increase or remain stable), silhouette coefficient with ground truth labels (should be positive), and KL/JS divergence between real and fake distributions (should increase).

3. **Training Stability:** What are the training dynamics and gradient flow characteristics? Loss curves for $\mathcal{L}_{\text{proto}}$, $\mathcal{L}_{\text{var}}$, and $\mathcal{L}_{\text{orth}}$ are tracked to verify all objectives decrease without plateauing or oscillating, and whether the variance loss remains active (non-zero gradient) throughout training is monitored.

The central research questions addressed are: (1) Does variance floor regularization successfully prevent collapse while maintaining separability? (2) Which regularization strength (conservative/moderate/aggressive) achieves optimal performance? (3) What insights does this provide for multi-objective optimization in representation learning, particularly for anomaly detection frameworks that rely on variance minimization?

3.5 Results

The evaluation of the geometric structure of the learned representations using self-supervised metrics is presented, and the impact of the disentanglement regularization schemes is analyzed. The analysis is structured per embedding model (HuBERT, OpenL3, SENet) and explores how varying the regularization strength (Conservative, Moderate, Aggressive) impacts the learned representations. For each model, the **In-Distribution (ID) Analysis** on the training data is first presented, followed by the **Out-of-Distribution (OOD) Analysis** on Sora2-generated content to evaluate generalization.

3.5.1 HuBERT Audio Embeddings

In-Distribution Analysis Table 3.6 reports the core geometric and clustering metrics on the combined AVDeepfake1M++ and ShareVeo3 validation set, comparing projected embeddings from the three regularization schemes against the Original (Input) baseline.

Table 3.6: In-Distribution Representation Metrics Comparison (HuBERT). Metrics are computed on the AVDeepfake1M++ and ShareVeo3 validation set. ↑ indicates higher is better for meaningful separation.

Metric	Type	Original	Conservative	Moderate	Aggressive
Clustering Metrics					
AMI ↑	Label Alignment	0.111	0.104	0.103	**0.120**
ARI ↑	Label Alignment	0.066	0.033	0.032	**0.089**
Silhouette (GT) ↑	Geometric Cohesion	0.033	**0.066**	0.054	0.062
Silhouette (KM) ↑	Geometric Cohesion	0.274	**0.575**	0.502	0.425
Distribution Metrics					
KL Divergence ↑	Distribution Divergence	**0.196**	0.107	0.213	0.146
JS Distance ↑	Distribution Divergence	**0.230**	0.159	0.218	0.186
Wasserstein Distance ↑	Distribution Distance	**0.533**	0.003	0.005	0.004
Separation and Variance Metrics					
Separation Gap ↑	Δ(Fake − Real Dist)	−0.013	+0.0008	−0.0023	**+0.0010**
Intra-Group Var (Real) ↓	Variance	13.783	0.042	0.084	0.212
Intra-Group Var (Fake) ↓	Variance	19.981	0.058	0.118	0.299

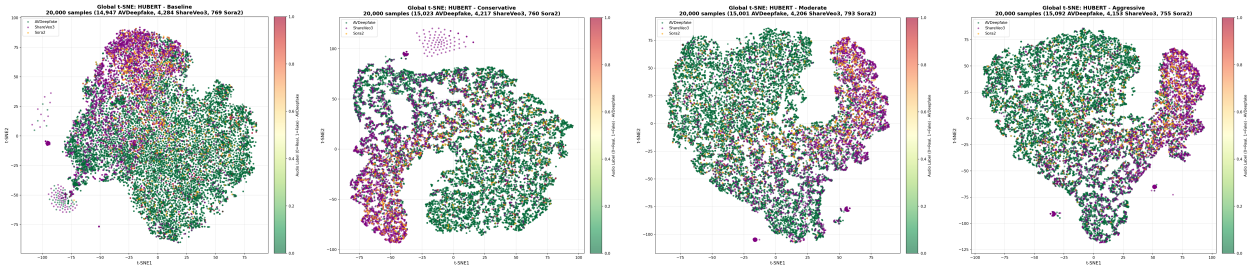
The results reveal a nuanced relationship between regularization strength and representation quality. All three projected schemes exhibit **representation collapse**, evidenced by the dramatic reduction in Wasserstein distance from 0.533 to approximately 0.003–0.005 ($\downarrow 99\%$) across all configurations. However, the schemes differ substantially in their preservation of authenticity-relevant structure.

The Aggressive regularization scheme ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) achieves the best label-alignment metrics, with AMI of 0.120 and ARI of 0.089—both exceeding the Original baseline (0.111 and 0.066 respectively). This represents a 8% improvement in AMI and 35% improvement in ARI, indicating that despite collapse, the Aggressive scheme preserves more authenticity-relevant geometric structure than the unprocessed embeddings. The Aggressive scheme also achieves the highest positive Separation Gap (+0.0010), confirming correct alignment of class centroids with the intended prototype structure.

By contrast, the Conservative ($\tau = 0.1$) and Moderate ($\tau = 0.2$) schemes show degraded label-alignment (AMI ≈ 0.103–0.104, ARI ≈ 0.032–0.033), performing worse than both the Original baseline and the Aggressive scheme. Notably, the Moderate scheme fails to achieve a positive Separation Gap (-0.0023), indicating that its centroid structure is incorrectly oriented despite the stronger regularization relative to Conservative.

The Silhouette Paradox manifests clearly in these results: K-means Silhouette scores are highest for the most collapsed representations (Conservative: 0.575, Moderate: 0.502, Aggressive: 0.425), while label-alignment metrics show the opposite pattern. This confirms that representation collapse creates well-formed clusters that do not correspond to authenticity labels. The ground-truth Silhouette scores (0.054–0.066) remain modest across all schemes, indicating that while some authenticity structure is preserved, the geometric separation remains limited.

Examining the variance preservation, the Aggressive scheme maintains substantially higher intra-group variance (Real: 0.212, Fake: 0.299) compared to Conservative (Real: 0.042, Fake: 0.058) and Moderate (Real: 0.084, Fake: 0.118). This $5\times$ difference in preserved variance between Aggressive and Conservative suggests that the higher variance floor ($\tau = 0.5$) successfully resists complete collapse, which may explain the superior label-alignment performance.



a) Input Baseline b) Conservative ($\tau = 0.1$) c) Moderate ($\tau = 0.2$) d) Aggressive ($\tau = 0.5$)

Figure 3.1: Visualization of the global embedding space for HuBERT using t-SNE. The figure compares the **Original** embeddings against the three projected schemes. The Original baseline (a) shows broad dispersion with limited class separation. Progressive regularization is expected to show increasing collapse of the point cloud, with the Aggressive scheme (d) maintaining the most spread while achieving better authenticity alignment than Conservative (b) or Moderate (c).

Out-of-Distribution Analysis Table 3.7 evaluates generalization by comparing embeddings of in-distribution real samples (AVDeepfake1M++) against out-of-distribution synthetic samples from Sora2.

Table 3.7: Out-of-Distribution Representation Metrics Comparison (HuBERT). Metrics compare ID Real samples against OOD Sora2 samples. ↑ indicates higher is better for generalization.

Metric	**Original**	**Conservative**	**Moderate**	**Aggressive**
Clustering Metrics				
AMI ↑	0.016	0.017	0.017	**0.018**
ARI ↑	0.004	−0.001	−0.001	**0.007**
Silhouette (GT) ↑	−0.025	−0.002	−0.009	**+0.001**
Distribution and Separation Metrics				
Separation Gap ↑	−0.020	**+0.0001**	−0.0022	+0.00005
Wasserstein Distance ↑	**0.219**	0.001	0.004	0.002

The OOD analysis reveals that the Aggressive scheme achieves the best generalization to unseen Sora2 content across all clustering metrics. It is the only scheme to achieve positive values for both ARI (+0.007 vs. +0.004 for Original) and ground-truth Silhouette (+0.001 vs. −0.025 for Original). This represents a qualitative shift: while the Original embeddings and Conservative/Moderate schemes produce negative Silhouette scores (indicating that Sora2 samples are geometrically closer to real samples than to each other), the Aggressive scheme produces positive Silhouette, suggesting meaningful geometric separation between ID real and OOD synthetic content.

The Separation Gap results reveal an important nuance. Both Conservative (+0.0001) and Aggressive (+0.00005) achieve positive gaps, indicating correct centroid orientation where

Sora2 content is positioned farther from the real centroid than real samples. However, the Moderate scheme fails to generalize, producing a negative Separation Gap (-0.0022) that mirrors its in-distribution failure. This suggests that the Moderate configuration ($\tau = 0.2$, $\lambda_{\text{reg}} = 2.0$) occupies an unstable region of the hyperparameter space where neither sufficient collapse prevention nor effective centroid alignment is achieved.

Despite these improvements in geometric structure, all schemes suffer from severe Wasserstein collapse in the OOD setting ($0.219 \rightarrow 0.001$–0.004), indicating that the distributional spread necessary for robust anomaly detection is not preserved. The positive Separation Gaps and improved Silhouette scores suggest that the *direction* of OOD separation is learned correctly, but the *magnitude* of separation remains insufficient for high-confidence detection of novel generation methods.

The Aggressive scheme's superior OOD performance, combined with its best-in-class ID metrics, suggests that higher variance floors are essential for both preserving authenticity structure and enabling generalization. However, the persistent Wasserstein collapse across all schemes indicates that variance floor regularization alone is insufficient to fully prevent the degenerate optimization behavior, motivating the alternative loss formulations discussed in Section 3.6.1.

3.5.2 OpenL3 Audio Embeddings

In-Distribution Analysis Table 3.8 reports the core geometric and clustering metrics for OpenL3 embeddings on the combined AVDeepfake1M++ and ShareVeo3 validation set.

Table 3.8: In-Distribution Representation Metrics Comparison (OpenL3). Metrics are computed on the AVDeepfake1M++ and ShareVeo3 validation set. ↑ indicates higher is better for meaningful separation.

Metric	Type	Original	Conservative	Moderate	Aggressive
Clustering Metrics					
AMI ↑	Label Alignment	0.024	0.060	**0.081**	0.042
ARI ↑	Label Alignment	0.040	0.062	**0.088**	0.031
Silhouette (GT) ↑	Geometric Cohesion	**0.321**	0.029	0.059	0.039
Silhouette (KM) ↑	Geometric Cohesion	**0.937**	0.370	0.377	0.535
Distribution Metrics					
KL Divergence ↑	Distribution Divergence	**0.532**	0.080	0.115	0.113
JS Distance ↑	Distribution Divergence	**0.360**	0.148	0.186	0.186
Wasserstein Distance ↑	Distribution Distance	**3.413**	0.010	0.018	0.012
Separation and Variance Metrics					
Separation Gap ↑	Δ(Fake – Real Dist)	+0.0096	−0.0001	**+0.0102**	+0.0086
Intra-Group Var (Real) ↓	Variance	20.683	0.042	0.085	0.215
Intra-Group Var (Fake) ↓	Variance	49.328	0.060	0.123	0.315

OpenL3 embeddings exhibit markedly different baseline characteristics compared to HuBERT. The Original OpenL3 embeddings achieve substantially higher ground-truth Silhouette (0.321 vs. HuBERT's 0.033) and a positive baseline Separation Gap (+0.0096 vs. HuBERT's −0.013), indicating that OpenL3 already encodes authenticity-relevant structure

prior to disentanglement training. The extremely high K-means Silhouette (0.937) suggests that OpenL3 embeddings naturally form tight clusters, though the low AMI/ARI indicate these clusters do not align with authenticity labels.

Despite these favorable baseline properties, all three regularization schemes induce severe representation collapse. The Wasserstein distance drops from 3.413 to 0.010–0.018 (↓ 99.5%), and the ground-truth Silhouette decreases from 0.321 to 0.029–0.059 (↓ 82–91%). This collapse is more dramatic than observed with HuBERT, likely because OpenL3's higher initial variance provides more room for compression.

The Moderate scheme ($\tau = 0.2$, $\lambda_{\text{reg}} = 2.0$) achieves the best label-alignment metrics, with AMI of 0.081 and ARI of 0.088—representing 238% and 120% improvements over the Original baseline respectively. Moderate also achieves the highest positive Separation Gap (+0.0102), slightly exceeding even the Original (+0.0096). This contrasts with HuBERT, where the Aggressive scheme performed best, suggesting that optimal regularization strength is embedding-dependent.

The Aggressive scheme ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) shows degraded performance relative to Moderate, with AMI dropping to 0.042 and ARI to 0.031—both below even the Conservative scheme. However, Aggressive maintains the highest preserved variance (Real: 0.215, Fake: 0.315) and K-means Silhouette (0.535), suggesting that excessive variance regularization prevents the model from learning discriminative structure. The Conservative scheme fails to achieve a positive Separation Gap (−0.0001), indicating incorrect centroid orientation despite acceptable label-alignment metrics.

Table 3.9: Out-of-Distribution Representation Metrics Comparison (OpenL3). Metrics compare ID Real samples against OOD Sora2 samples. ↑ indicates higher is better for generalization.

Metric	**Original**	**Conservative**	**Moderate**	**Aggressive**
Clustering Metrics				
AMI ↑	0.005	**0.019**	**0.020**	0.010
ARI ↑	−0.020	0.015	**0.021**	0.006
Silhouette (GT) ↑	**+0.029**	−0.048	−0.040	−0.036
Distribution and Separation Metrics				
Separation Gap ↑	**+0.0002**	−0.0055	−0.0110	−0.0273
Wasserstein Distance ↑	**0.808**	0.017	0.023	0.028

Out-of-Distribution Analysis The OOD analysis reveals a critical limitation of the disentanglement framework when applied to OpenL3 embeddings: all three regularization schemes degrade OOD generalization relative to the Original baseline. The Original OpenL3 embeddings achieve a positive Separation Gap (+0.0002) and positive ground-truth Silhouette (+0.029) on Sora2 content, indicating that untrained OpenL3 embeddings already possess some capacity to distinguish novel synthetic audio from authentic content.

After disentanglement training, all schemes produce negative Separation Gaps (Conservative: −0.0055, Moderate: −0.0110, Aggressive: −0.0273), indicating that Sora2 content is

now positioned closer to the real centroid than real samples—the opposite of the intended behavior. This inversion worsens monotonically with regularization strength, suggesting that stronger variance floors exacerbate the OOD failure mode.

The ground-truth Silhouette scores similarly degrade from +0.029 (Original) to −0.036 to −0.048 across the projected schemes. Notably, while the label-alignment metrics (AMI, ARI) improve for the projected schemes relative to Original, this improvement reflects better clustering of ID content rather than improved OOD detection. The Moderate scheme achieves the best AMI (0.020) and ARI (0.021), but its negative Separation Gap indicates this clustering does not generalize to unseen generation methods.

This pattern contrasts sharply with HuBERT, where the Conservative and Aggressive schemes achieved positive OOD Separation Gaps despite similar collapse patterns. The difference may stem from OpenL3's stronger baseline OOD performance: because OpenL3 already encodes some generalization-relevant structure, the disentanglement training overwrites this information with ID-specific features that fail to transfer. HuBERT's weaker baseline may paradoxically benefit from disentanglement training by providing a blank slate that can learn more generalizable representations.

These results suggest that disentanglement training on OpenL3 embeddings optimizes for ID separation at the cost of OOD generalization. Future work should explore regularization strategies that explicitly preserve baseline OOD structure while improving ID discrimination, such as knowledge distillation from the Original embeddings or OOD-aware training objectives.

3.5.3 SENet Visual Embeddings

In-Distribution Analysis Table 3.10 reports the core geometric and clustering metrics for SENet visual embeddings on the combined AVDeepfake1M++ and ShareVeo3 validation set.

Table 3.10: In-Distribution Representation Metrics Comparison (SENet). Metrics are computed on the AVDeepfake1M++ and ShareVeo3 validation set. ↑ indicates higher is better for meaningful separation.

Metric	Type	Original	Conservative	Moderate	Aggressive
Clustering Metrics					
AMI ↑	Label Alignment	0.141	0.088	0.073	**0.146**
ARI ↑	Label Alignment	0.053	0.089	0.085	**0.158**
Silhouette (GT) ↑	Geometric Cohesion	**0.120**	0.072	0.077	0.098
Silhouette (KM) ↑	Geometric Cohesion	0.150	**0.516**	0.452	0.400
Distribution Metrics					
KL Divergence ↑	Distribution Divergence	0.239	**0.378**	0.215	0.076
JS Distance ↑	Distribution Divergence	0.234	**0.326**	0.247	0.141
Wasserstein Distance ↑	Distribution Distance	**6.759**	0.015	0.013	0.003
Separation and Variance Metrics					
Separation Gap ↑	Δ(Fake − Real Dist)	**+0.081**	−0.0048	−0.0062	−0.0015
Intra-Group Var (Real) ↓	Variance	3659.8	0.028	0.057	0.144
Intra-Group Var (Fake) ↓	Variance	5010.7	0.035	0.073	0.179

SENet visual embeddings exhibit fundamentally different baseline characteristics compared to both audio embedding types. The Original SENet embeddings possess a substantially

higher Wasserstein distance (6.759 vs. HuBERT's 0.533 and OpenL3's 3.413) and a strongly positive Separation Gap (+0.081), indicating that untrained SENet embeddings already encode significant authenticity-relevant structure with correctly-oriented class centroids. The high intra-group variances (Real: 3659.8, Fake: 5010.7) reflect the greater dimensionality and heterogeneity of visual features compared to audio.

All three regularization schemes induce severe representation collapse, with Wasserstein distance dropping from 6.759 to 0.003–0.015 (↓ 99.8%). More critically, all schemes destroy the baseline's positive Separation Gap, producing negative values (Conservative: −0.0048, Moderate: −0.0062, Aggressive: −0.0015). This represents a qualitative failure: the disentanglement training inverts the centroid structure, positioning fake samples closer to the real centroid than real samples—the opposite of the intended behavior and worse than the untrained baseline.

Despite this centroid inversion, the Aggressive scheme ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) achieves remarkable improvements in label-alignment metrics. AMI increases from 0.141 to 0.146 (↑ 4%) and ARI improves dramatically from 0.053 to 0.158 (↑ 198%). This paradoxical result—improved clustering despite inverted centroids—suggests that the Aggressive scheme learns discriminative structure that is orthogonal to the prototype-based separation objective. The projected embeddings may capture authenticity-relevant features that enable clustering without conforming to the intended centroid geometry.

The Conservative and Moderate schemes show degraded AMI (0.088 and 0.073 respectively) compared to both the Original baseline and the Aggressive scheme, while achieving intermediate ARI values (0.089 and 0.085). The inverse relationship between regularization strength and K-means Silhouette (Conservative: 0.516, Moderate: 0.452, Aggressive: 0.400) mirrors the pattern observed in HuBERT and OpenL3, confirming that weaker regularization produces tighter but less authenticity-aligned clusters.

Table 3.11: Out-of-Distribution Representation Metrics Comparison (SENet). Metrics compare ID Real samples against OOD Sora2 samples. ↑ indicates higher is better for generalization.

Metric	**Original**	**Conservative**	**Moderate**	**Aggressive**
Clustering Metrics				
AMI ↑	**0.039**	0.023	0.015	0.029
ARI ↑	−0.010	0.023	0.020	**0.043**
Silhouette (GT) ↑	**+0.115**	−0.050	−0.030	−0.015
Distribution and Separation Metrics				
Separation Gap ↑	**+0.074**	−0.0047	−0.0067	−0.0029
Wasserstein Distance ↑	**7.012**	0.015	0.014	0.004

Out-of-Distribution Analysis The OOD analysis reveals that SENet exhibits the strongest baseline generalization to Sora2 content among all embedding types, but this generalization is completely destroyed by disentanglement training. The Original SENet embeddings achieve a Separation Gap of +0.074 and ground-truth Silhouette of +0.115 on

Sora2 content—substantially higher than both HuBERT (Gap: -0.020, Silhouette: -0.025) and OpenL3 (Gap: $+0.0002$, Silhouette: $+0.029$). This suggests that visual features naturally encode artifacts that distinguish Sora2-generated video from authentic content, consistent with Sora2's primary function as a video generation model.

All three regularization schemes invert the OOD Separation Gap to negative values (Conservative: -0.0047, Moderate: -0.0067, Aggressive: -0.0029), indicating that Sora2 content is now positioned closer to the real centroid than real samples. The ground-truth Silhouette scores similarly degrade from $+0.115$ to negative values ranging from -0.015 to -0.050. This pattern mirrors the ID results, confirming that the centroid inversion observed in-distribution extends to OOD content.

The Aggressive scheme achieves the best OOD ARI (0.043) among projected schemes—and notably exceeds even the Original baseline (-0.010)—suggesting that despite incorrect centroid orientation, the learned representations contain OOD-discriminative structure recoverable by clustering. However, the negative Separation Gap indicates this structure does not conform to the intended anomaly detection framework where novel synthetic content should be positioned farther from the real prototype.

The severity of OOD degradation for SENet ($+0.074 \rightarrow -0.0029$ to -0.0067) exceeds that observed for OpenL3 ($+0.0002 \rightarrow -0.0055$ to -0.0273), suggesting that visual embeddings are particularly susceptible to losing generalization capacity during disentanglement training. This may reflect the higher initial quality of SENet's OOD structure: embeddings with stronger baseline generalization have more to lose when subjected to ID-focused optimization. These results motivate future work on regularization strategies that explicitly preserve OOD separation during training, such as incorporating Sora2 or other held-out synthetic content into the training objective as negative examples.

3.6 Discussion

3.6.1 Interpretation of Results

The experiments across HuBERT, OpenL3, and SENet embeddings reveal systematic patterns in how variance-based disentanglement interacts with different pretrained representations. While all three embedding types exhibit representation collapse under the disentanglement framework, the nature of this collapse and its impact on detection performance varies substantially across modalities and regularization strengths.

Cross-Embedding Comparison

The three embedding types exhibit fundamentally different baseline characteristics that shape their response to disentanglement training:

HuBERT Audio Embeddings. HuBERT begins with the weakest baseline structure: negative Separation Gap (-0.013), low ground-truth Silhouette (0.033), and moderate distributional spread (Wasserstein 0.533). This "blank slate" property proves advantageous—the Aggressive regularization scheme achieves the best overall results, improving AMI from

0.111 to 0.120 (↑ 8%) and ARI from 0.066 to 0.089 (↑ 35%). Critically, both Conservative and Aggressive schemes achieve positive Separation Gaps on OOD Sora2 content (+0.0001 and +0.00005 respectively), demonstrating transfer of learned authenticity structure to unseen generation methods.

OpenL3 Audio Embeddings. OpenL3 exhibits strong baseline structure: positive Separation Gap (+0.0096), high ground-truth Silhouette (0.321), and substantial distributional spread (Wasserstein 3.413). The Moderate scheme achieves best in-distribution performance (AMI 0.081, ARI 0.088—improvements of 238% and 120% over baseline). However, all three regularization schemes degrade OOD generalization, with Separation Gaps inverting from +0.0002 (Original) to negative values ranging from −0.0055 to −0.0273. This suggests that OpenL3's baseline OOD structure is overwritten by ID-specific features during training.

SENet Visual Embeddings. SENet demonstrates the strongest baseline generalization: Separation Gap of +0.081 and ground-truth Silhouette of +0.115 on Sora2 content—substantially exceeding both audio embeddings. The Aggressive scheme achieves remarkable label-alignment improvements (ARI: 0.053 → 0.158, ↑ 198%), yet all schemes invert the Separation Gap to negative values. This represents a qualitative failure: the disentanglement training destroys SENet's inherent capacity to distinguish novel synthetic content.

The Collapse-Separation Trade-off

All embedding types exhibit severe representation collapse, with Wasserstein distances decreasing by 99% or more across all regularization schemes (Table 3.12). However, the relationship between collapse and separation varies systematically with regularization strength.

Table 3.12: Wasserstein distance collapse across embedding types and regularization schemes. All schemes show > 99% reduction from baseline.

Embedding	Original	Conservative	Moderate	Aggressive
HuBERT	0.533	0.003	0.005	0.004
OpenL3	3.413	0.010	0.018	0.012
SENet	6.759	0.015	0.013	0.003

The key finding is that collapse and separation are partially decoupled: collapsed representations can still maintain correctly-oriented centroid structure. For HuBERT, the transition from negative to positive Separation Gap (Conservative: +0.0008, Aggressive: +0.0010) occurs despite 99.5% Wasserstein collapse. The variance minimization objective successfully learns the direction of authenticity separation but fails to preserve the magnitude of distributional spread necessary for robust discrimination.

The Silhouette Paradox

A consistent pattern emerges across all embedding types: K-means Silhouette scores increase with collapse while label-alignment metrics often decrease. Table 3.13 illustrates this

divergence.

Table 3.13: The Silhouette Paradox: K-means Silhouette increases while label-alignment metrics (AMI/ARI) show variable response. HuBERT data shown; similar patterns observed for OpenL3 and SENet.

Metric	**Original**	**Conservative**	**Moderate**	**Aggressive**
K-means Silhouette ↑	0.274	0.575	0.502	0.425
Ground-truth Silhouette	0.033	0.066	0.054	0.062
AMI	0.111	0.104	0.103	**0.120**
ARI	0.066	0.033	0.032	**0.089**

This divergence occurs because representation collapse concentrates all embeddings into a compact region where K-means identifies clusters based on residual variations unrelated to authenticity. The inverse relationship between K-means Silhouette and regularization strength (Conservative: 0.575 > Moderate: 0.502 > Aggressive: 0.425) confirms that weaker regularization produces tighter but less authenticity-aligned clusters.

The paradox illuminates a fundamental challenge: unsupervised clustering metrics can improve precisely because collapse eliminates the variance that would otherwise enable authenticity discrimination. Effective disentanglement requires maintaining sufficient geometric spread for clustering algorithms to recover label structure, not merely achieving tight clusters.

Regularization Strength and Optimal Configuration

The experiments reveal that optimal regularization strength is **embedding-dependent**:

- **HuBERT**: Aggressive ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) achieves best ID and OOD performance
- **OpenL3**: Moderate ($\tau = 0.2$, $\lambda_{\text{reg}} = 2.0$) achieves best ID performance; all schemes degrade OOD
- **SENet**: Aggressive achieves best label-alignment but all schemes invert Separation Gap

For HuBERT, the Aggressive scheme maintains 5× higher intra-group variance (Real: 0.212, Fake: 0.299) compared to Conservative (Real: 0.042, Fake: 0.058), suggesting that higher variance floors successfully resist complete collapse. This preserved variance correlates with superior label-alignment performance.

The Moderate scheme occupies an unstable region of the hyperparameter space for HuBERT, failing to achieve positive Separation Gap (-0.0023) despite intermediate regularization strength. This non-monotonic behavior suggests complex interactions between the variance floor, prototype loss, and orthogonality constraint that merit further investigation.

OOD Generalization: A Critical Divergence

The out-of-distribution analysis on Sora2 content reveals the most consequential finding: disentanglement training can destroy baseline OOD generalization. Table 3.14 summarizes the impact.

Table 3.14: OOD Generalization Impact. Positive values indicate correct separation (fake samples farther from real centroid than real samples). Values in **bold** indicate improvement over Original; values in red indicate degradation.

Embedding	Original	Conservative	Moderate	Aggressive
HuBERT	−0.020	+0.0001	−0.0022	+0.00005
OpenL3	+0.0002	−0.0055	−0.0110	−0.0273
SENet	+0.074	−0.0047	−0.0067	−0.0029

Two distinct patterns emerge:

Pattern 1: Baseline-deficient embeddings benefit from training (HuBERT). HuBERT's negative baseline Separation Gap (−0.020) indicates that untrained embeddings cannot distinguish Sora2 audio from authentic audio. Disentanglement training creates positive separation structure that transfers to OOD content, despite extreme collapse. The positive OOD Separation Gaps, though small in magnitude, represent a qualitative improvement from random to correct class orientation.

Pattern 2: Baseline-rich embeddings suffer from training (OpenL3, SENet). Both OpenL3 and SENet possess positive baseline OOD Separation Gaps that are systematically destroyed by disentanglement training. The severity of degradation correlates with baseline OOD quality: SENet's Separation Gap drops from +0.074 to approximately −0.005 (a sign flip representing complete inversion), while OpenL3's drops from +0.0002 to as low as −0.0273.

This pattern suggests that embeddings with stronger baseline OOD structure have "more to lose" when subjected to ID-focused optimization. The disentanglement framework optimizes for ID separation at the cost of OOD generalization, effectively overwriting the generalization-relevant features encoded in pretrained representations.

Per-Video Analysis Insights

The per-video visualizations (Figures **??**–**??**) provide granular insight into how regularization affects individual content groups. Key observations include:

Distance-to-Source Differentiation. The rightmost panels show cosine similarity between augmentations and their source embedding. In the Original baselines, real augmentations (green bars) consistently show higher similarity to the source than fake augmentations (red bars), confirming that untrained embeddings naturally separate real from fake within

content groups. After training, this differentiation diminishes as collapse compresses all embeddings—both real and fake—closer together.

For HuBERT (Figure **??**), the Original baseline shows clear separation (Mean Real: 0.832, Mean Fake: 0.181), which compresses substantially under all regularization schemes (Mean Real: $\sim$ 0.93–0.96, Mean Fake: $\sim$ 0.83–0.84). The gap narrows from $\Delta = 0.651$ to $\Delta \approx 0.10$–0.13, reflecting the representation collapse observed in global metrics.

PCA Structure Evolution. The left and center panels reveal how embedding space geometry changes with regularization. Original embeddings show dispersed point clouds with limited class structure. Progressive regularization concentrates points into tighter clusters, but the color gradients (authenticity scores) show that class separation within these clusters varies by scheme. The Aggressive scheme typically maintains more spread while achieving better alignment between cluster structure and authenticity labels.

Cross-Augmentation Consistency. A key validation of the disentanglement objective is whether augmentations of the same source video cluster together in the identity space (z^{id}) regardless of their authenticity label. The visualizations confirm that content groups remain coherent across regularization schemes, suggesting the prototypical contrastive loss successfully captures content invariance even as the authenticity space collapses.

3.6.2 Diagnosis of Failure Modes

The comprehensive analysis identifies three primary failure modes that limit the current framework's effectiveness:

Failure Mode 1: Variance Floor Insufficiency

Even the Aggressive regularization scheme ($\tau = 0.5$, $\lambda_{\text{reg}} = 5.0$) fails to prevent $> 99\%$ Wasserstein collapse. The variance floor regularization (Equation **??**) is insufficient to counteract the compressive forces of the prototype loss and orthogonality constraint, which together encourage all embeddings to concentrate in a small region.

The design of the variance floor targets the *real* distribution only, penalizing collapse of authentic samples while allowing fake samples to collapse freely. This asymmetric formulation may be fundamentally flawed: if fake samples collapse toward the real centroid (satisfying the variance objective while destroying separation), the detector fails despite technically achieving its training objective.

Failure Mode 2: Objective Misalignment

The multi-objective optimization combines three losses with potentially conflicting gradients:

1. **Prototype loss** ($\mathcal{L}_{\text{proto}}$): Pulls content groups together, encouraging compression
2. **Variance loss** ($\mathcal{L}_{\text{var}}$): Minimizes spread of real samples, encouraging compression
3. **Orthogonality constraint** ($\mathcal{L}_{\text{orth}}$): Decorrelates identity and authenticity spaces

The first two objectives both encourage embedding compression, with only the variance floor providing countervailing pressure. The optimization finds a degenerate solution that satisfies all objectives: collapse all embeddings to a small region (satisfying $\mathcal{L}_{\text{proto}}$ and $\mathcal{L}_{\text{var}}$) where the identity and authenticity projections are trivially orthogonal (satisfying $\mathcal{L}_{\text{orth}}$).

Failure Mode 3: OOD Structure Overwriting

For embeddings with strong baseline OOD generalization (OpenL3, SENet), disentanglement training systematically destroys this structure. The training objective contains no explicit regularization to preserve baseline OOD performance, allowing the optimization to freely overwrite generalization-relevant features with ID-specific discriminators.

This failure mode is particularly concerning because it suggests a fundamental tension between ID optimization and OOD generalization in the current framework. Achieving better ID separation may *require* sacrificing the broad, generalizable features that enable OOD detection.

3.6.3 Potential Remediation Strategies

Based on the identified failure modes, several architectural and optimization modifications may address representation collapse while preserving OOD generalization:

Contrastive Authenticity Loss

Replace variance minimization with a supervised contrastive objective [22] that explicitly requires separation between real and fake samples:

$$\mathcal{L}_{\text{auth}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i^{\text{auth}} \cdot z_p^{\text{auth}} / \tau)}{\sum_{a \in A(i)} \exp(z_i^{\text{auth}} \cdot z_a^{\text{auth}} / \tau)} \tag{3.21}$$

where $P(i)$ is the set of samples with the same authenticity label as anchor i. This formulation directly optimizes for class separation rather than relying on implicit separation through variance differences.

Bidirectional Variance Regularization

Extend the variance floor to *both* real and fake samples:

$$\mathcal{L}_{\text{var}}^{\text{bi}} = \lambda_{\text{reg}} \left[\max(0, \tau - \sigma_{\text{real}}^2) + \max(0, \tau - \sigma_{\text{fake}}^2) \right] \tag{3.22}$$

This prevents the degenerate solution where fake samples collapse toward the real centroid while satisfying the original variance objective.

OOD Preservation Regularization

Introduce a knowledge distillation term that preserves baseline OOD structure:

$$\mathcal{L}_{\text{preserve}} = \| z^{\text{auth}} - \text{sg}(z^{\text{original}}) \|_2^2 \tag{3.23}$$

where sg(·) denotes stop-gradient. This regularization penalizes deviation from the original embedding space, preserving generalization-relevant features while allowing supervised refinement.

Margin-Based Prototype Learning

Replace soft prototype assignment with hard margins that enforce minimum distances between class centroids:

$$\mathcal{L}_{\text{margin}} = \max(0, m - \|\mu_{\text{real}} - \mu_{\text{fake}}\|_2) \tag{3.24}$$

This prevents the centroid structure from collapsing even as individual embeddings concentrate, maintaining the geometric separation necessary for detection.

Staged Training

Pretrain the authenticity head on a binary classification objective to establish initial separation, then introduce disentanglement losses with the separation structure as an anchor. This two-stage approach prevents the optimization from finding degenerate collapsed solutions by starting from a non-collapsed initialization.

3.6.4 Broader Implications

The systematic failure of variance-based disentanglement across multiple embedding types suggests fundamental limitations of this approach for deepfake detection. The core assumption—that authentic content shares intrinsic properties distinguishable from synthetic content through variance minimization—may be valid, but the optimization dynamics prevent recovery of this structure.

More broadly, these results highlight the challenge of multi-objective representation learning when objectives can be trivially satisfied through degenerate solutions. The Silhouette Paradox demonstrates that standard clustering metrics can misleadingly indicate success even when the learned representations fail to capture task-relevant structure.

Future work in disentangled deepfake detection should prioritize: (1) loss formulations that explicitly prevent collapse while encouraging separation, (2) regularization strategies that preserve baseline OOD generalization, and (3) evaluation frameworks that distinguish between genuine representation quality and degenerate clustering behavior.

3.6.5 Future Directions

The results, while not achieving the intended disentanglement, provide valuable diagnostic information about the challenges of multi-objective representation learning for deepfake detection. Several extensions remain for future work.

Multimodal Representation Fusion

Audio and video modalities capture complementary manipulation artifacts with moderate correlation ($r \approx 0.46$ between HuBERT and SENet predictions). Several fusion strategies merit investigation:

Early Fusion. Concatenating audio and video embeddings before disentanglement would allow dual projection heads to learn joint authenticity representations capturing cross-modal inconsistencies—for instance, lip movements not matching audio content.

Late Fusion. Training separate disentanglement models for each modality, then combining authenticity embeddings ($z^{\text{auth}}_{\text{audio}}$ and $z^{\text{auth}}_{\text{video}}$) at classification, preserves modality-specific representations while enabling joint decision-making.

Attention-Based Fusion. Cross-modal attention could dynamically weight modality contributions based on per-sample reliability, emphasizing whichever modality exhibits clearer artifacts.

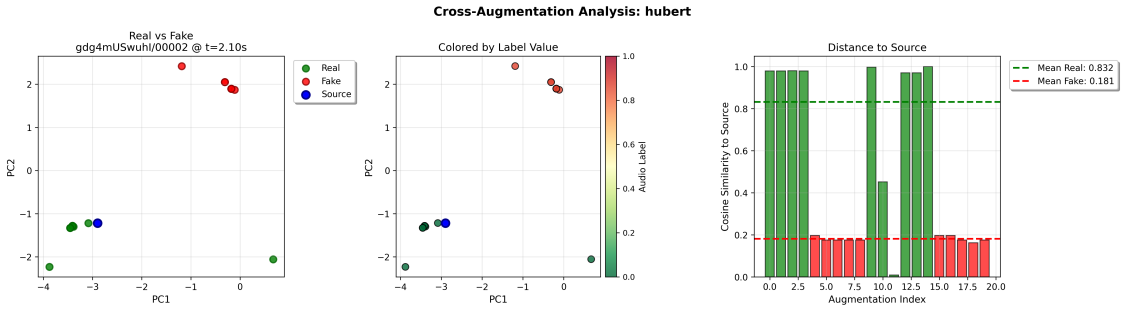
Alternative Loss Formulations

The identified failure modes motivate systematic exploration of alternative training objectives:

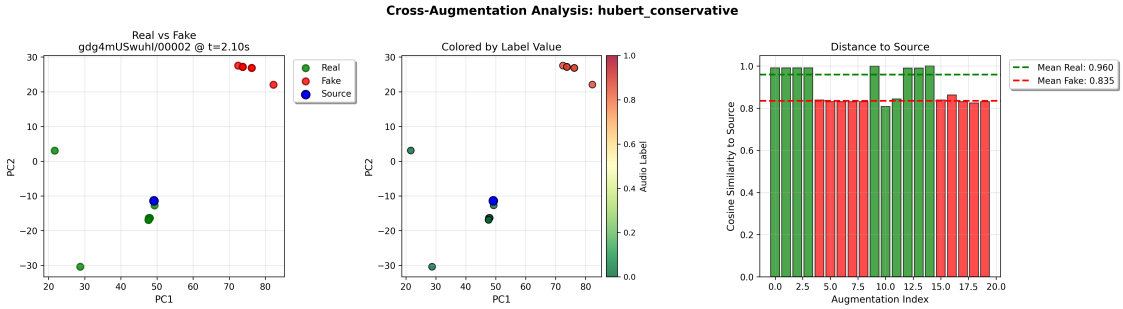
- **VICReg-style regularization** [93]: Combining variance, invariance, and covariance terms may better balance the competing objectives
- **Hyperspherical uniformity** [114]: Distributing embeddings uniformly on a hypersphere prevents collapse while maintaining separation
- **Spectral contrastive learning** [115]: Operating on the eigenspace of the similarity matrix may provide more stable gradients

Extended Evaluation

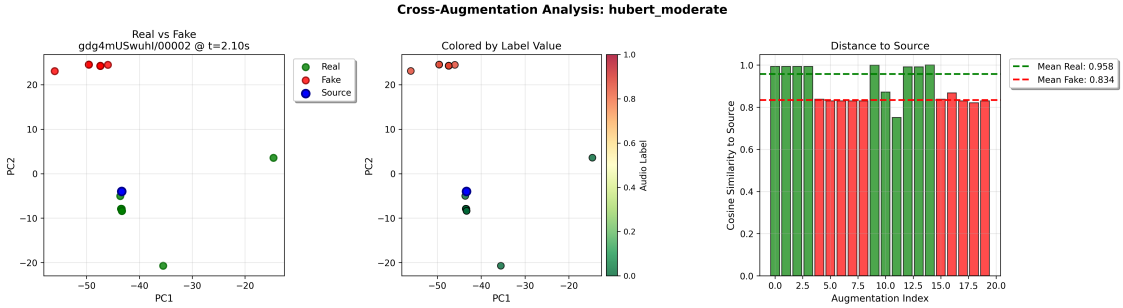
While Sora2 provides a challenging OOD test case, evaluation on additional unseen generators (Runway Gen-3, Pika, etc.) would further validate generalization claims. Additionally, ablation studies on the robustness to video compression, which can mask or mimic deepfake artifacts, is essential for practical deployment.



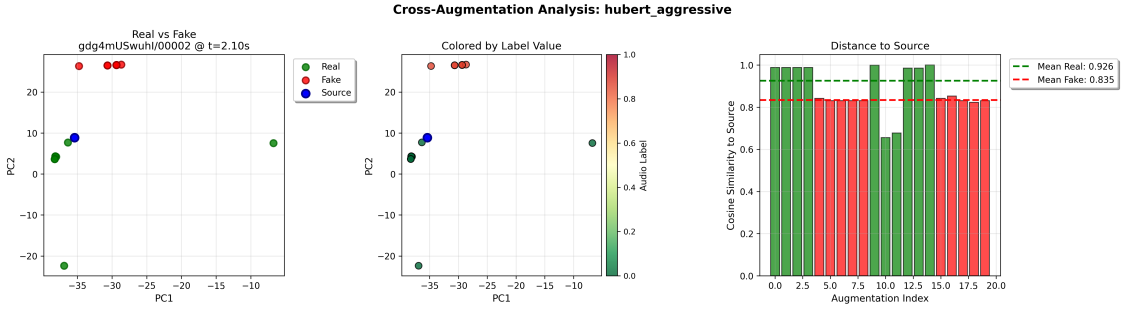
a) Original (Input) Baseline



b) Conservative Regularization ($\tau = 0.1$)

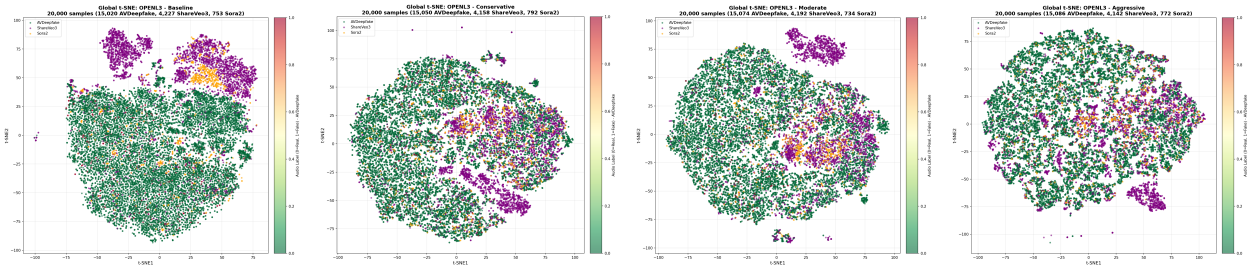


c) Moderate Regularization ($\tau = 0.2$)



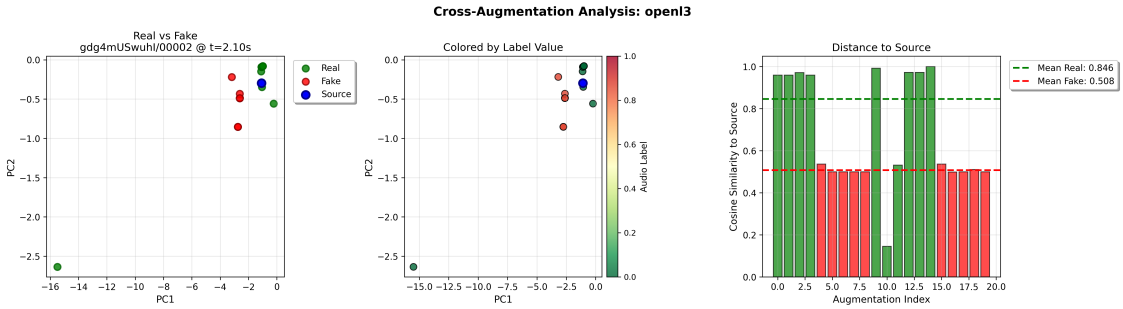
d) Aggressive Regularization ($\tau = 0.5$)

Figure 3.2: Per-video analysis for the projected embeddings across regularization schemes, shown via stacked visualization plots. Each row contains three panels: (left) PC1 vs. PC2 scatter plot colored by Real/Fake/Source labels, (center) the same scatter colored by continuous authenticity score, and (right) bar chart showing cosine similarity to the source embedding for each augmentation. The distance-to-source metric in the right panel should show greater differentiation between real (green) and fake (red) augmentations as regularization improves authenticity disentanglement.

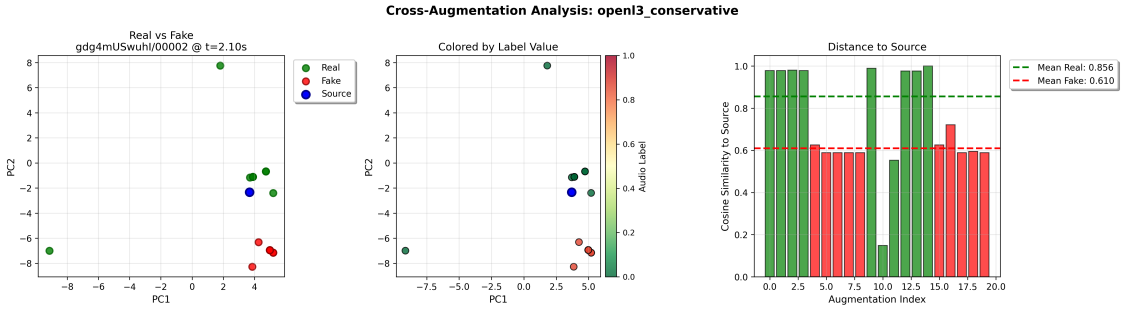


a) Input Baseline b) Conservative ($\tau = 0.1$) c) Moderate ($\tau = 0.2$) d) Aggressive ($\tau = 0.5$)

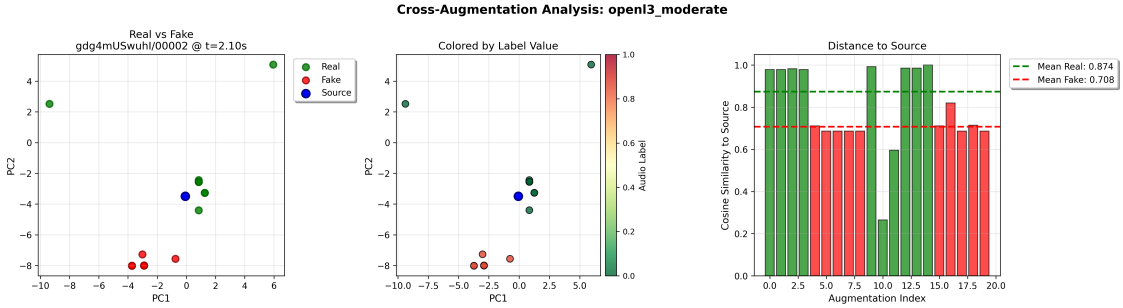
Figure 3.3: Visualization of the global embedding space for OpenL3 using t-SNE. The Original baseline (a) exhibits tighter clustering than HuBERT due to OpenL3's higher intrinsic similarity structure (K-means Silhouette 0.937). The projected schemes (b–d) are expected to show progressive collapse, with Moderate (c) achieving the best balance between compression and authenticity separation.



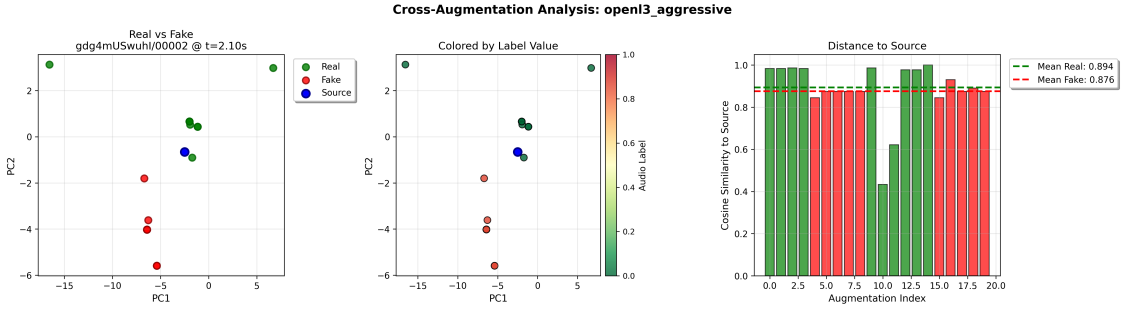
a) Original (Input) Baseline



b) Conservative Regularization ($\tau = 0.1$)

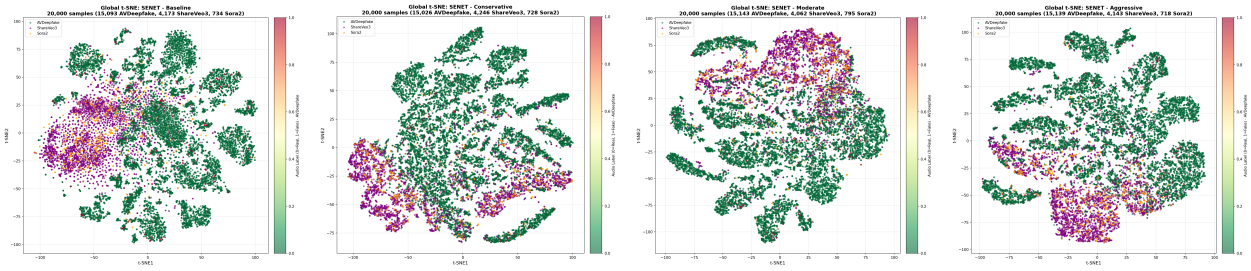


c) Moderate Regularization ($\tau = 0.2$)



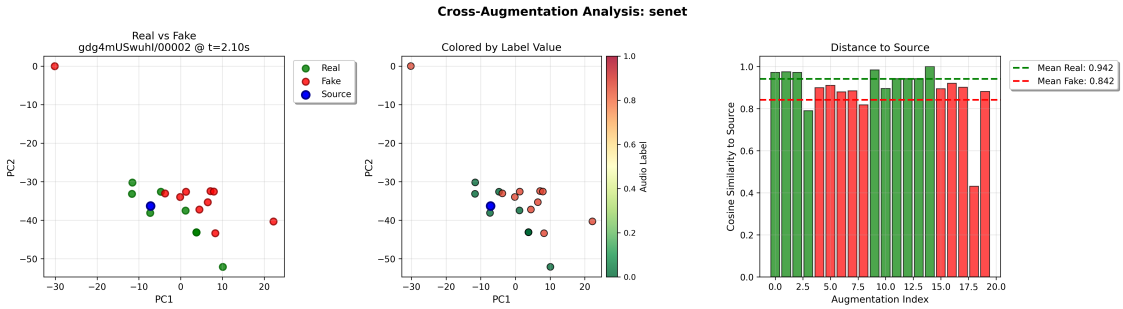
d) Aggressive Regularization ($\tau = 0.5$)

Figure 3.4: Per-video analysis for the projected OpenL3 embeddings across regularization schemes. Given OpenL3's superior baseline separation, the per-video plots should show clearer differentiation between real and fake augmentations in the Original baseline compared to HuBERT. The Moderate scheme (c) is expected to maintain this differentiation while the Conservative (b) and Aggressive (d) schemes may show degraded structure.

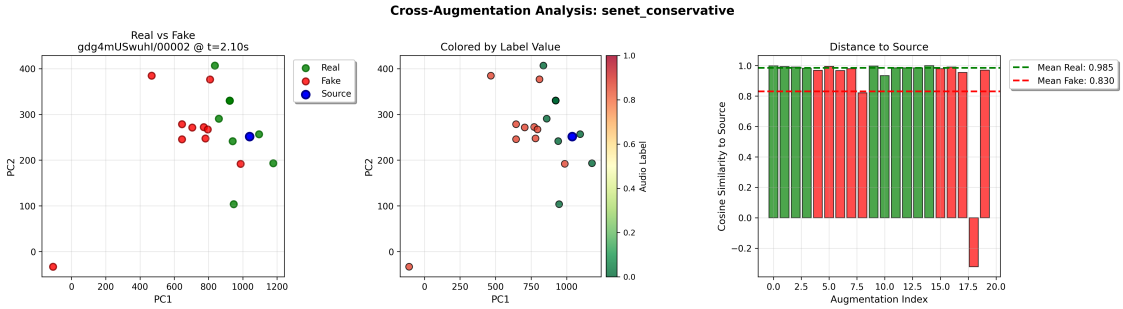


a) Input Baseline b) Conservative ($\tau = 0.1$) c) Moderate ($\tau = 0.2$) d) Aggressive ($\tau = 0.5$)

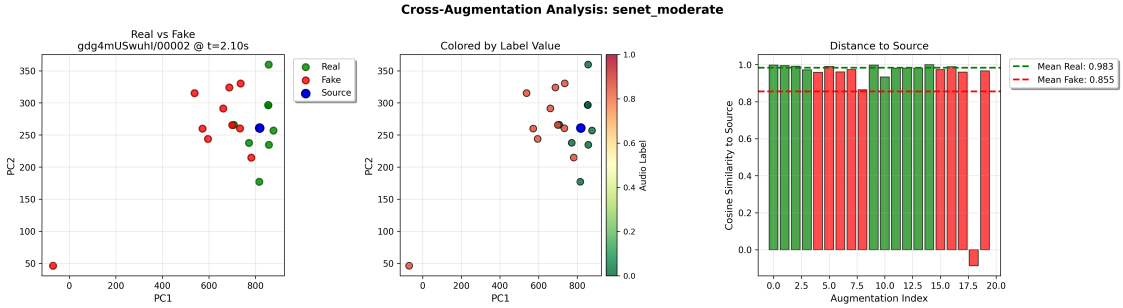
Figure 3.5: Visualization of the global embedding space for SENet using t-SNE. The Original baseline (a) exhibits the highest spread among all embedding types due to SENet's high-dimensional visual features. The projected schemes (b–d) are expected to show dramatic collapse, with Aggressive (d) maintaining slightly more spread while achieving the best label-alignment despite inverted centroid structure.



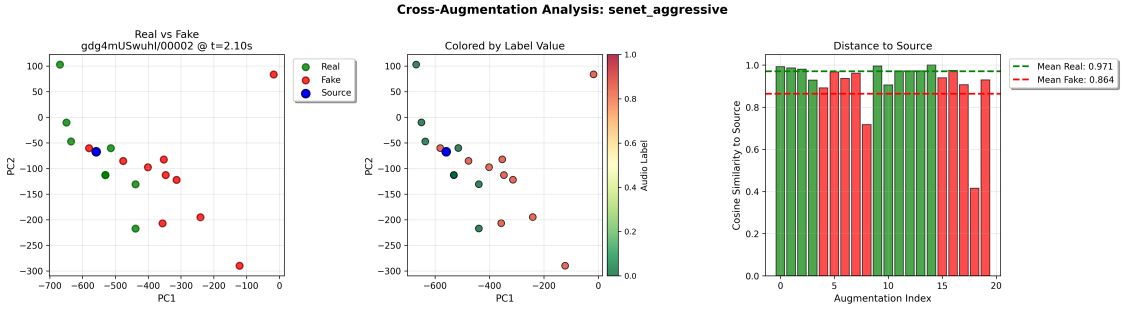
a) Original (Input) Baseline



b) Conservative Regularization ($\tau = 0.1$)



c) Moderate Regularization ($\tau = 0.2$)



d) Aggressive Regularization ($\tau = 0.5$)

Figure 3.6: Per-video analysis for the projected SENet embeddings across regularization schemes. Given SENet's strong baseline separation (+0.081 Separation Gap), the Original baseline (a) should show clear differentiation between real and fake augmentations in the distance-to-source bar chart. The projected schemes (b–d) may show reduced or inverted differentiation, reflecting the negative Separation Gaps observed in the quantitative metrics.

Chapter 4

Conclusion

4.1 Summary of Contributions

This thesis addressed two complementary challenges in visual deception detection, developing specialized approaches for textual homoglyph attacks and synthetic media identification.

For homoglyph detection, we introduced Visually-Aligned Text Embeddings (VA-TE), a contrastive learning framework that leverages vision-language model text encoders to capture visual properties of Unicode characters without requiring image rendering. Through curriculum-based hard negative mining and a lightweight projection architecture, VA-TE achieves 0.95 ROC-AUC as a standalone system. When combined with complementary string-similarity features in an ensemble, VA-TE attains state-of-the-art performance (0.98 ROC-AUC) while offering substantial advantages in scalability, memory efficiency, and deployment simplicity compared to prior image-based approaches. Building on recent advances in embedding refinement [116], this approach marks an important step toward multi-modal representations that integrate the visual characteristics of text, reflecting how humans read and interpret it.

For deepfake detection, we proposed a disentangled representation learning framework with dual projection heads and orthogonality constraints, designed to separate authenticity-relevant features from identity and content information. Our systematic evaluation across three embedding types (HuBERT, OpenL3, SENet) and three regularization schemes revealed both the promise and limitations of variance-based disentanglement. While all configurations exhibited representation collapse ($> 99\%$ Wasserstein distance reduction), the Aggressive regularization scheme on HuBERT embeddings achieved improved label-alignment (AMI: $0.111 \rightarrow 0.120$, ARI: $0.066 \rightarrow 0.089$) and positive out-of-distribution Separation Gaps on Sora2 content. However, for embeddings with strong baseline generalization (OpenL3, SENet), disentanglement training systematically degraded OOD performance, highlighting a fundamental tension between in-distribution optimization and generalization preservation. Additionally, we collected and annotated a novel out-of-distribution evaluation dataset comprising 150 videos (11,000+ segments) generated by OpenAI's Sora 2, providing a challenging benchmark for assessing generalization to state-of-the-art generation methods.

4.2 Methodological Insights

Several methodological insights emerged from this work that may inform future research in representation learning for security applications.

First, the success of VA-TE demonstrates that vision-language models encode visual characteristics of text beyond semantic meaning, and that these latent features can be effectively surfaced through targeted fine-tuning. The pairwise contrastive loss proved consistently superior to triplet, InfoNCE, and supervised contrastive alternatives, likely due to its alignment with the SigLIP encoder's original training objective.

Second, our deepfake detection experiments revealed what we term the *Silhouette Paradox*: K-means clustering metrics can improve substantially (e.g., $0.274 \rightarrow 0.575$ for HuBERT) even as label-alignment metrics degrade. This divergence occurs because representation collapse creates well-formed clusters based on residual variations unrelated to authenticity. This finding underscores the importance of evaluating learned representations with label-aware metrics rather than relying solely on unsupervised clustering quality.

Third, the multi-objective optimization underlying our disentanglement framework proved susceptible to degenerate solutions. The variance minimization objective, computed only on real samples, allows fake samples to collapse toward the real centroid while technically satisfying all training objectives. Notably, the optimal regularization strength proved embedding-dependent: Aggressive regularization ($\tau = 0.5$) performed best for HuBERT, while Moderate regularization ($\tau = 0.2$) achieved superior in-distribution results for OpenL3. This finding suggests that representation learning frameworks require careful hyperparameter tuning matched to the characteristics of their input embeddings.

Fourth, we observed a critical pattern in OOD generalization: embeddings with weak baseline structure (HuBERT, with negative baseline Separation Gap) benefited from disentanglement training, while embeddings with strong baseline structure (SENet, with +0.074 Separation Gap on Sora2) were harmed. This suggests that ID-focused optimization can overwrite the generalizable features encoded in pretrained representations, motivating future work on regularization strategies that explicitly preserve OOD structure.

Fifth, both projects benefited from combining learned representations with complementary features. For VA-TE, fusing visual embeddings with string-based metrics yielded a 3-point improvement in ROC-AUC. This pattern suggests that hybrid approaches leveraging multiple signal modalities remain valuable even as learned representations improve.

4.3 Broader Impact

Visual deception poses growing threats to financial systems, identity verification platforms, and information ecosystems. Homoglyph attacks enable deception at scale by exploiting the gap between human perception and computational string matching, while synthetic media undermines the trustworthiness of audio-visual content for authentication and evidence.

The methods developed in this thesis contribute to defending against these threats. VA-TE provides financial institutions with a scalable, deployable solution for detecting falsified account names and domain spoofing. The disentanglement framework, while requiring refinement to address representation collapse, establishes diagnostic tools and evaluation

metrics that clarify the challenges of multi-objective representation learning for deepfake detection. The identification of the collapse-separation trade-off as the primary failure mode offers a concrete target for future architectural improvements. More broadly, by focusing on learning stable properties of authentic content rather than chasing evolving adversarial signatures, both approaches embody a defensive paradigm better suited to the rapid pace of generative AI advancement.

We note that detection systems can also be misused—for instance, to identify which synthetic content evades detection, thereby improving adversarial attacks. Responsible deployment requires ongoing monitoring, regular model updates, and integration with broader security protocols rather than reliance on any single detection method.

4.4 Future Directions

Several promising directions extend this work. For VA-TE, applying the framework to non-Latin scripts (Chinese, Arabic, Cyrillic) and integrating with OCR pipelines for document verification represent natural next steps. The perceptual decoding extension proposed in Section 2.6.3 offers a path toward grounding text embeddings in low-level visual features.

For deepfake detection, the immediate priority is addressing representation collapse through alternative loss formulations. Supervised contrastive objectives that explicitly require separation between real and fake samples, bidirectional variance regularization that penalizes collapse of both distributions, and margin-based prototype learning that enforces minimum centroid distances all merit systematic evaluation. The strong baseline performance of SENet video embeddings (0.914 AUROC, +0.074 OOD Separation Gap) suggests that preserving, rather than overwriting, pretrained generalization structure may be more effective than aggressive representation learning.

Beyond addressing collapse, multimodal fusion strategies that combine audio and video authenticity signals offer substantial potential. The moderate correlation between HuBERT and SENet predictions ($r \approx 0.46$) indicates partially independent information that joint models could exploit. Attention-based fusion mechanisms that dynamically weight modality contributions based on per-sample reliability represent a promising direction.

Evaluation on additional unseen generators (Runway Gen-3, Pika, Kling) and across video compression levels will be essential for validating generalization claims before deployment. The Sora2 dataset introduced in this work provides one such benchmark, but broader coverage of the rapidly evolving generative model landscape is needed.

Finally, exploring unified architectures that address both textual and media-level visual deception within a single framework represents a longer-term opportunity, potentially leveraging shared principles of authenticity representation learning across modalities.

References

[1] Z. Cai, K. Kuckreja, S. Ghosh, A. Chuchra, M. H. Khan, U. Tariq, T. Gedeon, and A. Dhall. "Av-deepfake1m++: A large-scale audio-visual deepfake benchmark with real-world perturbations." In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 13686–13691.

[2] W. Wang and Y. Yang. "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models." *Advances in Neural Information Processing Systems*, **37**, 2024, pp. 65618–65642.

[3] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks." In: *International Conference on Machine Learning*. 2018, pp. 794–803.

[4] R. Dhamija, J. D. Tygar, and M. Hearst. "Why phishing works." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 581–590. ISBN: 1595933727. DOI: 10.1145/1124772.1124861. URL: https://doi.org/10.1145/1124772.1124861.

[5] A. Linari, F. Mitchell, D. Duce, and S. Morris. "Typo-Squatting: The Curse of Popularity." In: *WebSci'09: Society On-Line*. 2009.

[6] T. Liu, Y. Zhang, J. Shi, Y. Jing, Q. Li, and L. Guo. "Towards quantifying visual similarity of domain names for combating typosquatting abuse." In: *MILCOM 2016-2016 IEEE Military Communications Conference*. IEEE. 2016, pp. 770–775.

[7] R. Palacios, A. Sinha, and A. Gupta. "Automatic processing of brazilian bank checks." *Cited in US Pat*, (7,900,822), 2002.

[8] O. Goga, G. Venkatadri, and K. P. Gummadi. "The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks." In: *Proceedings of the 2015 Internet Measurement Conference*. IMC '15. Tokyo, Japan: Association for Computing Machinery, 2015, pp. 141–153. ISBN: 9781450338486. DOI: 10.1145/2815675.2815699. URL: https://doi.org/10.1145/2815675.2815699.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*, 2013.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision." In: *International conference on machine learning*. 2021, pp. 8748–8763.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[12] Y. Zhou, C. Li, G. Huang, Q. Guo, H. Li, and X. Wei. "A Short-Text Similarity Model Combining Semantic and Syntactic Information." *Electronics*, **12**(14), 2023. ISSN: 2079-9292. DOI: 10.3390/electronics12143126. URL: https://www.mdpi.com/2079-9292/12/14/3126.

[13] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." In: *Soviet Physics-Doklady*. Vol. 10. 8. 1966.

[14] K. P. Kalyanathaya, D. Akila, and G. Suseendren. "A fuzzy approach to approximate string matching for text retrieval in NLP." *J. Comput. Inf. Syst. USA*, **15**(3), 2019, pp. 26–32.

[15] R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality reduction by learning an invariant mapping." In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. Vol. 2. 2006, pp. 1735–1742.

[16] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi. "A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges." In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. June 2025, pp. 1587–1606.

[17] J. Woodbridge, H. S. Anderson, A. Ahuja, and D. Grant. "Detecting homoglyph attacks with a siamese neural network." In: *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*. Institute of Electrical and Electronics Engineers Inc., Aug. 2018, pp. 22–28. ISBN: 9780769563497. DOI: 10.1109/SPW.2018.00012.

[18] V. R. and S. K.P. "Siamese neural network architecture for homoglyph attacks detection." *ICT Express*, **6**, 1, Mar. 2020, pp. 16–19. ISSN: 24059595. DOI: 10.1016/j.icte.2019.05.002.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. PMLR, July 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

[20] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[21] A. van den Oord, Y. Li, and O. Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748*, 2018.

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. "Supervised contrastive learning." *Advances in neural information processing systems*, **33**, 2020, pp. 18661–18673.

[23] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. "Curriculum learning." In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Association for Computing Machinery, 2009, pp. 41–48. ISBN: 9781605585161. DOI: 10.1145/1553374.1553380. URL: https://doi.org/10.1145/1553374.1553380.

[24] F. J. Damerau. "A technique for computer detection and correction of spelling errors." *Commun. ACM*, **7**, 3, Mar. 1964, pp. 171–176. ISSN: 0001-0782. DOI: 10.1145/363958.363994. URL: https://doi.org/10.1145/363958.363994.

[25] A. Ginsberg and C. Yu. "Rapid Homoglyph Prediction and Detection." In: *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. 2018, pp. 17–23. DOI: 10.1109/ICDIS.2018.00010.

[26] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. "A neural probabilistic language model." *Journal of machine learning research*, **3**(Feb), 2003, pp. 1137–1155.

[27] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. "End-to-end learning of deep visual representations for image retrieval." *International Journal of Computer Vision*, **124**(2), 2017, pp. 237–254.

[28] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[29] C.-C. Wang, C.-T. Chiu, C.-T. Huang, Y.-C. Ding, and L.-W. Wang. "Fast and Accurate Embedded DCNN for Rgb-D Based Sign Language Recognition." In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 1568–1572. DOI: 10.1109/ICASSP40776.2020.9054076.

[30] N. V. Rao, A. Sastry, A. Chakravarthy, and P. Kalyanchakravarthi. "OPTICAL CHARACTER RECOGNITION TECHNIQUE ALGORITHMS." *Journal of Theoretical & Applied Information Technology*, **83**(2), 2016.

[31] P. Krishnan, K. Dutta, and C. Jawahar. "Deep Feature Embedding for Accurate Recognition and Retrieval of Handwritten Text." In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016, pp. 289–294. DOI: 10.1109/ICFHR.2016.0062.

[32] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew. "Deep Learning for Instance Retrieval: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(6), 2023, pp. 7270–7292. DOI: 10.1109/TPAMI.2022.3218591.

[33] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. "A survey on contrastive self-supervised learning." *Technologies*, **9**(1), 2020, p. 2.

[34] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji. "Dual contrastive learning for general face forgery detection." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 2022, pp. 2316–2324.

[35] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. "Signature Verification using a "Siamese" Time Delay Neural Network." In: *Advances in Neural Information Processing Systems*. Ed. by J. Cowan, G. Tesauro, and J. Alspector. Vol. 6. Morgan-Kaufmann, 1993. URL: https://proceedings.neurips.cc/paper_files/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf.

[36] Y. Zhai, X. Guo, Y. Lu, and H. Li. "In Defense of the Classification Loss for Person Re-Identification." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 1526–1535. DOI: 10.1109/CVPRW.2019.00194.

[37] A. Hermans, L. Beyer, and B. Leibe. "In defense of the triplet loss for person re-identification." *arXiv preprint arXiv:1703.07737*, 2017.

[38] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka. "CONTRASTIVE LEARNING WITH HARD NEGATIVE SAMPLES." In: *International Conference on Learning Representations (ICLR)*. 2021.

[39] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. "Curricularface: adaptive curriculum learning loss for deep face recognition." In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5901–5910.

[40] G. Chu, X. Wang, C. Shi, and X. Jiang. "CuCo: Graph representation with curriculum contrastive learning." In: *IJCAI*. 2021, pp. 2300–2306.

[41] X. Wang, Y. Chen, and W. Zhu. "A Survey on Curriculum Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 9, 2022, pp. 4555–4576. DOI: 10.1109/TPAMI.2021.3069908.

[42] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann. "Self-Paced Curriculum Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*, **29**, 1, Feb. 2015. DOI: 10.1609/aaai.v29i1.9608. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9608.

[43] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. "Automated Curriculum Learning for Neural Networks." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. PMLR, July 2017, pp. 1311–1320. URL: https://proceedings.mlr.press/v70/graves17a.html.

[44] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. "Multi-modal curriculum learning for semi-supervised image classification." *IEEE Transactions on Image Processing*, **25**(7), 2016, pp. 3249–3260.

[45] J. Wei, C. Huang, S. Vosoughi, Y. Cheng, and S. Xu. "Few-Shot Text Classification with Triplet Networks, Data Augmentation, and Curriculum Learning." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 5493–5500. DOI: 10.18653/v1/2021.naacl-main.434. URL: https://aclanthology.org/2021.naacl-main.434/.

[46] J. Zhang, J. Huang, S. Jin, and S. Lu. "Vision-Language Models for Vision Tasks: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 8, 2024, pp. 5625–5644. DOI: 10.1109/TPAMI.2024.3369699.

[47] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. "Multimodal neurons in artificial neural networks." *Distill*, **6**(3), 2021, e30.

[48] J. B. Clemens-Alexander and Denzler. "Not Just a Matter of Semantics: The Relationship Between Visual and Semantic Similarity." In: *Pattern Recognition*. Ed. by Simone, J. X. F. G. A., and Frintrop. Springer International Publishing, 2019, pp. 414–427. ISBN: 978-3-030-33676-9.

[49] F. Bordes et al. "An Introduction to Vision-Language Modeling." *arXiv preprint arXiv:2405.17247*, May 2024. URL: http://arxiv.org/abs/2405.17247.

[50] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. "Sigmoid Loss for Language Image Pre-Training." In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 11941–11952. ISBN: 979-8-3503-0718-4. DOI: 10.1109/ICCV51070.2023.01100.

[51] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision." In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916.

[52] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. "Flava: A foundational language and vision alignment model." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 15638–15650.

[53] J. Li, D. Li, C. Xiong, and S. Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.

[54] J. Li, D. Li, S. Savarese, and S. Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.

[55] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *Transactions on Machine Learning Research*, **Aug 2022**, 2022. URL: https://arxiv.org/abs/2205.01917.

[56] H. Liu, C. Li, Q. Wu, and Y. J. Lee. "Visual instruction tuning." *Advances in neural information processing systems*, **36**, 2023, pp. 34892–34916.

[57] OpenAI. *GPT-4V(ision) System Card*. Tech. rep. OpenAI, Sept. 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[58] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. "FLAVA: A Foundational Language And Vision Alignment Model." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 15617–15629. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.01519.

[59] S. Faizullah, M. S. Ayub, S. Hussain, and M. A. Khan. "A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges." *Applied Sciences*, **13**(7), 2023. ISSN: 2076-3417. DOI: 10.3390/app13074584. URL: https://www.mdpi.com/2076-3417/13/7/4584.

[60] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. "Accurate, large minibatch sgd: Training imagenet in 1 hour." *arXiv preprint arXiv:1706.02677*, 2017.

[61] E. Hoffer, I. Hubara, and D. Soudry. "Train longer, generalize better: closing the generalization gap in large batch training of neural networks." *Advances in neural information processing systems*, **30**, 2017.

[62] H. Khalid, M. Kim, S. Tariq, and S. S. Woo. "Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors." In: *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*. ACM. 2021, pp. 7–15. DOI: 10.1145/3476099.3484315.

[63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio. "Generative adversarial networks." en. *Communications of the ACM*, **63**(11), 2020, pp. 139–144.

[64] A. Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, and K. Kavukcuoglu. *Wavenet: A generative model for raw audio*. ca. arXiv preprint arXiv:1609.03499, 12, 1. 2016.

[65] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." en. In: *International Conference on Learning Representations*. Feb. 2018.

[66] T. Karras, S. Laine, and T. Aila. "A style-based generator architecture for generative adversarial networks." en. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[67] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. "Face2face: Real-time face capture and reenactment of rgb videos." en. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2387–2395.

[68] R. Chen, X. Chen, B. Ni, and Y. Ge. "Simswap: An efficient framework for high fidelity face swapping." en. In: *Proceedings of the 28th ACM international conference on multimedia*. Oct. 2020, pp. 2003–2011.

[69] R. Babaei, S. Cheng, R. Duan, and S. Zhao. "Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis." en. *Journal of Sensor and Actuator Networks*, **14**(1), 2025, p. 17.

[70] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics." en. In: *International conference on machine learning*. pmlr. June 2015, pp. 2256–2265.

[71] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models." en. *Advances in neural information processing systems*, **33**, 2020, pp. 6840–6851.

[72] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. "Leveraging frequency analysis for deep fake image recognition." en. In: *International conference on machine learning*. PMLR, Nov. 2020, pp. 3247–3258.

[73] Y. Li and S. Lyu. *Exposing deepfake videos by detecting face warping artifacts*. jv. arXiv preprint arXiv:1811.00656. 2018.

[74] F. Matern, C. Riess, and M. Stamminger. "Exploiting visual artifacts to expose deepfakes and face manipulations." en. In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW*. IEEE, Jan. 2019, pp. 83–92.

[75] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha. "Predicting heart rate variations of deepfake videos using neural ode." en. In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 2019, 0–0.

[76] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales. el-Latn. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400. 2020.

[77] U. Çiftçi, I. Demir, and L. Yin. "Deepfake source detection in a heart beat." en. *The Visual Computer*, **40**(4), 2024, pp. 2733–2750.

[78] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. "Lips don't lie: A generalisable and robust approach to face forgery detection." en. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5039–5049.

[79] S. Agarwal and H. Farid. "Detecting deep-fake videos from aural and oral dynamics." en. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 981–989.

[80] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. "Faceforensics++: Learning to detect manipulated facial images." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1–11.

[81] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. "Multi-attentional deepfake detection." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2185–2194.

[82] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. "Recurrent convolutional strategies for face manipulation detection in videos." *Interfaces (GUI)*, **3**(1), 2019, pp. 80–87.

[83] V. S. Katamneni and A. Rattani. "Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization." In: *2024 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2024, pp. 1–11. DOI: 10.48550/arXiv.2408.01532.

[84] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. "Emotions don't lie: An audio-visual deepfake detection method using affective cues." In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2823–2832.

[85] Y. Zhang, W. Lin, and J. Xu. "Joint audio-visual attention with contrastive learning for more general deepfake detection." *ACM Transactions on Multimedia Computing, Communications and Applications*, **20**(5), 2024, pp. 1–23.

[86] T. Wang and K. P. Chow. "Noise based deepfake detection via multi-head relative-interaction." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 14548–14556.

[87] Z. Liang, W. Liu, R. Wang, M. Wu, B. Li, Y. Zhang, and X. Yang. "Transfer Learning of Real Image Features with Soft Contrastive Loss for Fake Image Detection." en. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**(25), Apr. 2025, pp. 26281–26289.

[88] T. Reiss, B. Cavia, and Y. Hoshen. "Detecting Deepfakes Without Seeing Any." *arXiv preprint arXiv:2311.01458*, 2023.

[89] T. Qiao, S. Xie, Y. Chen, F. Retraint, and X. Luo. "Fully Unsupervised Deepfake Video Detection via Enhanced Contrastive Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**(7), 2024, pp. 4654–4668. DOI: 10.1109/TPAMI.2024.3356814.

[90] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 2013, pp. 1798–1828.

[91] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *International conference on learning representations*. 2017.

[92] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." *Advances in neural information processing systems*, **29**, 2016.

[93] A. Bardes, J. Ponce, and Y. LeCun. "Vicreg: Variance-invariance-covariance regularization for self-supervised learning." *arXiv preprint arXiv:2105.04906*, 2021.

[94] N. Tishby, F. C. Pereira, and W. Bialek. "The information bottleneck method." *arXiv preprint physics/0004057*, 2000.

[95] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. "Barlow twins: Self-supervised learning via redundancy reduction." In: *International conference on machine learning*. PMLR. 2021, pp. 12310–12320.

[96] OpenAI. *Sora 2: Next-Generation Text-to-Video Generation Model*. Tech. rep. OpenAI, 2025. URL: https://openai.com/index/sora-2/.

[97] Google DeepMind. *Veo: a text-to-video generation system*. Tech. rep. Google DeepMind, 2025. URL: https://deepmind.google/models/veo/.

[98] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello. "Look, listen, and learn more: Design choices for deep audio embeddings." In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 3852–3856.

[99] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 2021, pp. 3451–3460.

[100] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12449–12460.

[101] J. Hu, L. Shen, and G. Sun. "Squeeze-and-excitation networks." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141.

[102] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. "MARLIN: Masked Autoencoder for facial video Representation LearnINg." In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 1493–1504. DOI: 10.1109/CVPR52729.2023.00150.

[103] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. "ArcFace: Additive angular margin loss for deep face recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.

[104] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A unified embedding for face recognition and clustering." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 815–823.

[105] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. "MagFace: A universal representation for face recognition and quality assessment." In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14225–14234.

[106] N. X. Vinh, J. Epps, and J. Bailey. "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 1073–1080.

[107] L. Hubert and P. Arabie. "Comparing partitions." *Journal of classification*, **2**(1), 1985, pp. 193–218.

[108] P. J. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, **20**, 1987, pp. 53–65.

[109] S. Kullback and R. A. Leibler. "On information and sufficiency." *The annals of mathematical statistics*, **22**(1), 1951, pp. 79–86.

[110] J. Lin. "Divergence measures based on the Shannon entropy." *IEEE Transactions on Information theory*, **37**(1), 2002, pp. 145–151.

[111] C. Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2008.

[112] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein generative adversarial networks." In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[113] C. E. Shannon. "A mathematical theory of communication." *The Bell system technical journal*, **27**(3), 1948, pp. 379–423.

[114] T. Wang and P. Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In: *International conference on machine learning*. PMLR. 2020, pp. 9929–9939.

[115] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. "Provable guarantees for self-supervised deep learning with spectral contrastive loss." *Advances in neural information processing systems*, **34**, 2021, pp. 5000–5011.

[116] R. Jha, C. Zhang, V. Shmatikov, and J. X. Morris. *Harnessing the Universal Geometry of Embeddings*. 2025. arXiv: 2505.12540 [cs.LG]. URL: https://arxiv.org/abs/2505.12540.