MATCHED: Multimodal Authorship-Attribution To Combat Human Trafficking in Escort-Advertisement Data

Anonymous ACL submission

Abstract

Human trafficking (HT) remains a critical issue, with traffickers increasingly leveraging online escort advertisements to advertise victims anonymously. Existing detection methods, including text-based Authorship Attribution (AA), overlook the multimodal nature of these ads, which combine text and images. To bridge this gap, we introduce MATCHED, a multimodal dataset comprising 27,619 unique text descriptions and 55,115 unique images sourced from Backpage across seven U.S. cities in four geographic regions. This study extensively benchmarks text-only, vision-only, and multimodal baselines for vendor identification and verification tasks, employing multitask (joint) training objectives that achieve superior classification and retrieval performance on in-sample and out-of-data distribution datasets. The results demonstrate that while text remains the dominant modality, integrating visual features adds stylistic cues that enrich model performance. Moreover, text-image alignment strategies like CLIP and BLIP2 struggle due to low semantic overlap and vague connections between the modalities of escort ads, with end-to-end multimodal training proving more robust. Our findings emphasize the potential of multimodal AA to combat HT, providing Law Enforcement Agencies with robust tools to link advertisements and disrupt trafficking networks.

1 Introduction

Human trafficking (HT) is a pervasive crime exploiting individuals of all ages and genders, with sex trafficking being particularly prevalent. Traffickers coerce victims into commercial sex through violence, threats, deception, and debt bondage, mostly affecting women and girls (EUROPOL, 2020; UNDOC, 2020; ILO, 2012). Furthermore, the rise of digital platforms has enabled traffickers to exploit online advertisements (ads) for anonymity, overwhelming manual tracking efforts and leaving many cases undetected (POLARIS, 2020, 2018).

While end-to-end classification methods show promise in detecting HT (Alvari et al., 2016; Tong et al., 2017; Alvari et al., 2017), reliance on expert-generated labels risks overfitting and poor generalization. Therefore, Law Enforcement Agencies (LEAs) and researchers have developed HT indicators for identifying suspicious ads (Ibanez and Suthers, 2014; Ibanez and Gazan, 2016; Lugo-Graulich and Meyer, 2021). However, these indicators require grouping ads linked to individuals or networks. Traditional methods rely on phone numbers and email addresses (Chambers et al., 2019), yet research reports that only 37% of ads contain such identifiers (Saxena et al., 2023a). Supervised (Nagpal et al., 2015; Li et al., 2022a; Liu et al., 2023) and unsupervised techniques (Rabbany et al., 2018; Nair et al., 2022; Vajiac et al., 2023) often depend on explicit similarities (e.g., names, phrases, or near-duplicates), limiting effectiveness when vendors alter details to evade detection.

Authorship Attribution (AA) offers a more holistic approach by identifying unique language patterns and stylistic features across ads from the same vendor or group. NLP-based AA methods have successfully linked ads by analyzing subtle written expressions, even when explicit markers differ (Ardakani, 2020; Saxena et al., 2023a). However, existing AA research largely overlooks the multimodal nature of escort ads, which typically include text (title, description) and images. Integrating visual cues can enhance AA by capturing stylistic consistencies, locations, or poses that uniquely characterize a vendor's profile. For instance, vendors in larger networks may reuse images with varying text or pair similar text with different images. While current AA methods require at least five ads per vendor (Saxena et al., 2023a), leveraging multimodal AA (MAA) can improve performance for vendors with fewer ads by utilizing the multiple images typically present in each ad. This work aims to support LEAs in building AA-driven knowledge graphs and enabling targeted investigations across extensive collections of escort ads by making the following contributions:

(i). MATCHED Dataset and Comprehensive Benchmarking: We introduce MATCHED, a novel multimodal dataset for MAA, comprising 27,619 unique text descriptions and 55,115 images collected from Backpage escort ads across 7 U.S. cities between December 2015–April 2016. We establish benchmarks for text, vision, and multimodal domains, evaluating performance on both in-sample and out-of-data (OOD) distribution datasets. MATCHED provides a robust foundation for future MAA research. Due to sensitivity, anonymized metadata is shared via Dataverse, with the full dataset restricted and only accessible through requests. Our code is available at MATCHED.

(ii). Enhanced Performance through Multitask Training: We propose a joint multitask framework that simultaneously optimizes vendor identification and verification, outperforming traditional single-task models by 1.61% (text) and 1.52% (vision) on macro-F1 score for classification and 1.68% (text) and 6.75% (vision) on R-Precision for retrieval task. Although these gains may seem subtle, this dual-focus approach empowers LEAs to identify known vendors and discover emerging ones in OOD ads, enhancing their investigative capabilities.

(iii). Advancements in Model Performance through Multimodal Training: Traditional AA methods rely heavily on textual data, often ignoring valuable stylistic cues from images and excluding vendors with fewer ads. Our multimodal approach integrates text and image data, improving performance even for vendors with limited postings. Pairing a single text description with multiple images (e.g., one text with five images produces five samples) expands the training set and enriches feature representation. While text remains the dominant modality, incorporating images with text enhances text-only results by 5.43% on retrieval R-Precision, marginally improves vision-only results by 0.75% on retrieval R-Precision, and increases classification macro-F1 by 32.62%—ultimately providing a more comprehensive and robust AA framework.

2 Related Research

AA in NLP has advanced from basic stylometric analysis (Bhargava et al., 2013; Ramnial et al., 2016) to sophisticated models detecting distinct linguistic patterns across text segments (Fabien et al., 2020; Ai et al., 2022; Wegmann et al., 2022). AA applications span forensic linguistics, aiding attributing authorship in legal contexts (Iqbal et al., 2008; Nirkhi and Dharaskar, 2013; Fobbe, 2021), to cybersecurity, where it tracks malicious actors and criminal activity across platforms (Zhang et al., 2019; Saxena et al., 2023b). However, applying AA to online criminal markets presents unique challenges: conventional models struggle to capture the specialized jargon, coded language, and noise prevalent in illicit environments like illegal criminal marketplaces (Choshen et al., 2019; Manolache et al., 2022). This gap highlights the need for fine-tuned models that adapt to the nuanced linguistic and stylistic shifts in these contexts.

Therefore, Ardakani (2020) proposed supervised neural networks for AA on Backpage escort ads, uncovering stylistic consistencies even when explicit identifiers are altered. Similarly, Saxena et al. (2023a) leverage transformer-based models for vendor identification and verification, effectively linking ads across 41 cities. In addition to text, images in criminal markets can also reveal recurring stylistic patterns, such as backgrounds, lighting, or object placement, complementing linguistic cues when text data is sparse or inconsistent (Cotogni et al., 2024; Wang et al., 2018). Multimodal AA (MAA) approaches leverage these text and images, enhancing accuracy by merging stylistic patterns across media and creating comprehensive vendor profiles (Zhang et al., 2019).

This research introduces a novel multimodal dataset, MATCHED, of escort ads collected from seven U.S. cities across four geographical regions. Using a multitask training approach on the MATCHED dataset, we establish benchmarks for text, vision, and multimodal domains in escort market ads, laying a foundation for future MAA research. Our models optimize vendor identification (classifying ads to specific vendors) and verification (assessing if two ads are from the same vendor) through this unified training objective. This enables LEAs to identify known vendors in closed-set environments and link emerging vendors across out-of-data distribution ads in open-set scenarios. Finally, our multimodal approach leverages textual and visual cues, enabling LEAs to track HT networks more precisely across various online markets and platforms, laying the groundwork for advanced AA research. Integrating this multimodal data, especially for vendors with limited text ads, further enhances model performance by creating multiple samples per ad.

3 Dataset

Regions	Ads	Text	Images	% Faces	Vendors
South	14088	13661	27423	0.4928	1450
Midwest	8564	8259	14883	0.5542	1008
West	3262	3153	5049	0.6052	507
Northeast	2599	2546	7760	0.6183	584
All	28513	27619	55115	0.5676	3549

Table 1: Number of advertisements, unique text descriptions, images, % of Faces in the image datasets, and vendors per region in the MATCHED dataset.

Lugo-Graulich and Meyer (2021) provides compelling evidence linking Backpage escort advertisements to HT, motivating our focus on Backpage ads. We curate a dataset of 28,513 ads, comprising 27,619 unique text descriptions and 55,115 unique images associated with 3,549 vendors. Approximately 56% of the images feature an escort's face, while the remaining 44% display partial body images (without faces). To establish ground truth for AA tasks, we follow Saxena et al. (2023a), extracting phone numbers using Chambers et al. (2019) and leveraging NetworkX (Hagberg et al., 2008) to form vendor communities. Each community is assigned a unique vendor label, enabling robust AA analysis. Since the vendor label generation process is based on existing literature, detailed steps for phone number extraction and vendor label creation are provided in Appendices A.2–A.3.

The dataset spans seven major U.S. cities—Chicago, Houston, Detroit, Dallas, San Francisco, New York, and Atlanta—representing four geographic regions: South, Midwest, West, and Northeast. These regions group ads by city, with average text sequence lengths of 125, 118, 113, and 132 tokens, respectively. Detailed statistics, including vendor overlap between regions, text and image ad similarity, sentence and character lengths, and the frequency of text, image, and multimodal ads per vendor, are provided in Appendix A.2 (Figure 2b). The South region dataset, containing the largest number of text and image ads, is the primary dataset for training and in-distribution evaluation. The Midwest, West, and Northeast datasets are used as OOD datasets to evaluate model generalization. Notably, many vendors appear across multiple regions, meaning the OOD datasets include ads from vendors present in the South dataset as well as additional region-specific vendors.

4 Experimental Setup

Our experiments address 2 AA tasks critical for disrupting HT networks: vendor identification (closed-set classification) and vendor verification (open-set metric learning). Vendor identification determines whether an ad originates from a known vendor in a predefined candidate set. In contrast, vendor verification assesses whether two ads belong to the same vendor, including vendors unseen during training. We evaluate these tasks using text-only, vision-only, and multimodal baselines on the South region dataset and test OOD generalization on Midwest, West, and Northeast datasets. Complete implementation details are provided in Appendix A.4.

(i). Vendor Identification Task: For vendor identification, we perform multi-class classification using pre-trained backbones with a classification head on the South region dataset. We optimize models with cross-entropy (CE) loss (Juola and Baayen, 2005) and a multitask joint objective combining CE with supervised contrastive (SupCon) (Ye et al., 2023) and triplet losses (Hu et al., 2020). These multitask joint training objectives, referred to as CE+SupCon and CE+Triplet, enhance feature discrimination by aligning representations of ads from the same vendor while separating those from different vendors.

(ii). Vendor Verification Task: Since the vendor verification task aims to compare vendor ads based on content similarity, we employ contrastive learning with triplet and SupCon losses (Kaya and Bilge, 2019; Wegmann et al., 2022) to learn discriminative ad embeddings. These embeddings cluster ads from the same vendor while separating those from different vendors, enabling retrieval of all ads linked to a vendor—including those outside the training set—via FAISS-based similarity search (Johnson et al., 2019).

(iii). Baselines: Following (Saxena et al., 2023a), text-only baselines utilizes Style-Embedding (Wegmann et al., 2022) and DeCLUTR-small (Giorgi et al., 2021) backbones, whereas vision-only baselines utilizes VGG-16 (Simonyan and Zisserman, 2015), ResNet-50 (He et al., 2015), DenseNet-121 (Huang et al., 2018), InceptionNetV3 (Szegedy et al., 2015), EfficientNetV2 (Tan and Le, 2021), ConvNext-small (Woo et al., 2023), and ViT-base-patch16-244 (Dosovitskiy et al., 2021) backbones. The text-only and vision-only baselines are fine-tuned with CE, CE+Triplet, and CE+SupCon objectives for vendor identification tasks and Triplet or SupCon objectives for vendor verification tasks.

The multimodal baselines utilize VisualBERT (Li et al., 2019), ViLT (Kim et al., 2021), and a custom DeCLUTR-ViT backbone (combining DeCLUTR for text and ViT for images) with four fusion strategies—concatenation (Gallo et al., 2018; Li et al., 2024b), mean pooling (Sleeman et al., 2022), self-attention (Kiela et al., 2020; Gan et al., 2024), and adaptive auto fusion via a neural network (Sahu and Vechtomova, 2021), enabling nuanced cross-modal interactions by combining complementary signals. Finally, we employ the DeCLUTR-ViT backbone to also perform image-text alignment pre-training task on the combined dataset from all regions, applying three alignment strategies: Image-Text Contrastive (ITC, aka CLIP) (Radford et al., 2021), ITC+ITM (Image-Text Contrastive and Image-Text Matching) (Villegas et al., 2024), and BLIP2 (Li et al., 2023). These alignment techniques ensure that text and images from the same ad are represented closely in the latent space, particularly when a single text ad is associated with multiple images. Once the pre-training is completed, these backbones are fine-tuned similarly to other baselines with CE and CE+SupCon for vendor identification on the South-region dataset.

(iv). Evaluation: All the baselines in our research are evaluated for classification and retrieval tasks. Due to the class imbalance in our datasets (Figure 2b), we evaluate our classifiers on the Macro-F1 metric. Additionally, we evaluate all our models on a retrieval task focused on assessing the model's ability to find stylometric similarities between writing and photometric styles in our escort ads. To perform retrieval, the dataset is split into training ("documents") and test ("queries") sets, with text, image, and multimodal embeddings generated by trained models to compute cosine similarity via FAISS-based similarity-search operation. Text-only and vision-only baselines extract embeddings directly from their respective encoders, while multimodal baselines, including the CLIP baseline with ITC objective, combine text and vision embeddings from the DeCLUTR-ViT backbone using a mean pooling strategy. For ITC+ITM and BLIP2-based baselines, we take these image embeddings from the QFormer encoder. The retrieval tasks are categorized as text-to-text, image-to-image, or multimodal based on whether query and document embeddings are derived from text, vision, or pooled multimodal representations.

All the retrieval tasks are evaluated using R-Precision@X, which measures precision when the number of retrieved items equals the number of relevant ads per vendor, with higher scores reflecting more accurate representations of vendor activity (Saxena et al., 2023a). Additionally, Mean Reciprocal Rank (MRR@10) evaluates the average ranking position of the first ten correctly retrieved ads for each query, with scores closer to 1 indicating higher relevance ranking, thereby reducing manual search efforts for LEAs (Striebel et al., 2024). Lastly, Macro-F1@X independently calculates and averages F1 scores for each vendor class, ensuring equal weight for all vendors regardless of sample size. In Macro-F1@X and R-Precision@X, X represents the cutoff, defined as the number of relevant items per vendor.

5 Results

This section evaluates text-only, vision-only, and multimodal baselines for vendor identification (classification) and verification (retrieval) tasks. Given the space constraints, we only compare the best-performing baselines in our manuscript. However, an extensive analysis of results from all the baselines is provided in Appendix Tables 4-11.

Model	Loss	Macro-F1
Text-Baseline		
DeCLUTR-small	CE	0.6379
	CE+Triplet	0.5503
	CE+SupCon	**0.6540**
Vision-Baseline		
ViT-base-patch16	CE	0.6142
	CE+Triplet	0.6378
	CE+SupCon	**0.6294**
Multimodal-Baselines		
End2End DeCLUTR-ViT	CE	0.9670
	CE+SupCon	**0.9802**
DeCLUTR-ViT	BLIP2+CE+SupCon	0.9420

Table 2: Macro-F1 performance of the text, vision, and multimodal classifiers on the south region dataset. The benchmarks are highlighted by **color**.

(i). Classification task: As illustrated in Table 2 and confirming prior findings (Saxena et al., 2023a), DeCLUTR (0.6379) outperforms Style-Embedding (0.5210) backbone with CE loss and achieves the highest macro-F1 (0.6540) with CE+SupCon amongst the text baselines. Amongst vision baselines, ResNet-50 with CE achieves the highest macro-F1 (0.6394), followed by EfficientNetV2

(0.6285), DenseNet-121 (0.6262), ConvNext-small (0.6215), and ViT-base-patch16 (0.6141). Despite its slight underperformance in classification tasks, insights from Appendix Table 6 reveal that ViT outperforms all other models in retrieval tasks for both in-sample and OOD distribution datasets. This finding aligns with prior research (Gkelios et al., 2021; El-Nouby et al., 2021), which highlights ViT's ability to produce rich, contextualized representations that capture global relationships and stylistic patterns, even across diverse visual data (e.g., images with or without faces), making ViT the most suitable backbone for our task. Finally, the ViT baseline trained with the CE+Triplet objective achieves the best macro-F1 of 0.6378, with CE+SupCon closely following at 0.6294.

Amongst the multimodal baselines, the end-to-end DeCLUTR-ViT backbone with mean pooling fusion achieves the highest macro-F1 (0.9670), surpassing VisualBERT (0.9355) and ViLT (0.7369). When fine-tuned, alignment baselines (CLIP, ITC+ITM, BLIP2) underperform compared to the end-to-end baseline, though BLIP2-pretrained DeCLUTR-ViT backbone comes closest to matching this performance (0.9420). When trained with the joint CE+SupCon objective, the DeCLUTR-ViT backbone performs exceptionally (0.9802) in capturing multimodal relationships. This performance is attributed to the dataset's structure, where each text ad is paired with multiple images and vice versa, ensuring the model encounters diverse combinations during training.

(ii). Retrieval Task: The retrieval task evaluates the effectiveness of metric learning (Triplet and SupCon losses) and joint-objective classifiers in clustering ad representations by vendor-specific stylometric patterns. The Zero-Shot (ZS) average reflects retrieval performance across datasets without task-specific training, and the OOD average measures the generalization of South-trained models to unseen regions.

Figure 1(A) compares the text-to-text retrieval performance of text-only pre-trained (●), fine-tuned, and multimodal baselines. Fine-tuning on the South region dataset significantly improves performance across all metrics. Among text-only baselines, the DeCLUTR backbone trained with the joint CE+SupCon objective (■) outperforms the CE-only baseline (■) and performs on-par with the SupCon-only baseline (■) on OOD avg score, while surpassing it on the training dataset. Given the consistent performance of the DeCLUTR backbone with CE+SupCon objective on the classification and retrieval task, we establish it as the benchmark for text-only modality. This benchmark is further compared against the text representations from the multimodal DeCLUTR-ViT backbone trained end-to-end with CE+SupCon (✚) and the fine-tuned DeCLUTR-ViT backbone, pre-trained for text-image alignment task using BLIP2 objective (✚). The multimodal backbone trained end-to-end with CE+SupCon consistently outperforms all baselines on training and OOD datasets.

Figure 1(B) highlights image-to-image retrieval performance, comparing vision-only pre-trained (●), fine-tuned, and multimodal baselines. Fine-tuning on image ads also improves retrieval performance. Amongst vision-only baselines, the ViT backbone trained with the CE+SupCon objective (■) achieves superior performance over other baselines on both training and OOD datasets, establishing itself as the benchmark for the vision-only modality. Despite the better performance of the ViT backbone with CE+Triplet objective (■) on classification, it underperforms on the retrieval task. We further compare this vision benchmark against the vision representations from the multimodal DeCLUTR-ViT backbone trained end-to-end with CE+SupCon (✚) and the fine-tuned DeCLUTR-ViT backbone, pre-trained for text-image alignment task using BLIP2 objective (✚). The end-to-end multimodal backbone with CE+SupCon objective consistently outperforms other baselines on OOD datasets. However, it underperforms the fine-tuned BLIP2 baseline on R-Precision and Macro-F1 metrics for the training dataset.

Figure 1(C) compares retrieval performance among multimodal baselines, evaluating the multimodal representation from the end-to-end multimodal DeCLUTR-ViT backbone trained end-to-end with CE+SupCon (✚) and the fine-tuned DeCLUTR-ViT backbone, pre-trained for text-image alignment task using BLIP2 objective (✚). The end-to-end multimodal backbone with CE+SupCon objective consistently outperforms the other baseline across the training and OOD datasets. Our analysis (Appendix Table 8) indicates that the low performance of the text-image alignment strategies can be attributed to the lack of semantic similarity between images and text, as images in escort ads often do not directly reflect the context of the accompanying text.

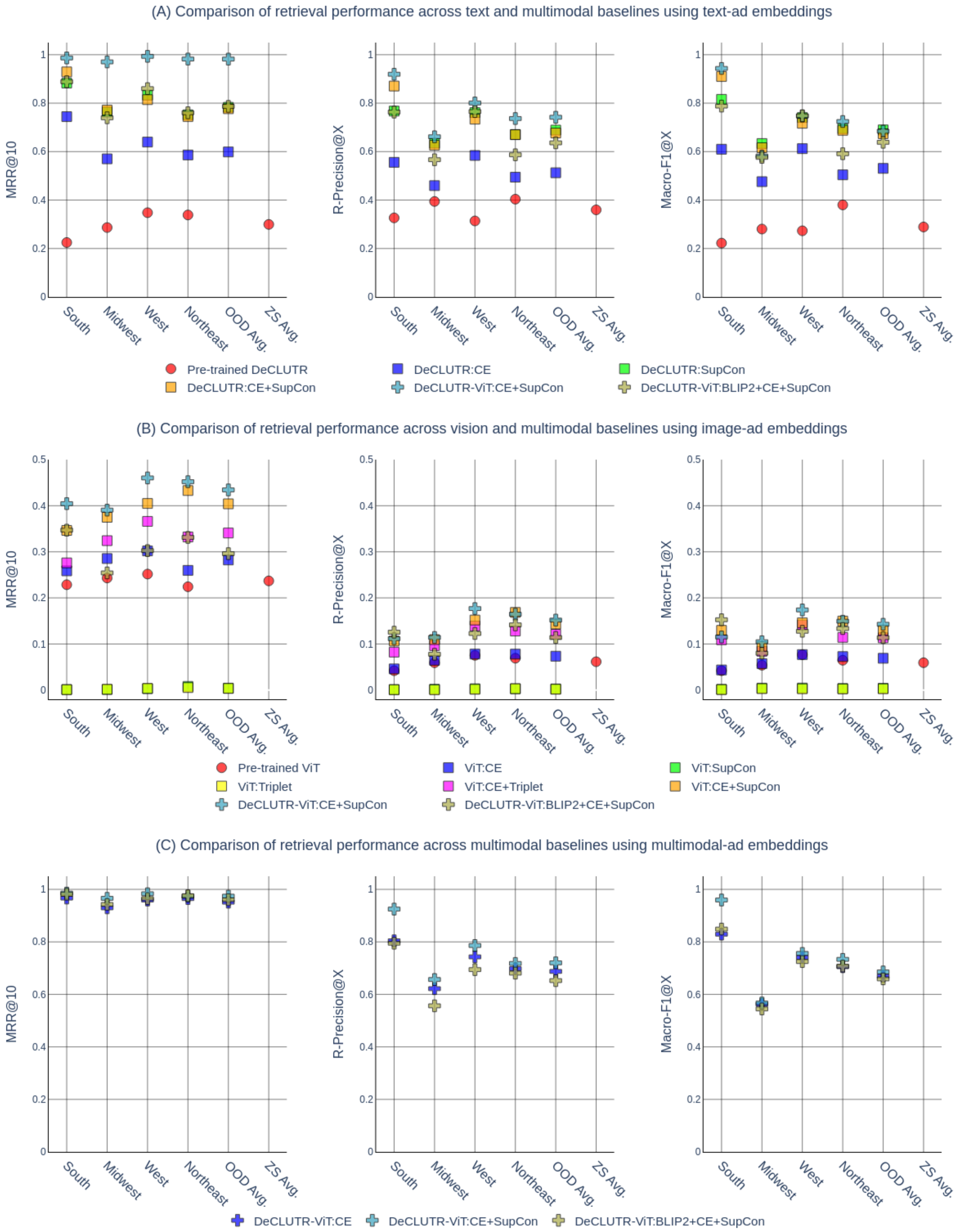


Figure 1: Comparison of retrieval performance across multiple baselines for text-to-text, image-to-image, and multimodal ads retrieval tasks on South, Midwest, West, and Northeast datasets. The text-to-text retrieval baselines include the pre-trained DeCLUTR checkpoint (●), DeCLUTR classifiers trained on CE (■) and CE+SupCon losses (■), and the DeCLUTR backbone trained with SupCon loss (■). Image-to-image retrieval baselines include the pre-trained ViT checkpoint (●), ViT classifiers trained on CE (■), CE+Triplet (■), and CE+SupCon losses (■), and ViT backbones trained with SupCon (■) and Triplet (■) losses. Multimodal baselines include End2End DeCLUTR-ViT classifiers trained with CE (✚), CE+SupCon (✚), and BLIP2-aligned DeCLUTR-ViT classifiers trained with CE+SupCon (✚) objectives.

6 Key Takeaways, Result Analysis, & Further Insights

(i). The experiments above demonstrate that fine-tuning on the MATCHED dataset significantly enhances retrieval performance, underscoring its value and exposing the limitations of existing pre-trained checkpoints in adapting to the unique linguistic and stylistic patterns of escort ads.

(ii). Given the dual objective of achieving in-sample and OOD distribution performance, the CE+SupCon joint objective consistently outperforms or matches other training objectives, demonstrating robustness and generalization. This dual focus enables models to effectively address closed-set vendor identification (linking ads to known vendors in LEA databases) and open-set vendor verification (connecting ads from emerging vendors on new platforms). While some baselines excel at one task, our benchmarks are established based on their ability to perform well across both objectives, ensuring practical utility for LEAs in tracking known and emerging HT networks.

(iii). Multimodal integration significantly enhances AA performance by leveraging complementary textual and visual features to capture richer authorship patterns. Beyond the quantitative improvements shown in Table 2 and Figure 1, our qualitative analysis in Appendix A.6 (Figure 3) reveals that multimodal training improves classification performance across all vendors, including those with lower class frequency. It also better connects images without faces, performs more effectively for vendors advertising multiple escorts, and increases true positive rates while reducing false positives. Similarly, observations in Appendix Figures 4-7 also confirm this pattern across retrieval tasks.

(iv). While integrating text and vision features enhances vision retrieval performance compared to vision-only baselines, vision remains less reliable (Figure 1(B)). Conversely, integrating vision features into text representations significantly boosts text retrieval performance, with text consistently outperforming vision and multimodal representations (Figure 1(A) and (C)). This highlights the superiority of the text representations from the DeCLUTR-ViT backbone, making it the most effective option for retrieval tasks on our dataset.

(v). While the multimodal DeCLUTR-ViT classifier achieves a strong macro-F1 score (0.9802), this performance reflects its ability to learn discriminative in-sample distribution patterns from paired text-image samples during training. However, retrieval results—particularly on OOD distribution—reveal the inherent challenges of generalization. As shown in Figure 1 and Appendix Tables 9–11, the model achieves average R-Precision scores of 0.7418 (text-to-text), 0.1518 (image-to-image), and 0.7202 (multimodal) for OOD retrieval, highlighting a notable performance gap. This discrepancy stems from the challenge of linking novel text-image combinations unseen during training. The model is trained by associating individual text descriptions with multiple images, learning stylistic and visual patterns across modalities. In OOD scenarios, the model encounters entirely new pairs, requiring it to infer authorship from subtle cross-modal cues rather than relying on memorized associations. For instance, a vendor might reuse a new image with text that shares stylistic similarities to prior ads. The model's retrieval performance under such conditions demonstrates its ability to leverage these complementary signals. This distinction between classification (closed-set identification) and retrieval (open-set verification) is critical for real-world applications. In practice, LEAs frequently encounter OOD cases where vendors alter content across platforms or regions to evade detection. The model's design—emphasizing OOD generalization and cross-modal linking—addresses a crucial gap in AA and HT investigations, where robustness to evolving evasion tactics is crucial.

(vi). Figure 2(a)(A) highlights significant vendor overlap across the four geographic regions, raising concerns about model generalization on OOD distribution. However, similarity analysis of the datasets (Figures 2(a)(B)-(C)) and retrieval performance on shared versus unique vendors (Appendix Table 12) demonstrate that the end-to-end multimodal DeCLUTR-ViT backbone performs equally on both shared and unique vendors. This indicates strong generalization capabilities in scenarios with overlapping or region-specific vendor activity.

(vii). To improve alignment between text descriptions and escort images, we experimented with three text-alignment strategies—CLIP, CLIP with an Image-Text Matching objective, and BLIP2. While these models show improved retrieval performance over pre-trained checkpoints (Appendix Tables 9–7), they consistently underperformed compared to our end-to-end DeCLUTR-ViT baseline, even after fine-tuning for vendor identification task. This underperformance is attributed to the low semantic overlap between the noisy text (vague descriptions) and images (e.g., partial or absent faces) in escort ads (Appendix Table 8), making the alignment difficult. Given these findings, using SoTA multimodal models like LLaVA-OV 7B (Li et al., 2024a), Gemini Flash 8B (Team, 2024), Pixtral 12B (Agrawal et al., 2024), etc. presents discouragements. These models have significantly larger parameter sizes, making them impractical within our computational constraints and unfair compared to our 169M-parameter DeCLUTR-ViT backbone. Additionally, they are optimized for unrelated tasks like knowledge reasoning and Q&A, which do not align with our AA objectives. Lastly, our BLIP2 results show that projecting visual features into language space, as used by models like LLaVA, does not resolve the alignment challenges caused by low semantic overlap. Therefore, we decide not to pursue these larger general-purpose multimodal models for our AA tasks.

7 Discussion

This research introduces a novel multimodal dataset and conducts extensive benchmarking to demonstrate that multitask joint objectives and multimodal data integration enhance AA performance on both in-sample and OOD distribution datasets. By linking escort ads through these techniques, we aim to assist researchers, investigators, and LEAs study HT indicators. Due to space constraints, the main manuscript focuses on critical claims and experimental results, while additional insights and detailed analyses are provided in the Appendix.

Specifically, Appendix sections A.2 and A.3 provide detailed information on data-specific statistics, preprocessing steps, label creation, and a datasheet following (Gebru et al., 2021). Due to the sensitive nature of our research, we cannot display data samples, publicly release our models, or provide model cards. Given the explicit sexual content in images and associated privacy concerns, qualitative examples cannot be provided in the paper. However, we conduct extensive qualitative and statistical analyses into model insights and learning in Appendix A.6. Further details on architectural design, training setup, and computational considerations are presented in Appendix A.4, while comprehensive performance metrics for all baselines are available in Appendix A.5. Lastly, Appendix A.7 explores the practical application of AA tasks in building knowledge graphs to support investigative efforts.

By structuring our paper this way, we balance clarity and depth. The main manuscript provides a concise overview, while the supplementary material ensures transparency, rigor, and accessibility, enabling domain experts and practitioners to derive actionable insights from our work.

8 Conclusion

Through this research, we demonstrate the potential of MAA in addressing the complexities of vendor identification and verification within online escort markets. Using our novel MATCHED dataset, we extensively benchmark text-only, vision-only, and multimodal approaches, showcasing the advantages of CE+SupCon multitask training objectives. Our analysis reveals that this dual-objective consistently outperforms single-task approaches across in-distribution and OOD datasets, enabling LEAs to identify known vendors while linking emerging ones in new markets. Additionally, multimodal integration significantly enhances model performance by capturing complementary patterns across text and images. While text remains the dominant modality, integrating image data along text descriptions adds stylistic cues that enrich the model's capabilities. Among text, vision, and multimodal representations, text representations from the DeCLUTR-ViT backbone emerge as the most effective for retrieval tasks, achieving the best results across all modalities. While pre-trained text-image alignment strategies like CLIP and BLIP2 fail to establish meaningful cross-modal connections due to low semantic overlap and ineffective use of stylistic features, end-to-end multitask training is a more robust approach for leveraging multimodal data in AA tasks. Finally, the performance gap between pre-trained checkpoints and fine-tuned baselines highlights the importance of domain-specific adaptations and task-specific training, providing a strong foundation for future research. By addressing real-world challenges and emphasizing scalability, we aim to equip LEAs with actionable tools to uncover and disrupt trafficking networks effectively.

9 Limitations

Assumption: Similar to existing research, our research assumes that each class label corresponds to a distinct vendor during the classification task, enabling the model to leverage domain knowledge effectively. However, our qualitative analysis identifies cases where the trained classifier misclassifies ads, likely due to similarities in writing style and content, suggesting the possibility that multiple vendors might belong to the same entity. While we lack definitive ground truth to confirm this hypothesis, it represents a notable challenge in ensuring label accuracy. We recognize that improving the quality of vendor labels would likely lead to enhanced benchmark performance and more robust model evaluations.

Dataset Limitations and Generalization Challenges: Our research utilizes escort ads collected from the Backpage platform between December 2015 and April 2016, spanning seven U.S. cities in four geographical regions. While this dataset provides valuable insights into AA for sex trafficking investigations, it also presents several limitations. Notably, there is significant vendor overlap across regions (Appendix Figure 2a), and the presence of near-duplicate ads—challenging to identify and remove due to noise and variability—complicates the evaluation of the model's generalization capabilities. Although this study evaluates OOD generalization, more comprehensive assessments would benefit from data collected from multiple escort platforms and diverse geographical regions to better simulate cross-platform generalization.

While the Global Organized Crime Index highlights regions worldwide for HT activities, HT manifests in various forms, such as labor, organ, and sex trafficking, as well as forced servitude. Our research is focused specifically on addressing sex trafficking within escort advertisements. While expanding data collection beyond US-based ads to encompass a wider range of geographical regions and demographics is crucial, identifying escort platforms directly linked to HT operations is a significant challenge, as such connections often require verification through law enforcement investigations or court rulings. To date, beyond Backpage and Craigslist, not many escort platforms have been explicitly linked to HT activities.

Finally, data collection for this study was conducted under strict ethical oversight. Approval from the ethics committee was obtained, largely due to the relatively dated nature of the dataset, which reduces privacy risks. It is suspected that many victims and perpetrators have since moved from these platforms or changed their personal information to avoid identification. Furthermore, our research is not an active investigation but rather an effort to develop tools that may assist LEAs in identifying and linking escort ads to disrupt trafficking networks. In future work, we aim to explore methods for ethically collecting data from additional escort platforms—particularly those with verifiable connections to HT operations—to enhance generalizability across diverse demographics and regions. This expansion will be crucial for developing more robust, globally representative AA models for HT investigations. That said, our current dataset remains a valuable benchmark for future research, offering critical insights into how traffickers facilitated HT on Backpage escort platforms during 2015-2016. It will serve as a reference point for analyzing how criminal behavior and evasion tactics have evolved over time and across platforms, helping researchers and LEAs track shifts in trafficking strategies and adapt investigative approaches accordingly.

Selective Feature Extraction and Fine-Tuning: In this work, we extract text and vision representations exclusively from the final layers of our models, which may not fully capture nuanced features learned at earlier layers. Representations extracted from intermediate layers could yield different or potentially better outcomes. Additionally, while fine-tuning pre-trained text-image alignment models, we fine-tune all layers uniformly, which may not be optimal. Techniques like Centered Kernel Alignment (CKA) (Kornblith et al., 2019) can provide insights into which layers learn the most relevant features, enabling more informed decisions about representation extraction and selective layer freezing during fine-tuning. Addressing these concerns is currently beyond the scope of this research, but we plan to explore these aspects in future work.

Computational Constraints: While our research employs relatively large model architectures and advanced training strategies, it is limited by the computing resources available to us. Larger model architectures could potentially enhance performance across classification and retrieval tasks. However, when applied to text-image alignment tasks, the computational demands of scaling these models exceeded our resource capacity. As a result, we opted for smaller, more efficient architectures that fit within our computational constraints, ensuring a fair and balanced comparison across baselines. Similarly, our research relies heavily on contrastive learning objectives, and prior studies (Gao et al., 2021; Vaessen and van Leeuwen, 2024) highlight the benefits of larger batch sizes for such tasks. However, to maintain consistency and fairness among baselines, we limited our batch size to 32, as larger sizes led to memory errors, particularly with text-image alignment models. This computational limitation also influenced our decision to forego fine-tuning pre-trained CLIP and BLIP2 checkpoints, as the memory requirements for fine-tuning BLIP2 architecture caused GPU crashes. These decisions reflect deliberate trade-offs made to ensure the reproducibility and fairness of our experimental comparisons while working within resource limitations.

Explainability: Although this research does not explicitly address the explainability or interpretability of our models, we recognize their critical role in fostering trust among researchers, investigators, and law enforcement agencies. Previous studies (Saxena et al., 2023a) have explored explainability in AA through local feature attribution techniques applied to text ads. However, while numerous frameworks exist for explainability in unimodal data (Ribeiro et al., 2016; Lundberg and Lee, 2017; Kokhlikyan et al., 2020), these methods cannot be directly extended to the multimodal AA context. Additionally, research highlights the limitations of existing explainability techniques, including their susceptibility to adversarial attacks, network sparsity, and inconsistencies in results (Das and Rad, 2020; Krishna et al., 2022; Saxena et al., 2023b). We aim to address these challenges in future work by developing a robust explainability framework tailored specifically for multimodal AA scenarios. Such a framework will help uncover the contributions of textual and visual features in decision-making processes, ensuring transparency and reliability in the application of multimodal AA models.

Generative Models: Vendors could potentially exploit advancements in generative technologies, such as Large Language Models (LLMs) like ChatGPT and vision-based generative models, to craft text ads with varying linguistic styles or manipulate images to inject obscuring identifiable stylistic cues, making AA more challenging. While such scenarios remain speculative—there is currently no concrete evidence that HT vendors are actively using LLMs or generative models to produce ads—the possibility poses significant challenges to AA systems. Detecting artificially generated content would require access to ground-truth information, which is difficult to obtain. Even if future datasets include ads suspected of being generated by LLMs, proving their artificial origins would remain a major challenge.

Although publicly available LLMs often restrict content generation for illegal purposes, open-source models could be fine-tuned or customized by vendors to evade detection by mimicking diverse stylistic patterns. These evolving capabilities could undermine the effectiveness of text- and vision-based AA systems, which depend on identifying unique stylometric and visual features. To address these potential threats, our future work plans to adapt our AA systems by recollecting and analyzing updated datasets, enabling them to differentiate between human-generated and machine-generated content. This will help ensure our models remain robust against emerging tactics that leverage generative technologies.

10 Ethical Considerations

10.1 Data Protocols

We collect our dataset from the Backpage Escort Markets spanning seven U.S. cities, posted between December 2015 and April 2016. Following ethical guidelines outlined by Krotov et al. (2020), which presents a framework of seven principles for responsible web scraping, we ensured our approach complied with these standards. The Backpage website's use policy does not explicitly prohibit data scraping.

10.2 Privacy Considerations and Potential Risks

In undertaking this research, we recognize the significant privacy concerns associated with using data from escort advertisements, particularly given that individuals within these ads may be at risk. However, the prevalence of human trafficking, a grave societal issue that affects countless lives, drives our commitment to contribute positively to anti-trafficking efforts. We believe our intentions align with the broader ethical imperative to support the fight against exploitation and to aid LEA in identifying and disrupting trafficking networks.

To address privacy concerns, we have extensively tried to mask personal identifiers within the dataset. Following methods from Saxena et al. (2023a), we mask phone numbers, email addresses, post IDs, dates, and links in text data, transforming them into generalized formats such as "<EMAILID-23>" or "<LINK>," which minimizes the risk of

reverse engineering and personal identification (please refer appendix section A.2 for more details). At the same time, we explored various entity recognition tools to mask names (Li et al., 2022a; Liu et al., 2023) and locations, the inherent noise in the data led to inaccuracies, with some false positives in entity predictions. Since research indicates that individuals in these ads often use pseudonyms (Carter et al., 2021; Lugo-Graulich) and Backpage ads are no longer publicly accessible after the 2016 seizure, we find it unlikely that masked text data could be misused for individual identification.
Privacy risks are more challenging to mitigate for the image data, as AA relies on preserving stylistic cues. Although we initially considered blurring faces to protect identities, we ultimately decided against it to avoid introducing biases that could compromise the authenticity of stylometric patterns. This decision was made after careful consideration of the potential impact on the accuracy and integrity of the AA task. Many ads already feature images with blurred or cropped faces, which suggests an attempt by individuals to maintain anonymity. For similar reasons, we also opted not to use other image augmentations, such as flipping or rotating, as these transformations could alter stylistic features tied to individual vendors, thus potentially impacting the accuracy and integrity of the AA task.

Our efforts to balance privacy with societal benefit align with the principles outlined in Article 6 of the General Data Protection Regulation (GDPR), the lawfulness of processing. By minimizing identifiable information and rigorously managing data access, we strive to uphold this balance.

To further safeguard against misuse, we have established strict access controls for the MATCHED dataset. Access will be limited to vetted researchers and organizations with legitimate research goals, particularly those focused on anti-trafficking and public welfare. Each access request will undergo a thorough review by an ethics review board, assessing the legitimacy of the research goals and the adequacy of the applicant's security measures. This process ensures that only those committed to ethical and secure usage standards gain access. Applicants must also sign non-disclosure and data protection agreements legally binding them to these standards. Any violation of these guidelines will result in legal consequences. Only metadata on the Dataverse platform will offer a high-level overview without compromising sensitive information.

Note: Our research has undergone ethical scrutiny within our institution, and we have received internal approval to proceed with the project. The ethical review details and additional documentation will be provided in the camera-ready version of this paper, demonstrating our commitment to transparency and responsibility in our efforts. We are guided by the principle that our work should ultimately serve to protect and support vulnerable individuals, advancing a cause deeply rooted in societal benefit.

10.3 Legal Impact

We acknowledge that the specific impact of our research on law enforcement processes is difficult to predict. Our primary goal is to support LEAs in better understanding vendor connections within online escort markets, offering a tool to assist in their investigative efforts. We strongly recommend that LEAs and researchers treat our analysis as an investigative aid rather than direct evidence for criminal prosecution. Our findings should be supplementary tools to guide investigations, not standalone proof of criminal activity.

10.4 Environmental Impact

Our experiments are conducted on a private infrastructure equipped with an NVIDIA H100 80GB GPU (TDP of 350W) and a carbon efficiency of 0.475 $kgCO_2eq/kWh$. Establishing all baselines required a cumulative training time of 45.79 hours. Using the Machine Learning Impact calculator from Lacoste et al. (2019), we estimate the total emissions for these experiments to be approximately 16.625 $kgCO_2eq$.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? learning to contrast for authorship attribution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1142–1157, Online only. Association for Computational Linguistics.

Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel

Werghi. 2023. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446.

Hamidreza Alvari, Paulo Shakarian, and J. E. Kelly Snyder. 2017. Semi-supervised learning for detecting human trafficking. *Preprint*, arXiv:1705.10786.

Hamidreza Alvari, Paulo Shakarian, and J.E. Kelly Snyder. 2016. A non-parametric learning approach to identify online human trafficking. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 133–138.

Hassan Marzoughi Ardakani. 2020. *Identifying Human Trafficking Networks in Louisiana by Using Authorship Attribution and Network Modeling*. Louisiana State University and Agricultural & Mechanical College.

Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings 2*, pages 37–47. Springer.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. *Preprint*, arXiv:1710.08092.

Pelham Carter, Matt Gee, Hollie McIlhone, Harkeeret Lally, and Robert Lawson. 2021. Comparing manual and computational approaches to theme identification in online forums: A case study of a sex work special interest community. *Methods in Psychology*, 5:100065.

Nathanael Chambers, Timothy Forman, Catherine Griswold, Kevin Lu, Yogaish Khastgir, and Stephen Steckler. 2019. Character-based models for adversarial phone extraction: Preventing human sex trafficking. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Leshem Choshen, Dan Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. 2019. The language of legal and illegal activity on the Darknet. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4271–4279, Florence, Italy. Association for Computational Linguistics.

Marco Cotogni, Marco Arazzi, and Claudio Cusano. 2024. Photostyle60: A photographic style dataset for photo authorship attribution and photographic style transfer. *IEEE Transactions on Multimedia*, pages 1–12.

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *Preprint*, arXiv:2006.11371.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. 2021. Training vision transformers for image retrieval. *Preprint*, arXiv:2102.05644.

EUROPOL. 2020. The challenges of countering human trafficking in the digital era.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Andrian Firmansyah, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. 2023. Comparison of face recognition accuracy of arcface, facenet and facenet512 models on deepface framework. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 535–539.

Eilika Fobbe. 2021. Text-linguistic analysis in forensic authorship attribution.

I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua. 2018. Image and encoded text fusion for multi-modal classification. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7.

Chenquan Gan, Xiang Fu, Qingdong Feng, Qingyi Zhu, Yang Cao, and Ye Zhu. 2024. A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications*, 242:122731.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Socratis Gkelios, Yiannis Boutalis, and Savvas A. Chatzichristofis. 2021. Investigating the vision transformer model for image retrieval tasks. *Preprint*, arXiv:2101.03771.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. 2020. Deepstyle: User style embedding for authorship attribution of short texts. In *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*, pages 221–229. Springer.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely connected convolutional networks. *Preprint*, arXiv:1608.06993.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Michelle Ibanez and Rich Gazan. 2016. Virtual indicators of sex trafficking to identify potential victims in online advertisements. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 818–824.

Michelle Ibanez and Daniel D. Suthers. 2014. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In *2014 47th Hawaii International Conference on System Sciences*, pages 1556–1565.

ILO. 2012. Ilo global estimate of forced labour.

Plotly Technologies Inc. 2015. Collaborative data science.

Farkhund Iqbal, Rachid Hadjidj, Benjamin C.M. Fung, and Mourad Debbabi. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:S42–S51. The Proceedings of the Eighth Annual DFRWS Conference.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Patrick Juola and R Harald Baayen. 2005. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl):59–67.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *Preprint*, arXiv:2009.07896.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *Preprint*, arXiv:2202.01602.

Vlad Krotov, Leigh Johnson, and Leiser Silva. 2020. Tutorial: Legality and ethics of web scraping.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Yifei Li, Pratheeksha Nair, Kellin Pelrine, and Reihaneh Rabbany. 2022a. Extracting person names from user generated text: Named-entity recognition for combating human trafficking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2854–2868, Dublin, Ireland. Association for Computational Linguistics.

Yifei Li, Pratheeksha Nair, Kellin Pelrine, and Reihaneh Rabbany. 2022b. Extracting person names from user generated text: Named-entity recognition for combating human trafficking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2854–2868, Dublin, Ireland. Association for Computational Linguistics.

Yushi Li, Xin Zheng, Ming Zhu, Jie Mei, Ziwen Chen, and Yunfei Tao. 2024b. Compact bilinear pooling and multi-loss network for social media multimodal classification. *Signal, Image and Video Processing*, 18(11):8403–8412.

Javin Liu, Hao Yu, Vidya Sujaya, Pratheeksha Nair, Kellin Pelrine, and Reihaneh Rabbany. 2023. SWEET - weakly supervised person name extraction for fighting human trafficking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3355–3367, Singapore. Association for Computational Linguistics.

Kristina Lugo-Graulich. Indicators of sex trafficking in online escort ads. https://www.ojp.gov/pdffiles1/nij/grants/305453.pdf.

Kristina Lugo-Graulich and Leah F. Meyer. 2021. Law enforcement guide on indicators of sex trafficking in online escort ads. *Justice Research and Statistics Association*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. Veridark: A large-scale benchmark for authorship verification on the dark web. *Preprint*, arXiv:2207.03477.

Chirag Nagpal, Kyle Miller, Benedikt Boecking, and Artur Dubrawski. 2017. An entity resolution approach to isolate instances of human trafficking online. *Preprint*, arXiv:1509.06659.

Chirag Nagpal, Kyle Miller, Benedikt Boecking, and Artur W. Dubrawski. 2015. An entity resolution approach to isolate instances of human trafficking online. In *NUT@EMNLP*.

Pratheeksha Nair, Yifei Li, Catalina Vajiac, Andreas Olligschlaeger, Meng-Chieh Lee, Namyong Park, Duen Horng Chau, Christos Faloutsos, and Reihaneh Rabbany. 2022. Vispad: Visualization and pattern discovery for fighting human trafficking. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 273–277, New York, NY, USA. Association for Computing Machinery.

Smita Nirkhi and Dr. R.V. Dharaskar. 2013. Comparative study of authorship identification techniques for cyber forensics analysis. *International Journal of Advanced Computer Science and Applications*, 4(5).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

POLARIS. 2018. Human trafficking statistics.

POLARIS. 2020. Polaris analysis of 2020 data from the national human trafficking hotline.

Rebecca S. Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. 2017. Backpage and bitcoin: Uncovering human traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1595–1604, New York, NY, USA. Association for Computing Machinery.

Reihaneh Rabbany, David Bayani, and Artur Dubrawski. 2018. Active search of connections for case building and combating human trafficking. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2120–2129.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Hoshiladevi Ramnial, Shireen Panchoo, and Sameerchand Pudaruth. 2016. Authorship attribution using stylometry and machine learning techniques. In *Intelligent Systems Technologies and Applications: Volume 1*, pages 113–125. Springer.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Gaurav Sahu and Olga Vechtomova. 2021. Adaptive fusion techniques for multimodal data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3156–3166, Online. Association for Computational Linguistics.

Vageesh Saxena, Benjamin Ashpole, Gijs van Dijck, and Gerasimos Spanakis. 2023a. IDTraffickers: An authorship attribution dataset to link and connect potential human-trafficking operations on text escort advertisements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8444–8464, Singapore. Association for Computational Linguistics.

Vageesh Saxena, Nils Rethmeier, Gijs van Dijck, and Gerasimos Spanakis. 2023b. VendorLink: An NLP approach for identifying & linking vendor migrants & potential aliases on Darknet markets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8619–8639, Toronto, Canada. Association for Computational Linguistics.

Sefik Ilkin Serengil and Alper Ozpinar. 2023. An evaluation of sql and nosql databases for facial recognition pipelines.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

William C. Sleeman, Rishabh Kapoor, and Preetam Ghosh. 2022. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Comput. Surv.*, 55(7).

Jacob Striebel, Abishek Edikala, Ethan Irby, Alex Rosenfeld, J. Gage, Daniel Dakota, and Sandra Kübler. 2024. Scaling up authorship attribution. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 295–302, Mexico City, Mexico. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *Preprint*, arXiv:1512.00567.

Kovács Tamás, Atzenhofer-Baumgartner, Florian, Aoun, Sandy, Nicolaou, Anguelos, Luger, Daniel, Decker, Franziska, Lamminger, Florian, Vogeler, and Georg. 2022. langdetect (revision 0215f72).

Mingxing Tan and Quoc V. Le. 2021. Efficientnetv2: Smaller models and faster training. *Preprint*, arXiv:2104.00298.

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

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1547–1556, Vancouver, Canada. Association for Computational Linguistics.

UNDOC. 2020. Global report on trafficking in persons.

Nik Vaessen and David A. van Leeuwen. 2024. The effect of batch size on contrastive self-supervised speech representation learning. *Preprint*, arXiv:2402.13723.

Catalina Vajiac, Duen Horng Chau, Andreas Olligschlaeger, Rebecca Mackenzie, Pratheeksha Nair, Meng-Chieh Lee, Yifei Li, Namyong Park, Reihaneh Rabbany, and Christos Faloutsos. 2023. Trafficvis: Visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):53–62.

Danae Sánchez Villegas, Daniel Preoţiuc-Pietro, and Nikolaos Aletras. 2024. Improving multimodal classification of social media posts by leveraging image-text auxiliary tasks. *Preprint*, arXiv:2309.07794.

Xiangwen Wang, Peng Peng, Chun Wang, and Gang Wang. 2018. You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 431–442, New York, NY, USA. Association for Computing Machinery.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Preprint*, arXiv:2301.00808.

Zhanhong Ye, Changle Zhong, Haoliang Qi, and Yong Han. 2023. Supervised contrastive learning for multi-author writing style analysis. In *CLEF (Working Notes)*, pages 2817–2822.

Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, WWW '19, page 3448–3454, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Responsible NLP Checklist

A.1.1 For every submission

Did you describe the limitations of your work? Yes, the limitations of our work are extensively described in Section 9.

Did you discuss any potential risks of your work? Yes, the potential privacy risks associated with our work are described in Section 10.2.

A.1.2 Did you use or create scientific artifacts?

Did you discuss the license or terms for use and / or distribution of any artifacts? The dataset will be released under Custom License Terms with restrictive access. Extensive details about the terms of use and/or distribution are mentioned in Appendix Section A.2. These terms will also be made available on the Dataverse portal once the dataset's meta-data is released publicly.

Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify the intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? Given the sensitivity of our dataset, access will be provided under restricted conditions to ensure ethical use. Interested parties must sign a Non-Disclosure Agreement (NDA) and Data Transfer Agreement (DTA) with our institution and the ethics committee. To minimize risks to individuals represented in the dataset, we have implemented strong anonymization techniques to remove private and personally identifiable information. We strictly prohibit using this dataset for any commercial or unethical purposes beyond the intended scope of our research. Violations of these guidelines will be subject to legal repercussions as outlined by the institution's policies and the ethics committee.

Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? Yes, we thoroughly detail the data collection and preprocessing steps, including the measures taken to identify and remove any private or personally identifiable information. Specifically, we anonymize sensitive content such as names, phone numbers, email addresses, advertisement IDs, dates, and ages of individuals to ensure privacy. These efforts and additional discussions are comprehensively reported in Appendix Section A.2-A.3.

Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? Yes, the details about the coverage of domains, languages, and geographical groups are presented in Section 3 and Appendix Sections A.2 - A.3.

Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Yes, these details are mentioned in Section 3 and Appendix Section A.4.

A.1.3 Did you run computational experiments?

Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? Yes, these details are attached in Appendix Table 4.

Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? Yes, these details are attached in Appendix Section A.4.

Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? Details about the effects of random initialization for our best-performing model, the end-to-end multimodal DeCLUTR-ViT baseline, are attached in Appendix Section A.4.

If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? All relevant details are described in Appendix Section A.4.

A.1.4 Did you use human annotators (e.g., crowdworkers) or research with human participants?

Do you use any human annotators? No.

Did you discuss whether and how consent was obtained from people whose data you're using/curating? Getting consent for our data is challenging due to the nature and timeline of our dataset. We have extensively described this problem in our Appendix Section A.3.

Was the data collection protocol approved (or determined exempt) by an ethics review board? Yes, the approval was granted by our institutional's ethics board. We plan to attach the approval in our camera-ready version.

A.1.5 Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?

While our research methodology, experiments, and results were developed independently without AI assistants, we utilized ChatGPT and Grammarly to improve our paper's readability, clarity, and flow. Importantly, we wrote the initial drafts, including all content. ChatGPT was used only to paraphrase sections for clarity and improve grammar. Additionally, for coding purposes, we employed ChatGPT solely to generate in-line comments for better code readability. Specifically, we passed handwritten functions and classes to ChatGPT and requested it to generate comments without altering any logic or structure in the code.

This information is transparently described here and is not included in the main manuscript because the AI assistance was limited to minor paraphrasing, grammar improvement, and in-line code comments, with no role in generating methodology, experiments, or results.

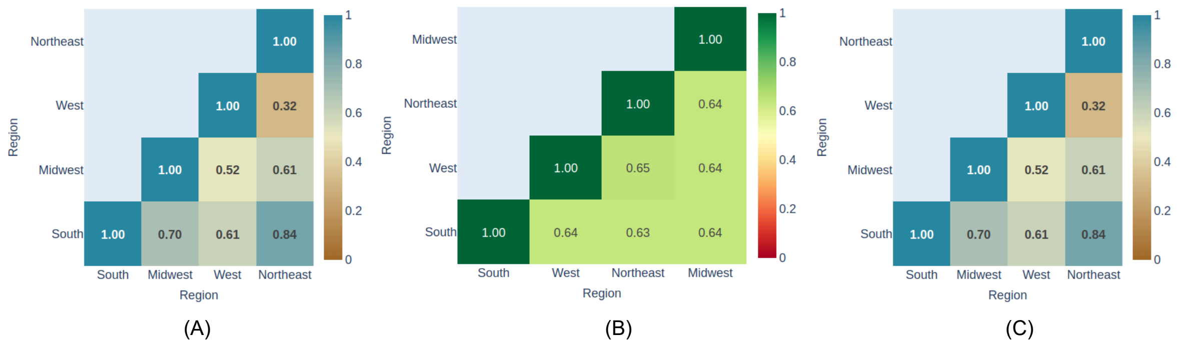
A.2 Dataset

(i) Data Analysis: Figure 2a(A) illustrates the % of shared vendors across different datasets. As can be observed, many vendors post ads across multiple geographical regions, which aligns with existing findings that the Backpage escort marketplace was often flagged for HT activities, with vendors frequently advertising their services across various regions (Lugo-Graulich and Meyer, 2021). This cross-regional vendor activity also highlights a limitation in our OOD generalization experiments, which are designed to test the ability of our models to make predictions on data distribution that is different from the data it was trained on. These experiments may not fully capture real-world conditions. To properly assess true OOD generalization, future work would need to collect ads from an entirely separate escort platform to evaluate our models' adaptability to a new distribution of ads—an approach that lies outside the scope of this research.

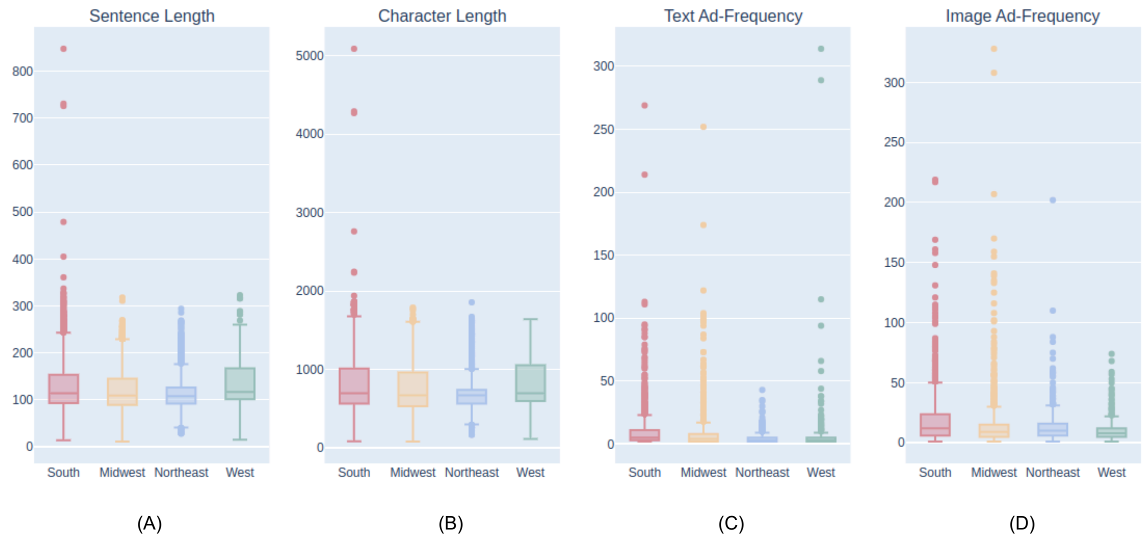
Figure 2a(B) examines the average text-to-text similarity between ads from different datasets. Using a pre-trained DeCLUTR-small model, we compute the similarity by generating sentence embeddings for each ad and calculating the cosine similarity between pairs from different datasets. Given the high level of vendor overlap across regions, the text content is expected to exhibit considerable similarity. Similarly, figure 2a(C) shows the average image-to-image cosine similarity across ads from different datasets, calculated using representations from a pre-trained ViT-base-patch16 model. Compared to the relatively high text similarity, the image similarity is lower. This suggests that, while vendors often maintain consistent writing styles across regions, they tend to vary the images posted, potentially to depict different escorts.

Figure 2b(A) and (B) illustrate the sentence and character length distributions of text ads within our datasets. Sentence length is measured by counting the total number of tokens generated by the pre-trained DeCLUTR-small checkpoint after tokenization, while character length is the count of characters in each text ad. As shown, most text ads have a sentence length of fewer than 512 tokens. Therefore, we truncate all text ads to a maximum length of 512 tokens, also the maximum sequence length allowed by most transformers-based models. Figure 2b(C) depicts the text-ad frequency, i.e., the number of text ads posted per vendor. As evident, most vendors post between 1 and 20 text ads. Unlike other authorship attribution (AA) approaches applied to criminal markets, which require a minimum of 5 (Saxena et al., 2023a) or 20 (Saxena et al., 2023b) ads for effective AA implementation, our research explores the applicability of AA techniques for vendors with as few as two ads. This distribution of ad frequency highlights a class imbalance in our dataset, prompting us to prioritize Macro-F1 performance to ensure equal weighting across all classes in our classification tasks. Similarly, 2b(D) depicts the image-ad frequency or the number of image ads posted per vendor. As evident, most vendors post between 5 and 24 image ads. A detailed analysis of the frequency of text, image, and multimodal ads per vendor is attached in Figure 2c [1]. Finally, our language analysis using the LangDetect model (Tamás et al., 2022) reveals that the vast majority of text ads are in English: 99.65% in the South dataset, 99.98% in the Midwest dataset, 99.88% in the West dataset, and

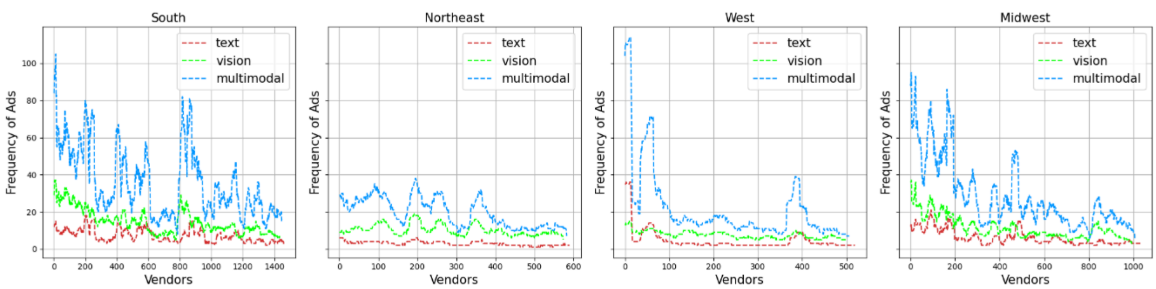
[1] Note that these line plots are plotted with a smoothing applied to window size of 30 for better readability.



(a) Figure (A) shows the % of vendors shared between different datasets. Figures (B) and (C) show the average text-text and image-image cosine similarity between datasets computed on the ad representations from the pre-trained available checkpoints of DeCLUTR-small and ViT-base-patch16 backbones, respectively.



(b) Figures (A) and (B) showcase distributions of sentence and character length for text advertisements in the datasets. Figures (C) and (D) show a distribution of text-ad and image-ad frequency for each dataset, i.e., the number of text and image ads per vendor.



(c) Frequency of text, image, and multimodal ads in South, Northeast, West, and Midwest region datasets.

Figure 2

99.85% in the Northeast dataset.

(ii) Data Pre-Processing: As described in Section 3, our dataset is sourced from Backpage escort ads posted across seven US cities between December 2015 and April 2016. We scrape titles, descriptions, and images for each ad. The text sequence for each entry is created by combining the title and description separated with a "[SEP]" token. Since ads may contain multiple images, we duplicate the text sequence for each associated image to prepare the dataset for multimodal training.

To establish ground truth, we follow Saxena et al. (2023a) and utilize tools from Nagpal et al. (2017); Chambers et al. (2019) and Hagberg et al. (2008) to extract phone numbers and form vendor communities, aka vendor labels. Consistent with Saxena et al. (2023a), we mask most personal information, including phone numbers, escort ages, measurements, ad IDs, and posting dates. Despite attempts to mask all identifiable information, existing Named Entity Recognizers (NER) (Li et al., 2022a; Liu et al., 2023) struggle to extract escort names from the ads reliably. However, since escorts generally use pseudonyms in these ads (Carter et al., 2021; Lugo-Graulich), the potential for misuse of personal data is already minimal.

For image anonymization, we initially considered blurring faces to protect escort identities. However, manual inspection revealed that many images with blurred or cropped faces are anonymously posted. To preserve these stylistic elements, we opted not to add artificial blurring, which could introduce visual biases. Similarly, we avoided other image augmentation techniques, as transformations such as flipping or rotating could alter stylistic cues linked to specific vendors. Some ads naturally feature mirrored or rotated images, which are retained to prevent misattribution. To further analyze model behavior, we categorized the image dataset into "Face" and "No Face" subsets for each of the four regions—South, West, Midwest, and Northeast—using a pre-trained FaceNet model (Firmansyah et al., 2023). FaceNet detects and generates bounding boxes around faces in images, which are then assigned to that region's "Face" dataset.

(iii) Language Distribution: Our analysis reveals that approximately 99.84% of our dataset's vocabulary is English. Given that only a small fraction of our dataset's vocabulary lies outside English, we anticipate that employing multilingual models would have a negligible effect on model performance. These statistics are obtained using the LangDetect (Tamás et al., 2022) python model.

A.3 Datasheet

Following Gebru et al. (2021), we provide the datasheet for our MATCHED dataset below:

A.3.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. The MATCHED dataset was created to support LEAs, investigators, and researchers in identifying vendor connections within online escort ads. Traditional methods often rely on explicit personal identifiers such as phone numbers and email addresses. However, existing research shows that only a small fraction of ads include this information, limiting the effectiveness of these approaches. In response, Saxena et al. (2023a) introduced AA methods to connect escort vendors through stylistic similarity in text, providing an alternative way to link ads without direct identifiers. Our dataset fills a critical gap by incorporating textual descriptions and images associated with escort ads, enabling researchers to move beyond text-only analysis. This multimodal dataset allows for the exploration of multimodal training strategies that integrate both text and images, aimed at improving the robustness and generalizability of AA in the context of HT detection.

A.3.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. The instances in the MATCHED dataset represent individual ads from online escort services. Each ad instance comprises two main components: (1) a raw text sequence created by merging the title and description of the escort ad with a [SEP] token separating them, and (2) one or more images associated with the ad, typically depicting the escort being advertised. Each ad instance is then connected to a vendor ID, a unique identifier representing the individual or organization responsible for posting the ad. This vendor ID enables the grouping of ads by their source, supporting the AA task and facilitating the connection of ads linked to the same vendor.

How many instances are there in total (of each type, if appropriate)? What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? Is there a label or target associated with each instance? The MATCHED dataset consists of 28,513 ad instances, including 27,619 unique text descriptions and 55,115 escort images linked to 3549 unique vendors. Each instance in the dataset comprises "raw" data from unprocessed text and images. The dataset is provided as a pandas DataFrame in a .csv format, with three main columns: "TEXT," "IMAGES," and "VENDOR." The "TEXT" column contains the input text sequence in string format, created by merging the title and description of the ad. The "IMAGES" column holds the local file path for each image associated with the ad. The "VENDOR" column includes the class labels, represented as integer IDs corresponding to specific vendors. Further details on dataset composition and split are outlined in Table 1.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). The MATCHED dataset represents a sample of the broader Backpage escort market data, with ads collected from seven cities across five U.S. states. To ensure a reliable ground truth for AA tasks, we filtered the ads to include only those with phone numbers (used to establish vendor labels) and at least one image. This filtering process resulted in a final set of 28,513 ads. Consequently, while the dataset does not fully represent the entire Backpage escort market, it focuses on instances where both text and image modalities are available, which is essential for exploring MAA.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. No

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Relationships between instances in our dataset are established by extracting and grouping phone numbers found within ads. Using the TJBatchExtractor (Nagpal et al., 2017) and CNN-LSTM-CRF classifier (Chambers et al., 2019), we identify phone numbers that act as identifiers for vendors. These identifiers are then used to construct vendor communities via NetworkX (Hagberg et al., 2008), where each community corresponds to a unique vendor label. This approach links ads to individual or organizational entities (vendors) by grouping ads associated with the same phone number, creating a structured relationship among instances in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please describe these splits, explaining the rationale behind them. We split our dataset into training, validation, and test sets using a 0.75:0.05:0.20 split ratio. This allocation is intended to provide a substantial training set (75%) for effective model learning, a validation set (5%) for tuning model hyperparameters and avoiding overfitting, and a test set (20%) to assess model generalization and in-distribution performance.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. As indicated by Saxena et al. (2023a), a considerable amount of noise is present in the Backpage escort ads. In the text data, vendors often add extra punctuation, emojis, irregular white spaces, and random characters, likely as a tactic to circumvent automated detection systems. These irregularities can impact text processing and add complexity to data-cleaning efforts. Our manual inspection of the image data also reveals visual noise, including intentionally blurred areas and white noise, which further complicates the analysis. However, quantifying the extent of this noise in images remains challenging. Despite these issues, the noise and irregularities reflect the original conditions in which the data was originally posted, providing a realistic foundation for developing robust AA models that can handle similar situations in real-world applications.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restric-

tions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate No. The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description. Building on the guidelines by Saxena et al. (2023a), we also include measures to minimize privacy risks and mitigate data misuse. We anonymize sensitive details in text by replacing digits with the letter "N" and substituting email addresses with $< EMAIL_ID >$, post IDs with $POST_ID : NNNNN$, dates with $< DATES >$, and links with $< LINK >$. Attempts were made to mask escort names and locations using NER models (Li et al., 2022a; Liu et al., 2023), but noise in the data led to inaccurate predictions. Nevertheless, as previous studies suggest that escorts often use pseudonyms (Carter et al., 2021; Lugo-Graulich), the potential for misuse of personal details in text ads is low.

That said, we recognize that identities could still be inferred from images. Initially, we considered blurring faces to enhance anonymity. However, manual inspection showed that many images already had faces blurred or cropped by the posters. To retain these natural stylistic cues, we decided against additional blurring, as it could interfere with AA tasks and introduce unintended biases in the visual data. Additionally, a sanity check using the FairFace (Karkkainen and Joo, 2021) and DeepFace (Serengil and Ozpinar, 2023) models demonstrated that these tools, when applied to our noisy dataset, were unable to extract any ethnicity or age-related information from the dataset's images.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. Yes, the dataset comprises text and (semi-nude) images from escort advertisements that contain sexual descriptions.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Our dataset does not explicitly identify subpopulations by age, as all age information has been masked in the text ads. However, some ads include descriptions of the escorts' ethnicities, which remain unmasked to preserve the original stylometric features for AA tasks. Additionally, most ads in our dataset correspond to women-based escort services. It is important to note that while we have not provided age or ethnicity labels, malicious users could potentially infer such details by applying automated systems to the images. This potential for inference underscores the importance of responsible dataset usage and adherence to ethical guidelines to prevent misuse.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. While we cannot entirely rule out the possibility of identifying individuals through our dataset, we have followed extensive privacy measures pointed out by (Saxena et al., 2023a) to minimize this risk. In the text ads, we have masked private identifiers, such as phone numbers, email addresses, and other personal information, to protect the identities of individuals. The dataset comprises ads from the Backpage escort market collected between December 2015 and April 2016, a period for which there are no longer public records since the website was seized. However, there remains a risk associated with the images in our dataset, as they may still allow for indirect identification of individuals.

To mitigate this risk, we will restrict access to the MATCHED dataset, allowing only approved researchers or agencies with legitimate research objectives—specifically those focused on combating HT or conducting academic (non-commercial) research related to AA. Access will be granted through a data portal, Dataverse, subject to approval from our ethics review board, which ensures that the dataset is used solely for its intended purposes. Unauthorized use of the dataset, particularly for purposes beyond AA or HT research, is strictly prohibited under our ethical guidelines and will have legal repercussions.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union

memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. Despite our masking efforts, our dataset still contains sensitive information. While we have successfully masked certain private identifiers, such as phone numbers and email addresses, challenges remain in masking other potentially sensitive details, including escort names, ad locations, ethnicities, and sexual orientations. These details are present in the ads' text descriptions and could be extracted from the images using automated systems. The inherent noise in the data further complicates the accurate masking of these elements. As a result, while we have taken significant precautions, there remains a possibility that sensitive information could be inferred from the dataset.

A.3.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The data for each instance was acquired from raw text and images associated with escort ads posted on the Backpage market. Following Saxena et al. (2023a), we utilized the TJ-BatchExtractor (Nagpal et al., 2017) and a CNN-LSTM-CRF classifier (Chambers et al., 2019) to extract phone numbers from these ads, which serve as identifiers to group ads into vendor communities. NetworkX (Hagberg et al., 2008) was subsequently used to build these communities, assigning a unique label ID to each vendor.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? The raw data is provided to us from Bashpole Software, Inc..

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. The MATCHED dataset contains ads from seven US cities and is scraped from online posted ads between December 2015 and April 2016 on the Backpage Escort Markets. The raw data is provided to us from Bashpole Software, Inc..

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. The individuals in our ads were not notified about the data collection. Given that the ads were posted on Backpage between December 2015 and April 2016, obtaining consent from these individuals is infeasible. Since the Backpage escort market was seized and shut down, reconnecting with these individuals—many of whom used pseudonyms and transient contact information like phone numbers or email addresses—is impractical after such a long period. Additionally, as Backpage no longer exists as a platform, contacting the original poster would be challenging and unlikely to yield responses.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. No.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). NA

A.3.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. To prioritize privacy and reduce the risk of misuse, we implemented extensive preprocessing and cleaning procedures to protect sensitive information within the text descriptions in our dataset. This involved masking identifiable elements, including phone numbers, email addresses, age details, post IDs, dates, and links mentioned in the ads. The images are not processed or cleaned to maintain their original stylometric cues. Finally, since the goal of our research is MAA, we removed all instances that did not contain phone numbers or images.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. No.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. No.

A.3.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. This research introduces MATCHED, a novel dataset comprising text descriptions and images from Backpage escort markets, specifically developed for MAA. While MATCHED has not been utilized in any previous studies, several works have reportedly used text descriptions or images from Backpage escort marketplaces for similar analyses (Alvari et al., 2016; Portnoff et al., 2017; Alvari et al., 2017; Saxena et al., 2023a), etc. However, due to the unavailability of these datasets, we could not verify whether any ads overlap with those in MATCHED.

What (other) tasks could the dataset be used for? The MATCHED dataset is strictly intended for use in AA tasks related to combating human trafficking or conducting academic research within ethical boundaries. Our ethics review board has implemented strict guidelines prohibiting using this dataset beyond these purposes. Consequently, we discourage any other applications, as they could risk potential misuse or ethical concerns that are not aligned with the dataset's purpose and ethical considerations.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? Although we have taken extensive precautions to mask sensitive information, our dataset still includes details like escort pseudonyms, posted locations, ethnicity, and sexual preferences, which could be potentially sensitive. While these details are unlikely to be used to harm individuals directly, we strongly caution against any unethical applications, particularly those that could lead to re-identifying individuals or otherwise compromising their privacy. This includes any research or commercial use aimed at profiling, targeting, or stereotyping. To mitigate these risks, we advise dataset consumers to strictly adhere to ethical guidelines, focusing solely on the dataset's intended purpose of combating human trafficking through academic research. Additionally, we encourage the users to implement further anonymization techniques, especially if using images, and avoid practices that could unintentionally expose or unfairly represent individuals or groups in the dataset.

A.3.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. Yes, we plan to make our dataset accessible to third parties via the Dataverse data repository. To mitigate risks of illegal or unethical use, access will be granted under specific conditions, including mandatory signing of a non-disclosure agreement (NDA) and data protection agreements. Each application for access will be evaluated by our ethics committee to ensure alignment with the dataset's intended purpose. These agreements will prohibit data redistribution and restrict its use exclusively to ethical, non-commercial research, especially in contexts that support combating HT.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? Yes

When will the dataset be distributed? The MATCHED dataset will be released after the final decision from the ACL committee, along with the camera-ready version.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No

A.3.7 Maintenance

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? We are committed to enhancing the dataset by exploring advanced NLP-based entity extraction techniques to protect individual privacy further. Specifically, we aim to implement more effective methods for masking escort pseudonyms, posted locations, and ethnicities. Additionally, we plan to expand the dataset by including ads from multiple escort platforms, enabling us to evaluate our models' generalization on real-to-close-world OOD datasets. These updates aim to improve the dataset's privacy measures and its utility for robust, cross-platform AA tasks. Progress and updates will be communicated through research publications, and detailed updates will be made to the dataset's description on the Dataverse portal.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. We encourage researchers to collaborate with us to extend and improve the dataset through extensions, augmentations, or related enhancements. To safeguard the privacy and well-being of individuals in the dataset, we have restricted sharing rights, meaning contributors cannot freely distribute the dataset. However, we invite researchers to work with us directly, and we are open to reviewing and integrating validated contributions to improve the dataset's utility responsibly. We ensure that all validated contributions and updates will be acknowledged and communicated to the research community.

A.4 Infrastructure & Schedule

Split Ratio: We split the dataset into training, validation, and test sets using a standard ratio of 0.75:0.05:0.20 for our experiments. During this process, we set the seed parameter to 1111 for reproducibility.

Training: We conduct model training and evaluation on an NVIDIA H100 GPU with 80 GB of memory. For optimization, we use the Adam optimizer configured with $\beta1$ and $\beta2$ values of 0.9 and 0.999, respectively, along with an L2 weight decay of 0.01. We experiment with learning rates of 0.01, 0.001, and 0.0001, ultimately finding the best performance at a learning rate of 0.001. Additionally, we apply a warm-up strategy for the first 100 steps, followed by a linear decay schedule.

Architectures & Hyperparameters: Considering our computational constraints, we initialize text baselines using pre-trained model checkpoints from DeCLUTR-small and Style-Embedding architectures. Similarly, vision baselines are initialized using pre-trained checkpoints from VGG-16, ResNet-50, DenseNet-121, InceptionNetV3, EfficientNetV2, ConvNext-small, and ViT-base-patch16-244 architectures. We also explore face recognition models such as VGG-Face2 (Cao et al., 2018), ArcFace, FaceNet512 (Firmansyah et al., 2023), and GhostFaceNet (Alansari et al., 2023) from DeepFace (Serengil and Ozpinar, 2023) for the vision baselines. However, these models struggle with vendor identification and verification tasks, likely because they focus solely on facial features, making it challenging to connect multiple faces to a single vendor. We further experimented by training these face recognition models on the face (images with faces) and no face (images without faces) subsets of our dataset. However, the results remained consistent, confirming their unsuitability for these tasks. Finally, the multimodal baselines are initialized by combining the DeCLUTR-small and ViT-base-patch-244 baselines to process text and vision modalities. Each model is equipped with a sequence classification head to perform classification tasks. Due to resource limitations, all models are trained with a batch size of 32, the maximum feasible size, and training continues until convergence.

During model training, we use five in-batch negatives for contrastive objectives such as Triplet, SupCon, CE+Triplet, and CE+SupCon. Increasing the number of in-batch negatives did not improve performance, likely constrained by the fixed batch size of 32 for the classification task. For the text-image alignment pre-training task, we employ the Normalized Temperature-Scaled Cross-Entropy (NT-XENT) loss (Chen et al., 2020) for the Image-Text Contrastive (ITC) objective, sampling negatives from regions outside the training dataset. In all multimodal experiments, negatives are strictly non-associated, ensuring text-image pairs are unrelated ads. We also experiment with temperature coefficient values of 0.01, 0.1, and 0.3 for the NT-XENT loss, finding the best performance at 0.1.

The experiments are implemented in Python 3.10 using frameworks such as scikit-learn (Pedregosa et al., 2011), PyTorch (Paszke et al., 2019), Hugging Face, timm, and Lightning 2.0 (Falcon and The PyTorch Lightning team, 2019). The plots in the research are developed using Matplotlib (Hunter, 2007) and Plotly (Inc., 2015).

Computational Details: Table 4 provides an overview of the number of trainable parameters, training time, and convergence epochs for all the classifiers evaluated in our experiments. Additionally, we dedicated 8 hours 21 minutes and 51 seconds, 1 hour 51 minutes and 6 seconds, and 3 hours 52 minutes and 12 seconds to pre-train our text-image alignment models using ITC (CLIP), ITC+ITM (Image Text Matching loss), and ITC+ITM+Text Generation Loss (BLIP2) training strategies, respectively.

Seed	Acc.	Weighted-F1	Micro-F1	Macro-F1
100	0.9670	0.9862	0.9878	0.9630
500	0.9761	0.9914	0.9921	0.9755
1111	0.9823	0.9911	0.9916	0.9802
Mean	0.9751	0.9896	0.9905	0.9729
Std.	0.0077	0.0029	0.0024	0.0089

Table 3: Influence of random initialization on DeCLUTR-ViT classifier's performance

Random Initialization: Due to limited resources, we only examine the effects of different initializations on our model's performance for the established DeCLUTR-ViT benchmark with the CE+SupCon objective. Table 3 displays the mean and standard deviation in the model's performance against balanced accuracy, Micro-F1, Weighted-F1, and Macro-F1 scores. The results indicate minimal to no effects on these scores across different initializations.

A.5 Model Performance

This section provides detailed insights into our experiments' training and evaluation results, as summarized in the appendix tables. Table 4 outlines the performance of text-only, vision-only, and multimodal classifiers on the vendor identification task. These classifiers were trained on the South region dataset and evaluated using Balanced Accuracy, Weighted-F1, Micro-F1, and Macro-F1 metrics. Given the class imbalance in our datasets, we emphasize Macro-F1 as the primary metric to assess model performance effectively. The models were trained with various objectives, including CE, Triplet, SupCon, CE+Triplet, and CE+SupCon, allowing a comprehensive comparison of their capabilities.

For retrieval tasks, results are detailed in Tables 5, 6, 7, 9, 10, and 11, covering text-to-text, image-to-image, and multimodal retrieval scenarios. While we analyze all three retrieval metrics—MRR@10, R-Precision, and Macro-F1@X—our emphasis is on R-Precision. This metric reflects the model's ability to retrieve all relevant ads linked to a query ad from the same vendor, offering a direct measure of retrieval effectiveness.

As explained in the main manuscript, the Zero-Shot (ZS) performance refers to the capability of pre-trained models to perform retrieval tasks without prior AA training. Pre-trained text-only model is represented in Table 5, vision-only models in Table 7, and text-image alignment models, as Aligned DeCLUTR-ViT, in Tables 9, 10, and 11. These models are evaluated on the South, Midwest, West, and Northeast region datasets without specific AA task training, making them ideal for understanding baseline performance in unseen contexts. Conversely, the Out-of-Data (OOD) average performance measures how well AA models trained for vendor identification or verification tasks generalize to unseen datasets from the Midwest, West, and Northeast regions. This evaluation highlights the models' robustness in handling diverse, previously unseen ads and vendors, offering critical insights into their cross-region generalization capabilities. By contrasting ZS and OOD performance, we assess both the initial adaptability of pre-trained models and the impact of AA-specific training. All the vendor verification metrics are represented in x ± y format, where x and y represent the mean and standard deviation of performance across all vendor classes.

A.5.1 Text-only Modality

The text-baseline results presented in Table 4 demonstrate that the DeCLUTR-small architecture significantly outperforms the Style-Embedding model in terms of Macro-F1 score for the vendor identification task. As a result, the DeCLUTR-small architecture is exclusively used for further experiments involving joint objectives. Among all text-only baselines, the DeCLUTR backbone trained with the CE+SupCon objective achieves the highest performance across all vendor identification metrics, showcasing its effectiveness. For the vendor verification task, retrieval results in Table 5 reveal that the DeCLUTR backbone trained with the CE+SupCon objective consistently outperforms the CE objective and performs comparably to the SupCon-only objective. Additionally, the smaller standard deviation in performance between the CE+SupCon and CE objectives highlights the model's enhanced consistency across all vendor classes, further underscoring the robustness of the CE+SupCon objective for text-only baselines.

A.5.2 Vision-only Modality

The vision baselines in Table 4 highlight that ResNet-50 with CE loss achieves the highest performance among classifiers for the vendor identification task. However, retrieval results in Table 6 show that, despite slightly lower classification performance, the ViT-base-patch16 backbone consistently outperforms other models on both training and OOD datasets for the image-to-image retrieval task. Given our research's dual objectives of vendor identification and verification, we establish the ViT-base-patch16 backbone as the most suitable choice for further experiments. Consistent with the text-only modality findings, Table 7 indicates that using a joint objective with CE+SupCon loss delivers the best results across all vision-only baselines, reinforcing its effectiveness in both classification and retrieval tasks.

A.5.3 Multimodal Modality

The multimodal baselines in Table 4 consistently outperform their text-only and vision-only counterparts on the classification task. Among the fusion techniques explored, mean pooling proves to be the most effective for merging text and vision representations. However, despite pre-training on text-image alignment tasks, the fine-tuned multimodal baselines show limited vendor identification and verification performance. Table 8 highlights the text-to-image retrieval performance of these pre-trained baselines, where, given a query text ad, the goal is to retrieve its associated images from the original ad. The underperformance of these models stems from the lack of semantic alignment in escort ads, as the visual content often fails to correspond meaningfully to the accompanying text. In contrast, as demonstrated in Tables 4, 9, 10, and 11, the DeCLUTR-ViT backbone trained end-to-end with the CE+SupCon objective (without pre-training) achieves superior performance across all tasks, reinforcing the effectiveness of end-to-end training for multimodal AA in this domain.

Model	Param	Loss	Fusion	Epochs	Accuracy	Weighted F1	Micro F1	Macro F1	Time (hrs.)
				Text-Baselines					
Style-Embedding	128M	CE	-	28	0.6582	0.6883	0.6897	0.5210	01:07:12
DeCLUTR-small	**86M**	CE		21	0.7647	0.7772	0.7777	0.6379	0:12:19
		CE+Triplet		10	0.6905	0.7068	0.7074	0.5503	0:07:32
		CE+SupCon		**15**	**0.7786**	**0.7891**	**0.7898**	**0.6540**	**0:06:33**
				Vision-Baselines					
VGG-16	138M	CE	-	9	0.6823	0.6873	0.6884	0.5262	0:15:33
ResNet-50	25M			19	0.7741	0.7777	0.7789	0.6394	0:23:14
DenseNet-121	7M			13	0.7624	0.7656	0.7673	0.6262	0:27:01
InceptionNetV3	23M			12	0.7471	0.7510	0.7524	0.6047	0:20:26
EfficientNetV2	23M			12	0.7652	0.7690	0.7703	0.6285	0:29:29
ConvNeXT-small	50M			7	0.7593	0.7625	0.7646	0.6215	0:16:52
ViT-base-patch16	**86M**	CE		8	0.7559	0.7593	0.7606	0.6142	0:13:16
		CE+Triplet		13	0.7729	0.7765	0.7771	0.6378	0:30:35
		CE+SupCon		**13**	**0.7711**	**0.7709**	**0.7716**	**0.6294**	**0:31:41**
				Multimodal-Baselines					
ViLT	112M	CE	-	12	0.8454	0.8327	0.8291	0.7369	01:18:00
VisualBERT	197M			11	0.9652	0.9637	0.9641	0.9355	01:10:17
DeCLUTR-ViT	171M	CE	auto	11	0.9344	0.9578	0.9565	0.9121	03:41:44
			attention	14	0.8774	0.9184	0.9217	0.8451	03:45:15
	169M		concat	15	0.9422	0.9762	0.9781	0.9411	03:52:36
			mean	16	0.9713	0.9857	0.9861	0.9670	01:02:16
		CE+SupCon	**mean**	**17**	**0.9823**	**0.9911**	**0.9916**	**0.9802**	**01:15:56**
		ITC+CE		18	0.9463	0.9744	0.9760	0.9466	01:17:20
	307M	ITC+ITM+CE		10	0.8456	0.9010	0.8995	0.8443	01:07:17
		BLIP2+CE		11	0.9101	0.9620	0.9644	0.9128	01:14:19
		BLIP2+CE+SupCon		13	0.9450	0.9702	0.9722	0.9420	01:30:57

Table 4: Performance metrics (Balanced Accuracy, Weighted-F1, Micro-F1, and Macro-F1) and computational details for text, vision, and multimodal classifier baselines trained on the South region dataset. Pre-training strategies—ITC, ITC+ITM, and BLIP2—are applied to DeCLUTR-small and ViT-base-patch16 models to align text and images from the same advertisement. Fine-tuning is then conducted for the vendor identification task on the South region dataset, with classifiers optimized using CE, CE+Triplet, and CE+SupCon loss objectives.

Loss	South	Midwest	West	Northeast	OOD Avg.	ZS Avg.
			MRR@10			
Pre-trained	0.2248 ± 0.30	0.2866 ± 0.36	0.3479 ± 0.41	0.3385 ± 0.38	-	0.2995 ± 0.36
CE	0.7445 ± 0.39	0.5703 ± 0.46	0.6394 ± 0.45	0.5862 ± 0.48	0.5986 ± 0.46	-
Triplet	0.4282 ± 0.45	0.3200 ± 0.43	0.4074 ± 0.46	0.3503 ± 0.45	0.3592 ± 0.45	-
SupCon	0.8829 ± 0.29	0.7636 ± 0.39	0.8331 ± 0.35	0.7520 ± 0.42	0.7829 ± 0.39	-
CE+Triplet	0.8891 ± 0.28	0.6410 ± 0.45	0.6969 ± 0.43	0.6561 ± 0.45	0.6647 ± 0.44	-
CE+SupCon	**0.9290 ± 0.23**	**0.7716 ± 0.38**	**0.8145 ± 0.36**	**0.7449 ± 0.42**	**0.7770 ± 0.39**	-
			R-Precision@X			
Pre-trained	0.3265 ± 0.47	0.3943 ± 0.49	0.3139 ± 0.46	0.4037 ± 0.49	-	0.3596 ± 0.48
CE	0.5557 ± 0.36	0.4596 ± 0.40	0.5842 ± 0.41	0.4944 ± 0.43	0.5127 ± 0.41	-
Triplet	0.3200 ± 0.34	0.2443 ± 0.33	0.3365 ± 0.38	0.3032 ± 0.38	0.2947 ± 0.36	-
SupCon	0.7673 ± 0.29	0.6346 ± 0.37	0.7612 ± 0.35	0.6707 ± 0.41	0.6888 ± 0.38	-
CE+Triplet	0.8055 ± 0.30	0.5000 ± 0.40	0.5890 ± 0.4	0.5410 ± 0.42	0.5433 ± 0.41	-
CE+SupCon	**0.8706 ± 0.24**	**0.6264 ± 0.38**	**0.7339 ± 0.37**	**0.6699 ± 0.41**	**0.6767 ± 0.39**	-
			Macro-F1@X			
Pre-trained	0.2224 ± 0.30	0.2804 ± 0.36	0.2731 ± 0.36	0.3801 ± 0.39	-	0.2890 ± 0.37
CE	0.6098 ± 0.35	0.4760 ± 0.38	0.6123 ± 0.35	0.5042 ± 0.42	0.5308 ± 0.38	-
Triplet	0.4135 ± 0.37	0.2892 ± 0.35	0.4337 ± 0.35	0.3121 ± 0.39	0.3450 ± 0.36	-
SupCon	0.8157 ± 0.27	0.6333 ± 0.36	0.7408 ± 0.31	0.6950 ± 0.39	0.6897 ± 0.35	-
CE+Triplet	0.8680 ± 0.26	0.5198 ± 0.39	0.5789 ± 0.35	0.5612 ± 0.41	0.5533 ± 0.38	-
CE+SupCon	**0.9102 ± 0.21**	**0.6162 ± 0.37**	**0.7169 ± 0.33**	**0.6879 ± 0.40**	**0.6737 ± 0.37**	-

Table 5: Comparison of text-to-text retrieval performance for the text-only benchmark, DeCLUTR-small backbone, with different objectives (losses), evaluated across MRR@10, R-Precision@X, and Macro-F1@X metrics.

Loss	South	Midwest	West	Northeast	OOD Avg.
			MRR@10		
VGG16	0.0069 ± 0.05	0.0098 ± 0.07	0.0491 ± 0.19	0.0172 ± 0.1	0.0254 ± 0.12
ResNet50	0.1026 ± 0.22	0.1569 ± 0.29	0.221 ± 0.35	0.125 ± 0.26	0.1676 ± 0.30
Densenet121	0.218 ± 0.32	0.2465 ± 0.35	0.2669 ± 0.37	0.1889 ± 0.32	0.2341 ± 0.35
InceptionNetV3	0.0477 ± 0.15	0.0583 ± 0.19	0.0684 ± 0.2	0.0625 ± 0.19	0.0631 ± 0.19
EfficientNetV2	0.2305 ± 0.32	0.2468 ± 0.35	0.2523 ± 0.36	0.2276 ± 0.35	0.2422 ± 0.35
ConvNext-small	0.0588 ± 0.17	0.0851 ± 0.22	0.0854 ± 0.23	0.0917 ± 0.24	0.0874 ± 0.23
ViT-base-patch16	**0.2587 ± 0.33**	**0.2854 ± 0.37**	**0.3019 ± 0.39**	**0.2597 ± 0.36**	**0.2823 ± 0.37**
			R-Precision@X		
VGG16	0.0063 ± 0.03	0.0074 ± 0.03	0.0165 ± 0.05	0.0139 ± 0.06	0.0126 ± 0.05
ResNet50	0.0267 ± 0.05	0.0415 ± 0.09	0.0599 ± 0.1	0.0452 ± 0.09	0.0489 ± 0.09
Densenet121	0.0413 ± 0.08	0.0618 ± 0.11	0.0849 ± 0.11	0.0671 ± 0.11	0.0713 ± 0.11
InceptionNetV3	0.0084 ± 0.02	0.0176 ± 0.07	0.0224 ± 0.06	0.0143 ± 0.04	0.0181 ± 0.06
EfficientNetV2	0.0417 ± 0.07	0.0609 ± 0.1	0.0752 ± 0.11	0.0692 ± 0.11	0.0684 ± 0.11
ConvNext-small	0.0157 ± 0.04	0.026 ± 0.06	0.0299 ± 0.07	0.0291 ± 0.06	0.0283 ± 0.06
ViT-base-patch16	**0.0459 ± 0.07**	**0.0645 ± 0.11**	**0.0781 ± 0.11**	**0.078 ± 0.13**	**0.0735 ± 0.12**
			Macro-F1@X		
VGG16	0.0091 ± 0.03	0.0151 ± 0.04	0.0171 ± 0.06	0.0158 ± 0.05	0.0160 ± 0.05
ResNet50	0.0276 ± 0.06	0.0407 ± 0.08	0.0565 ± 0.1	0.0468 ± 0.09	0.0479 ± 0.09
Densenet121	0.04 ± 0.07	0.0535 ± 0.09	0.0823 ± 0.12	0.0641 ± 0.11	0.0666 ± 0.11
InceptionNetV3	0.0083 ± 0.02	0.0154 ± 0.05	0.0215 ± 0.06	0.0147 ± 0.04	0.0172 ± 0.05
EfficientNetV2	0.042 ± 0.07	0.0546 ± 0.09	0.0764 ± 0.12	0.0648 ± 0.1	0.0653 ± 0.10
ConvNext-small	0.0159 ± 0.04	0.028 ± 0.06	*0.0312 ± 0.07*	0.0301 ± 0.06	0.0298 ± 0.06
ViT-base-patch16	**0.0436 ± 0.07**	**0.0574 ± 0.09**	**0.077 ± 0.12**	**0.0727 ± 0.11**	**0.0690 ± 0.11**

Table 6: Comparison of image-to-image retrieval performance for the vision-baselines trained on south region image ads with CE loss, evaluated on MRR@10, R-Precision@X, and Macro-F1@X metrics

Loss	South	Midwest	West	Northeast	OOD Avg.	ZS Avg.
			MRR@10			
Pre-trained	0.2286 ± 0.32	0.2432 ± 0.35	0.2517 ± 0.36	0.2242 ± 0.35	-	0.2369 ± 0.35
CE	0.2587 ± 0.33	0.2854 ± 0.37	0.3019 ± 0.39	0.2597 ± 0.36	0.2823 ± 0.37	-
SupCon	0.0010 ± 0.03	0.0013 ± 0.03	0.0031 ± 0.03	0.0079 ± 0.08	0.0041 ± 0.05	-
Triplet	0.0010 ± 0.03	0.0016 ± 0.04	0.0035 ± 0.06	0.0054 ± 0.07	0.0035 ± 0.06	-
CE+Triplet	0.2760 ± 0.35	0.3242 ± 0.39	0.366 ± 0.41	0.3322 ± 0.39	0.3408 ± 0.40	-
CE+SupCon	**0.3464 ± 0.37**	**0.3749 ± 0.40**	**0.4049 ± 0.42**	**0.4330 ± 0.42**	**0.4041 ± 0.41**	**-**
			R-Precision@X			
Pre-trained	0.0420 ± 0.07	0.0593 ± 0.10	0.0754 ± 0.11	0.0691 ± 0.11	-	0.0615 ± 0.10
CE	0.0459 ± 0.07	0.0645 ± 0.11	0.0781 ± 0.11	0.078 ± 0.13	0.0735 ± 0.12	-
SupCon	0.0010 ± 0.01	0.0018 ± 0.01	0.0028 ± 0.01	0.0028 ± 0.02	0.0025 ± 0.01	-
Triplet	0.0009 ± 0.01	0.0007 ± 0.01	0.0017 ± 0.02	0.003 ± 0.02	0.0018 ± 0.02	-
CE+Triplet	0.0824 ± 0.14	0.0963 ± 0.15	0.139 ± 0.19	0.1281 ± 0.17	0.1211 ± 0.17	-
CE+SupCon	**0.1064 ± 0.16**	**0.1095 ± 0.16**	**0.1519 ± 0.20**	**0.1685 ± 0.21**	**0.1433 ± 0.19**	**-**
			Macro-F1@X			
Pre-trained	0.0421 ± 0.07	0.0539 ± 0.09	0.0767 ± 0.12	0.0647 ± 0.1	-	0.0594 ± 0.10
CE	0.0436 ± 0.07	0.0574 ± 0.09	0.077 ± 0.12	0.0727 ± 0.11	0.0690 ± 0.11	-
SupCon	0.0015 ± 0.01	0.0041 ± 0.01	0.0043 ± 0.02	0.0034 ± 0.02	0.0039 ± 0.02	-
Triplet	0.0011 ± 0.01	0.0031 ± 0.01	0.0028 ± 0.01	0.0026 ± 0.01	0.0028 ± 0.01	-
CE+Triplet	0.1091 ± 0.20	0.0842 ± 0.14	0.1413 ± 0.2	0.1143 ± 0.17	0.1133 ± 0.17	-
CE+SupCon	**0.1296 ± 0.21**	**0.0948 ± 0.14**	**0.1460 ± 0.20**	**0.1497 ± 0.20**	**0.1302 ± 0.18**	**-**

Table 7: Comparison of image-to-image retrieval performance for the vision-only benchmark, ViT-base-patch16 backbone, with different objectives (losses), evaluated on MRR@10, R-Precision@X, and Macro-F1@X metrics.

Loss	South	Midwest	West	Northeast	Avg.
Alignment MRR@10					
ITC	0.0001 ± 0.01	0.0001 ± 0.01	0.0003 ± 0.02	0.0004 ± 0.02	0.0002 ± 0.01
ITC+ITM	0.0001 ± 0.01	0.0001 ± 0.01	0.0003 ± 0.02	0.0008 ± 0.03	0.0003 ± 0.02
BLIP2	0.001 ± 0.03	0.0027 ± 0.05	0.0063 ± 0.08	0.0098 ± 0.10	0.0050 ± 0.07
Alignment R-Precision@X					
ITC	0.0001 ± 0.01	0.0002 ± 0.01	0.0013 ± 0.03	0.0005 ± 0.01	0.0005 ± 0.01
ITC+ITM	0.0002 ± 0.01	0.0002 ± 0.01	0.0006 ± 0.01	0.0007 ± 0.01	0.0004 ± 0.01
BLIP2	0.0017 ± 0.02	0.0049 ± 0.04	0.0103 ± 0.06	0.0104 ± 0.06	0.0068 ± 0.05
Alignment Macro-F1@X					
ITC	0.0001 ± 0.01	0.0002 ± 0.01	0.0013 ± 0.03	0.0005 ± 0.01	0.0005 ± 0.02
ITC+ITM	0.0002 ± 0.01	0.0002 ± 0.01	0.0006 ± 0.01	0.0007 ± 0.01	0.0004 ± 0.01
BLIP2	0.0017 ± 0.02	0.0049 ± 0.04	0.0103 ± 0.06	0.0104 ± 0.06	0.0068 ± 0.05

Table 8: Text-to-Image retrieval results from the multimodal DeCLUTR-ViT backbone pre-trained on the text-image alignment task using CLIP (ITC), ITC+ITM (Image text matching loss), BLIP2 (ITC+ITM+Text generation loss).

Backbone	Loss	South	Midwest	West	Northeast	OOD Avg.	ZS Avg.
Text MRR@10							
DeCLUTR	**CE+SupCon**	0.9290 ± 0.23	0.7716 ± 0.38	0.8145 ± 0.36	0.7449 ± 042	0.7770 ± 0.39	
End2End	**CE**	0.9850 ± 0.10	0.9693 ± 0.14	0.9900 ± 0.07	0.9778 ± 0.12	0.9790 ± 0.11	-
DeCLUTR-ViT	**CE+SupCon**	**0.9866 ± 0.09**	**0.9704 ± 0.14**	**0.9932 ± 0.07**	**0.9821 ± 0.12**	**0.9819 ± 0.11**	-
Aligned DeCLUTR-ViT	**ITC**	0.4097 ± 0.43	0.4289 ± 0.45	0.5404 ± 0.47	0.5034 ± 0.47	-	0.4909 ± 0.46
	ITC+ITM	0.8192 ± 0.37	0.7990 ± 0.39	0.8600 ± 0.35	0.5914 ± 0.48	-	0.7674 ± 0.40
	BLIP2	0.7551 ± 0.41	0.7226 ± 0.44	0.8400 ± 0.37	0.5376 ± 0.49	-	0.7140 ± 0.43
	BLIP2-Cond	0.7672 ± 0.41	0.7203 ± 0.44	0.8400 ± 0.37	0.4946 ± 0.49	-	0.7055 ± 0.43
Fine-tuned DeCLUTR-ViT	**ITC+CE**	0.8613 ± 0.34	0.6623 ± 0.46	0.8600 ± 0.35	0.6263 ± 0.48	0.7162 ± 0.43	-
	ITC+ITM+CE	0.4239 ± 0.39	0.2851 ± 0.37	0.3417 ± 0.42	0.3600 ± 0.41	0.3289 ± 0.40	-
	BLIP2+CE	0.8866 ± 0.30	0.7226 ± 0.44	0.8400 ± 0.37	0.7292 ± 0.44	0.7639 ± 0.42	-
	BLIP2+CE+SupCon	0.8886 ± 0.31	0.7397 ± 0.43	0.8600 ± 0.35	0.7604 ± 0.42	0.7867 ± 0.40	-
Text R-Precision@X							
DeCLUTR	**CE+SupCon**	0.8706 ± 0.24	0.6264 ± 0.38	0.7339 ± 0.37	0.6699 ± 0.41	0.6767 ± 0.39	
End2End	**CE**	0.8687 ± 0.19	0.6500 ± 0.30	0.7934 ± 0.24	0.7300 ± 0.28	0.7245 ± 0.27	-
DeCLUTR-ViT	**CE+SupCon**	**0.9193 ± 0.16**	**0.6612 ± 0.31**	**0.8008 ± 0.25**	**0.7365 ± 0.28**	**0.7418 ± 0.28**	-
Aligned DeCLUTR-ViT	**ITC**	0.2337 ± 0.28	0.2936 ± 0.34	0.4035 ± 0.37	0.3779 ± 0.38	-	0.3583 ± 0.36
	ITC+ITM	0.4964 ± 0.34	0.5679 ± 0.38	0.7093 ± 0.33	0.4818 ± 0.45	-	0.5639 ± 0.38
	BLIP2	0.4230 ± 0.34	0.5094 ± 0.39	0.6354 ± 0.37	0.3913 ± 0.41	-	0.4898 ± 0.38
	BLIP2-Cond	0.4341 ± 0.35	0.5142 ± 0.39	0.6644 ± 0.36	0.3729 ± 0.42	-	0.4964 ± 0.38
Fine-tuned DeCLUTR-ViT	**ITC+CE**	0.6378 ± 0.33	0.4885 ± 0.37	0.6825 ± 0.35	0.3770 ± 0.35	0.5160 ± 0.36	-
	ITC+ITM+CE	0.1462 ± 0.19	0.0818 ± 0.14	0.1292 ± 0.18	0.1359 ± 0.19	0.1156 ± 0.17	-
	BLIP2+CE	0.7131 ± 0.32	0.5569 ± 0.39	0.7280 ± 0.36	0.5627 ± 0.41	0.6159 ± 0.39	-
	BLIP2+CE+SupCon	0.7632 ± 0.32	0.5666 ± 0.40	0.7652 ± 0.31	0.5869 ± 0.40	0.6362 ± 0.37	-
Text Macro-F1@X							
DeCLUTR	**CE+SupCon**	0.9102 ± 0.21	0.6162 ± 0.37	0.7169 ± 0.33	0.6879 ± 0.40	0.6737 ± 0.37	
End2End	**CE**	0.8726 ± 0.20	0.5653 ± 0.33	0.7374 ± 0.26	0.7261 ± 0.31	0.6763 ± 0.30	-
DeCLUTR-ViT	**CE+SupCon**	**0.9433 ± 0.16**	**0.5819 ± 0.34**	**0.7466 ± 0.26**	**0.7242 ± 0.31**	**0.6841 ± 0.30**	-
Aligned DeCLUTR-ViT	**ITC**	0.3039 ± 0.31	0.3756 ± 0.35	0.4887 ± 0.33	0.4173 ± 0.39	-	0.4272 ± 0.36
	ITC+ITM	0.5079 ± 0.32	0.5659 ± 0.36	0.7281 ± 0.29	0.5136 ± 0.44	-	0.5946 ± 0.35
	BLIP2	0.4283 ± 0.33	0.5279 ± 0.38	*0.6552 ± 0.34*	0.4216 ± 0.39	-	0.5605 ± 0.38
	BLIP2-Cond	0.4356 ± 0.34	0.5251 ± 0.38	0.6720 ± 0.34	0.4249 ± 0.42	-	0.5125 ± 0.38
Fine-tuned DeCLUTR-ViT	**ITC+CE**	0.6805 ± 0.32	0.5054 ± 0.37	0.6877 ± 0.32	0.3790 ± 0.34	0.5240 ± 0.34	-
	ITC+ITM+CE	0.1438 ± 0.20	0.0748 ± 0.12	0.1218 ± 0.18	0.1214 ± 0.17	0.1060 ± 0.16	-
	BLIP2+CE	0.7215 ± 0.31	0.5774 ± 0.38	0.7391 ± 0.32	0.5499 ± 0.39	0.6221 ± 0.36	-
	BLIP2+CE+SupCon	0.7879 ± 0.29	0.5762 ± 0.39	0.7482 ± 0.29	0.5912 ± 0.38	0.6385 ± 0.35	-

Table 9: Comparison of text-to-text retrieval performance for the multimodal, DeCLUTR-ViT backbone, evaluated on the text-only modality using MRR@10, R-Precision@X, and Macro-F1@X metrics. The DeCLUTR-small model serves as the text-only baseline. End2End baselines denote DeCLUTR-ViT models trained directly for vendor identification tasks, while Aligned baselines represent DeCLUTR-ViT backbone pre-trained for text-image alignment tasks using ITC, ITC+ITM, and BLIP2 objectives. Fine-tuned baselines build upon pre-trained aligned models by fine-tuning them for vendor identification tasks on the South region ads.

Backbone	Loss	South	Midwest	West	Northeast	OOD Avg.	ZS Avg.
Vision MRR@10							
ViT	**CE+SupCon**	0.3464 ± 0.37	0.3749 ± 0.40	0.4049 ± 0.42	0.4330 ± 0.42	0.4041 ± 0.41	-
End2End	**CE**	0.2257 ± 0.33	0.1716 ± 0.32	0.2142 ± 0.35	0.1866 ± 0.32	0.2575 ± 0.33	-
DeCLUTR-ViT	**CE+SupCon**	**0.4045 ± 0.38**	**0.3905 ± 0.40**	**0.4603 ± 0.45**	**0.4521 ± 0.42**	**0.4343 ± 0.42**	-
Aligned DeCLUTR-ViT	**ITC**	0.2329 ± 0.30	0.2336 ± 0.33	0.2984 ± 0.39	0.2964 ± 0.37	-	0.2761 ± 0.36
	ITC+ITM	0.3281 ± 0.37	0.3434 ± 0.39	0.3683 ± 0.43	0.3442 ± 0.40	-	0.3324 ± 0.38
	BLIP2	0.2119 ± 0.32	0.2055 ± 0.33	0.2674 ± 0.40	0.2858 ± 0.39	-	0.2425 ± 0.36
	BLIP2-Cond	0.2049 ± 0.32	0.1855 ± 0.31	0.2488 ± 0.39	0.2450 ± 0.36	-	0.2211 ± 0.35
Fine-tuned DeCLUTR-ViT	**ITC**	0.4157 ± 0.38	0.3512 ± 0.39	0.3818 ± 0.43	0.3792 ± 0.41	0.3707 ± 0.41	-
	ITC+ITM	0.4239 ± 0.39	0.2851 ± 0.37	0.3417 ± 0.42	0.3600 ± 0.41	0.3289 ± 0.40	-
	BLIP2	0.3677 ± 0.38	0.2629 ± 0.36	0.3229 ± 0.41	0.3128 ± 0.39	0.2995 ± 0.39	-
	BLIP2-CE+SupCon	0.3470 ± 0.38	0.2542 ± 0.35	0.3026 ± 0.41	0.3312 ± 0.39	0.2960 ± 0.39	-
Vision R-Precision@X							
ViT	**CE+SupCon**	0.1064 ± 0.16	0.1095 ± 0.16	0.1519 ± 0.20	0.1685 ± 0.21	0.1433 ± 0.19	-
End2End	**CE**	0.0862 ± 0.16	0.0567 ± 0.12	0.0915 ± 0.14	0.0676 ± 0.11	0.0719 ± 0.12	-
DeCLUTR-ViT	**CE+SupCon**	**0.1115 ± 0.15**	**0.1141 ± 0.16**	**0.1768 ± 0.21**	**0.1646 ± 0.19**	**0.1518 ± 0.19**	-
Aligned DeCLUTR-ViT	**ITC**	0.0537 ± 0.09	0.0752 ± 0.13	0.1275 ± 0.17	0.1143 ± 0.16	-	0.1057 ± 0.16
	ITC+ITM	0.0650 ± 0.10	0.0826 ± 0.14	0.1218 ± 0.17	0.1003 ± 0.14	-	0.0924 ± 0.14
	BLIP2	0.0645 ± 0.15	0.0641 ± 0.13	0.1197 ± 0.20	0.1492 ± 0.24	-	0.0994 ± 0.18
	BLIP2-Cond	0.0563 ± 0.13	0.0569 ± 0.13	0.1001 ± 0.18	0.1115 ± 0.20	-	0.0812 ± 0.16
Fine-tuned DeCLUTR-ViT	**ITC**	0.1247 ± 0.17	0.0957 ± 0.15	0.1461 ± 0.18	0.1383 ± 0.17	0.1267 ± 0.17	-
	ITC+ITM	0.1462 ± 0.19	0.0818 ± 0.14	0.1292 ± 0.18	0.1359 ± 0.19	0.1156 ± 0.17	-
	BLIP2	0.1370 ± 0.19	0.0775 ± 0.14	0.1217 ± 0.18	0.1393 ± 0.21	0.1128 ± 0.18	-
	BLIP2-CE+SupCon	0.1256 ± 0.19	0.0777 ± 0.14	0.1228 ± 0.17	0.1414 ± 0.20	0.1140 ± 0.17	-
Vision Macro-F1@X							
ViT	**CE+SupCon**	0.1296 ± 0.21	0.0948 ± 0.14	0.1460 ± 0.20	0.1497 ± 0.20	0.1302 ± 0.18	-
End2End	**CE**	0.1028 ± 0.21	0.0600 ± 0.11	0.0960 ± 0.15	0.0657 ± 0.11	0.0859 ± 0.14	-
DeCLUTR-ViT	**CE+SupCon**	**0.1152 ± 0.17**	**0.1049 ± 0.14**	**0.1739 ± 0.21**	**0.1493 ± 0.18**	**0.1427 ± 0.19**	-
Aligned DeCLUTR-ViT	**ITC**	0.0689 ± 0.11	0.0892 ± 0.14	0.1415 ± 0.19	0.1072 ± 0.15	-	0.1118 ± 0.18
	ITC+ITM	0.0614 ± 0.10	0.0675 ± 0.11	0.1070 ± 0.15	0.0933 ± 0.13	-	0.0837 ± 0.13
	BLIP2	0.0938 ± 0.20	0.0908 ± 0.17	0.1281 ± 0.22	0.1458 ± 0.24	-	0.1146 ± 0.21
	BLIP2-Cond	0.0805 ± 0.18	0.0776 ± 0.16	0.1074 ± 0.20	0.1088 ± 0.19	-	0.0936 ± 0.18
Fine-tuned DeCLUTR-ViT	**ITC**	0.1319 ± 0.18	0.0914 ± 0.14	0.1485 ± 0.20	0.1333 ± 0.17	0.1244 ± 0.17	-
	ITC+ITM	0.1438 ± 0.20	0.0748 ± 0.12	0.1218 ± 0.18	0.1214 ± 0.17	0.1060 ± 0.16	-
	BLIP2	0.1517 ± 0.23	0.0837 ± 0.14	0.1277 ± 0.19	0.1367 ± 0.20	0.1160 ± 0.18	-
	BLIP2-CE+SupCon	0.1526 ± 0.24	0.0799 ± 0.14	0.1276 ± 0.19	0.1335 ± 0.20	0.1137 ± 0.18	-

Table 10: Comparison of image-to-image retrieval performance for the multimodal, DeCLUTR-ViT backbone, evaluated on the vision-only modality using MRR@10, R-Precision@X, and Macro-F1@X metrics. The ViT-base-patch16-244 model serves as the vision-only baseline. End2End baselines denote DeCLUTR-ViT models trained directly for vendor identification tasks, while Aligned baselines represent DeCLUTR-ViT backbone pre-trained for text-image alignment tasks using ITC, ITC+ITM, and BLIP2 objectives. Fine-tuned baselines build upon pre-trained aligned models by fine-tuning them for vendor identification tasks on the South region ads.

Backbone	Loss	South	Midwest	West	Northeast	OOD Avg.	ZS Avg.
Multimodal MRR@10							
End2End DeCLUTR-ViT	**CE**	0.9669 ± 0.13	0.9297 ± 0.20	0.9592 ± 0.17	0.9650 ± 0.14	0.9513 ± 0.17	-
	CE+SupCon	**0.9859 ± 0.10**	**0.9658 ± 0.15**	**0.9834 ± 0.11**	**0.9735 ± 0.13**	**0.9742 ± 0.13**	-
Aligned DeCLUTR-ViT	**ITC**	0.6574 ± 0.35	0.6822 ± 0.36	0.7396 ± 0.36	0.6750 ± 0.38	-	0.6886 ± 0.36
	ITC+ITM	0.9375 ± 0.18	0.9389 ± 0.19	0.9601 ± 0.16	0.9715 ± 0.14	-	0.9520 ± 0.17
	BLIP2	0.6142 ± 0.36	0.6136 ± 0.39	0.6108 ± 0.41	0.5921 ± 0.42	-	0.6077 ± 0.40
	BLIP2-Cond	0.6052 ± 0.36	0.6006 ± 0.39	0.5975 ± 0.41	0.5657 ± 0.42	-	0.5923 ± 0.40
Fine-tuned DeCLUTR-ViT	**ITC**	0.9650 ± 0.13	0.8331 ± 0.29	0.7313 ± 0.36	0.7641 ± 0.34	0.7762 ± 0.33	-
	ITC+ITM	0.9739 ± 0.12	0.9285 ± 0.20	0.9498 ± 0.19	0.9655 ± 0.15	0.9480 ± 0.23	-
	BLIP2	0.9774 ± 0.11	0.9378 ± 0.20	0.9559 ± 0.18	0.9690 ± 0.14	0.9542 ± 0.17	-
	BLIP2-CE+SupCon	0.9814 ± 0.10	0.9426 ± 0.19	0.9648 ± 0.15	0.9759 ± 0.12	0.9602 ± 0.19	
Multimodal R-Precision@X							
End2End DeCLUTR-ViT	**CE**	0.8040 ± 0.20	0.6217 ± 0.26	0.7429 ± 0.24	0.6980 ± 0.27	0.6875 ± 0.26	-
	CE+SupCon	**0.9248 ± 0.14**	**0.6567 ± 0.30**	**0.7861 ± 0.25**	**0.7178 ± 0.30**	**0.7202 ± 0.28**	-
Aligned DeCLUTR-ViT	**ITC**	0.1797 ± 0.16	0.2373 ± 0.20	0.3330 ± 0.23	0.3076 ± 0.24	-	0.2644 ± 0.21
	ITC+ITM	0.4939 ± 0.24	0.5705 ± 0.26	0.7046 ± 0.23	0.6747 ± 0.26	-	0.6109 ± 0.25
	BLIP2	0.1708 ± 0.22	0.1847 ± 0.22	0.2182 ± 0.24	0.2841 ± 0.32	-	0.2145 ± 0.25
	BLIP2-Cond	0.1455 ± 0.20	0.1602 ± 0.20	0.1830 ± 0.21	0.2324 ± 0.29	-	0.1803 ± 0.23
Fine-tuned DeCLUTR-ViT	**ITC**	0.7377 ± 0.21	0.3716 ± 0.22	0.2844 ± 0.22	0.3700 ± 0.26	0.3420 ± 0.23	-
	ITC+ITM	0.7282 ± 0.22	0.4968 ± 0.23	0.6109 ± 0.23	0.6419 ± 0.27	0.5832 ± 0.24	-
	BLIP2	0.7723 ± 0.2	0.5524 ± 0.25	0.6759 ± 0.23	0.6691 ± 0.27	0.6325 ± 0.25	-
	BLIP2-CE+SupCon	0.7950 ± 0.19	0.5564 ± 0.25	0.6943 ± 0.23	0.6809 ± 0.26	0.6524 ± 0.25	
Multimodal Macro-F1@X							
End2End DeCLUTR-ViT	**CE**	0.8294 ± 0.21	0.5618 ± 0.29	0.7408 ± 0.24	0.7053 ± 0.29	0.6693 ± 0.27	-
	CE+SupCon	**0.9595 ± 0.12**	**0.5671 ± 0.33**	**0.7560 ± 0.26**	**0.7333 ± 0.30**	**0.6855 ± 0.29**	-
Aligned DeCLUTR-ViT	**ITC**	0.2519 ± 0.23	0.3254 ± 0.26	0.4687 ± 0.27	0.3493 ± 0.26	-	0.3488 ± 0.26
	ITC+ITM	0.4809 ± 0.27	0.5239 ± 0.28	0.7023 ± 0.23	0.6934 ± 0.27	-	0.6001 ± 0.26
	BLIP2	0.3263 ± 0.35	0.3408 ± 0.35	0.4612 ± 0.37	0.4190 ± 0.38	-	0.3868 ± 0.37
	BLIP2-Cond	0.2724 ± 0.32	0.2850 ± 0.32	0.3649 ± 0.33	0.3353 ± 0.35	-	0.3144 ± 0.33
Fine-tuned DeCLUTR-ViT	**ITC**	0.7698 ± 0.23	0.4008 ± 0.25	0.4003 ± 0.27	0.3881 ± 0.28	0.3964 ± 0.27	-
	ITC+ITM	0.7313 ± 0.25	0.4538 ± 0.26	0.6275 ± 0.24	0.6591 ± 0.28	0.5801 ± 0.27	-
	BLIP2	0.7973 ± 0.22	0.5325 ± 0.28	0.7050 ± 0.24	0.6944 ± 0.29	0.6440 ± 0.27	-
	BLIP2-CE+SupCon	0.8487 ± 0.20	0.5446 ± 0.29	0.7250 ± 0.24	0.7077 ± 0.29	0.6591 ± 0.27	

Table 11: Comparison of multimodal retrieval performance for the DeCLUTR-ViT backbone evaluated on the multimodal (text and image) ads using MRR@10, R-Precision@X, and Macro-F1@X metrics. The End2End baselines represent the DeCLUTR-ViT backbone trained directly on the vendor identification task, while the Pre-trained baselines involve an image-text alignment task aligning text and images from the same advertisements. The Fine-tuned baselines build upon the Pre-trained models by performing vendor identification on the South region multimodal ads.

A.6 Further Insights

This section evaluates the multimodal DeCLUTR-ViT backbone trained with the CE+SupCon objective, generating comprehensive insights into model learning and retrieval performance. All line plots have been smoothed for clarity and readability by setting the window size to 30.

A.6.1 Insights from the Multimodal Classifier on the South Region Dataset

Figure 3(i) compares the average F1 performance of the DeCLUTR-small text-only, ViT-base-patch16-244 vision-only, and multimodal DeCLUTR-ViT classifiers for vendors in the South region dataset. The results show that the multimodal classifier consistently outperforms text- and vision-only baselines across all vendors. Further analysis, supported by the vendor frequency distribution in Table 2c and 3(ii), indicates that many vendors in the text-only and vision-only datasets have very few ads, likely contributing to the lower model performance. In contrast, the multimodal classifier benefits from more training examples per vendor (at least five examples when combining text and vision data). This expanded training set allows the model to capture a broader range of stylistic and visual patterns, resulting in better performance. The findings underscore the importance of multimodal integration in enhancing model effectiveness to capture richer and more complementary stylometric cues, particularly for vendors with sparse data in individual modalities.

Figure 3(iii) compares the average number of true positives and false positives achieved by the text-only, vision-only, and multimodal DeCLUTR-ViT baselines across all vendors in the South region dataset. The results reveal a clear advantage for the multimodal baseline, which yields significantly more true positives while maintaining fewer false positives than the other baselines. The results emphasize the superiority of multimodal approaches in minimizing errors and improving the reliability of predictions.

Figure 3(iv) illustrates the average F1 performance of the text-only, vision-only, and multimodal baselines as a function of the number of names per vendor present in the text ads. Since multiple escort names likely represent different individuals, this analysis assesses the models' ability to link varying text descriptions and facial features to a single vendor. To extract escort names from the text ads, we utilized (Li et al., 2022b), though manual inspection revealed that it often failed to extract names accurately. However, the extracted entities remained consistent, allowing us to use them as unique identifiers representing escort names. The results indicate that the multimodal baseline consistently outperforms the text-only and vision-only baselines, demonstrating resilience and robust performance even as the number of escort names per vendor increases.

Finally, Figure 3(v) and (vi) compare the average F1 performance of the vision-only and multimodal baselines as a function of the number of images with and without faces per vendor. In Figure (v), as the number of images with faces increases, the multimodal baseline performs worse than the vision-only baseline up to approximately 120 images. Beyond this threshold, the multimodal baseline either outperforms or performs on par with the vision-only baseline, indicating its ability to adapt as the data volume increases. In contrast, Figure (vi) shows that for images without faces, the multimodal baseline consistently outperforms the vision-only baseline, demonstrating its superior capacity to effectively leverage text and image features, even when facial features are absent.

A.6.2 Insights from the Multimodal Retriever on the OOD Datasets

In this section, we analyze retrieval performance by comparing our multimodal DeCLUTR-ViT baseline against the text-only (DeCLUTR-small) and vision-only (ViT-base-patch16-244) baselines, all trained with the CE+SupCon objective on the South (Figure 4), Midwest (Figure 5), West (Figure 6), and Northeast (Figure 7) region datasets. To further contextualize our findings, we also evaluate the text-only (M-Text) and vision-only (M-Vision) representations extracted from the multimodal baseline, comparing their performance against the standalone text-only and vision-only baselines. Additionally, we assess the Vision-Face and Multimodal-Face baselines, which analyze the performance of vision-only and multimodal models, specifically on images with and without faces. Below, we present the consolidated insights across all regions, structured according to the key factors influencing performance: vendors, ad frequency, number of names, and the presence or absence of faces in images.

Performance per Vendor: Across all regions, the multimodal baseline consistently outperforms text-only and vision-only baselines for both

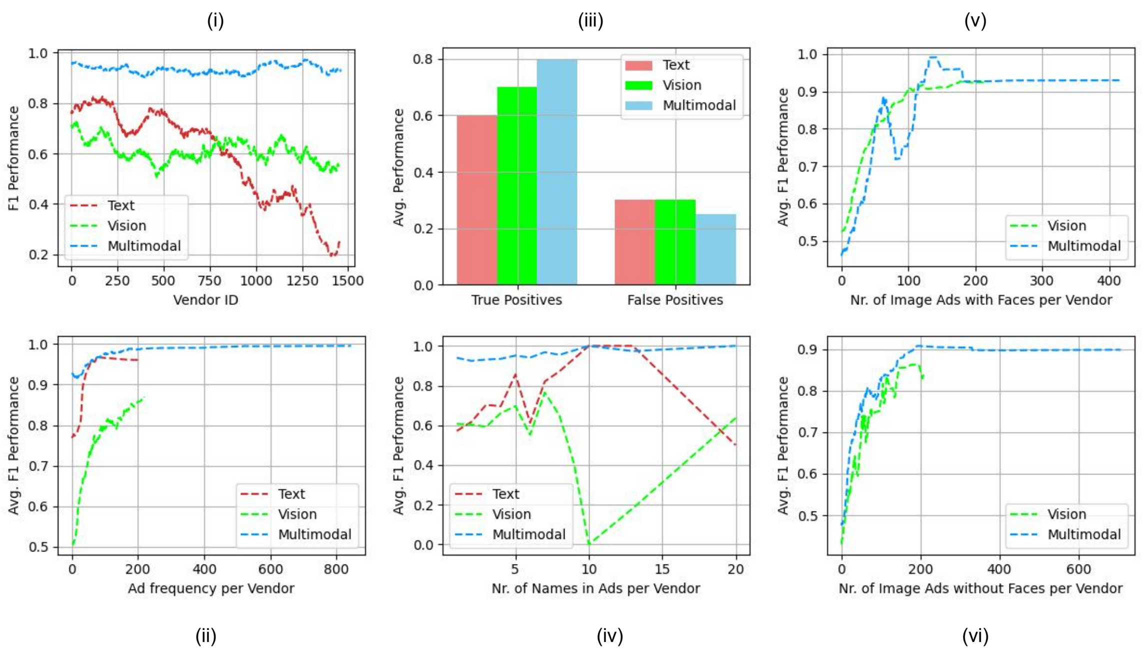


Figure 3: Comparison of model performance among text-only, vision-only, and multimodal classifiers trained on the South region test dataset: (i) F1 score across different vendor IDs, (ii) Average F1 score for vendors with varying ad frequencies, (iii) Analysis of true and false positives, (iv) Average F1 score relative to the number of escort names (potentially representing different individuals) in vendor ads, and (v, vi) Average F1 score based on the number of vendor images with and without faces.

MRR@10 and R-Precision@X. This performance advantage underscores the power of integrating textual and visual cues, which capture complementary information. The M-Text and M-Vision representations, extracted from the multimodal model, also outperform their respective standalone baselines. Notably, the text-only baseline performs better than the vision-only baseline, emphasizing the dominant role of text in vendor identification and retrieval tasks. However, the multimodal baseline demonstrates lower performance variability than unimodal approaches, indicating its robustness across diverse vendors. This consistency is critical for addressing real-world applications where vendor behaviors vary significantly.

Performance by Ad Frequency: The relationship between retrieval performance and the frequency of ads per vendor remains consistent across regions. The multimodal baseline achieves high performance across all ad frequencies, particularly excelling for vendors with lower ad frequencies. This suggests that multimodal integration effectively compensates for data sparsity by leveraging both textual and visual features. The M-Text representation follows closely, showing a significant improvement over the standalone text-only baseline, particularly as ad frequency increases. While the vision-only baseline struggles with sparse data, the M-Vision representation extracted from the multimodal model provides a noticeable improvement, albeit still trailing behind M-Text. These results reinforce the strength of multimodal baselines in handling scenarios with limited vendor representation.

Performance by Number of Names: As mentioned earlier, analyzing retrieval performance by the number of names associated with each vendor reveals the robustness of the multimodal baseline in linking ads with varied linguistic and visual patterns. Across all regions, the multimodal baseline maintains superior performance as the number of names increases, outperforming text-only and vision-only baselines. The M-Text representation consistently surpasses the standalone text-only baseline, demonstrating that multimodal training enhances the textual representation's robustness. While the vision-only baseline experiences noticeable drops in performance with increasing names, the M-Vision representation extracted from the multimodal model maintains steadier performance. These findings highlight the ability of multimodal baselines to capture stylistic and semantic

Retrieval	Metric	Midwest	West	Northeast
Text-to-Text	**MRR@10**	Shared: 0.7164 ± 0.41 Unique: 0.7910 ± 0.37	Shared: 0.8581 ± 0.33 Unique: 0.8498 ± 0.33	Shared: 0.7859 ± 0.38 Unique: 0.7013 ± 0.42
	R-Precision@X	Shared: 0.5027 ± 0.38 Unique: 0.6251 ± 0.37	Shared: 0.7128 ± 0.36 Unique: 0.7234 ± 0.36	Shared: 0.6553 ± 0.40 Unique: 0.5817 ± 0.44
Image-to-Image	**MRR@10**	Shared: 0.3462 ± 0.36 Unique: 0.3583 ± 0.38	Shared: 0.3506 ± 0.37 Unique: 0.3728 ± 0.37	Shared: 0.3031 ± 0.38 Unique: 0.2432 ± 0.32
	R-Precision@X	Shared: 0.0673 ± 0.09 Unique: 0.0914 ± 0.14	Shared: 0.0896 ± 0.12 Unique: 0.1168 ± 0.16	Shared: 0.0816 ± 0.13 Unique: 0.0807 ± 0.14
Multimodal	**MRR@10**	Shared: 0.7862 ± 0.36 Unique: 0.8355 ± 0.31	Shared: 0.8909 ± 0.28 Unique: 0.8693 ± 0.29	Shared: 0.8138 ± 0.35 Unique: 0.7920 ± 0.29
	R-Precision@X	Shared: 0.5026 ± 0.35 Unique: 0.6196 ± 0.34	Shared: 0.7103 ± 0.33 Unique: 0.7266 ± 0.33	Shared: 0.6436 ± 0.37 Unique: 0.5550 ± 0.41

Table 12: Text-to-Text, Image-to-Image, and multimodal retrieval performance for shared and unique vendors between South and Midwest, West, and Northeast region dataset. All the representations are extracted from the multimodal DeCLUTR-ViT backbone trained with CE+SupCon objective on the South region dataset.

variations better than unimodal baselines, which is crucial for identifying vendors with diverse aliases.

Performance by Images with and without Faces: The analysis of retrieval performance based on the presence or absence of faces in images provides critical insights into the multimodal baseline's ability to leverage facial features. Across all regions, the Multimodal-Face baseline consistently outperforms the Vision-Face baseline for both MRR@10 and R-Precision@X, demonstrating its effectiveness in combining facial and textual cues. For images with faces, the multimodal baseline initially struggles as faces increase but eventually outperforms the vision-only baseline when more visual data becomes available. This trend reflects the model's ability to adapt and utilize visual information effectively when sufficient samples are present. For images without faces, the Multimodal-Face baseline consistently surpasses the Vision-Face baseline, leveraging non-facial visual patterns and textual information to improve retrieval performance.

A.6.3 Multimodal Retrieval Performance on Shared and Unseen Vendors in OOD Datasets

Here, the evaluation focuses on a retrieval task, distinguishing between shared vendors—those present in the South and OOD datasets—and unknown vendors exclusive to the OOD datasets. While Figure 2a highlights an overlap of vendors between the South and OOD datasets, it is important to note that the OOD datasets were never exposed to the model during training. Table 12 presents a detailed analysis of the model's MRR@10 and R-Precision@X performance across text-to-text, image-to-image, and multimodal retrieval tasks for both shared and unseen vendors. Representations for these evaluations are derived from the multimodal DeCLUTR-ViT classifier trained on the South dataset. The results confirm the model's robust performance on shared and unseen vendors, showcasing its ability to generalize effectively to unseen scenarios. This demonstrates the model's capability to link ads to vendors, further underscoring its practical utility in real-world HT applications regardless of prior exposure to vendors and ads.

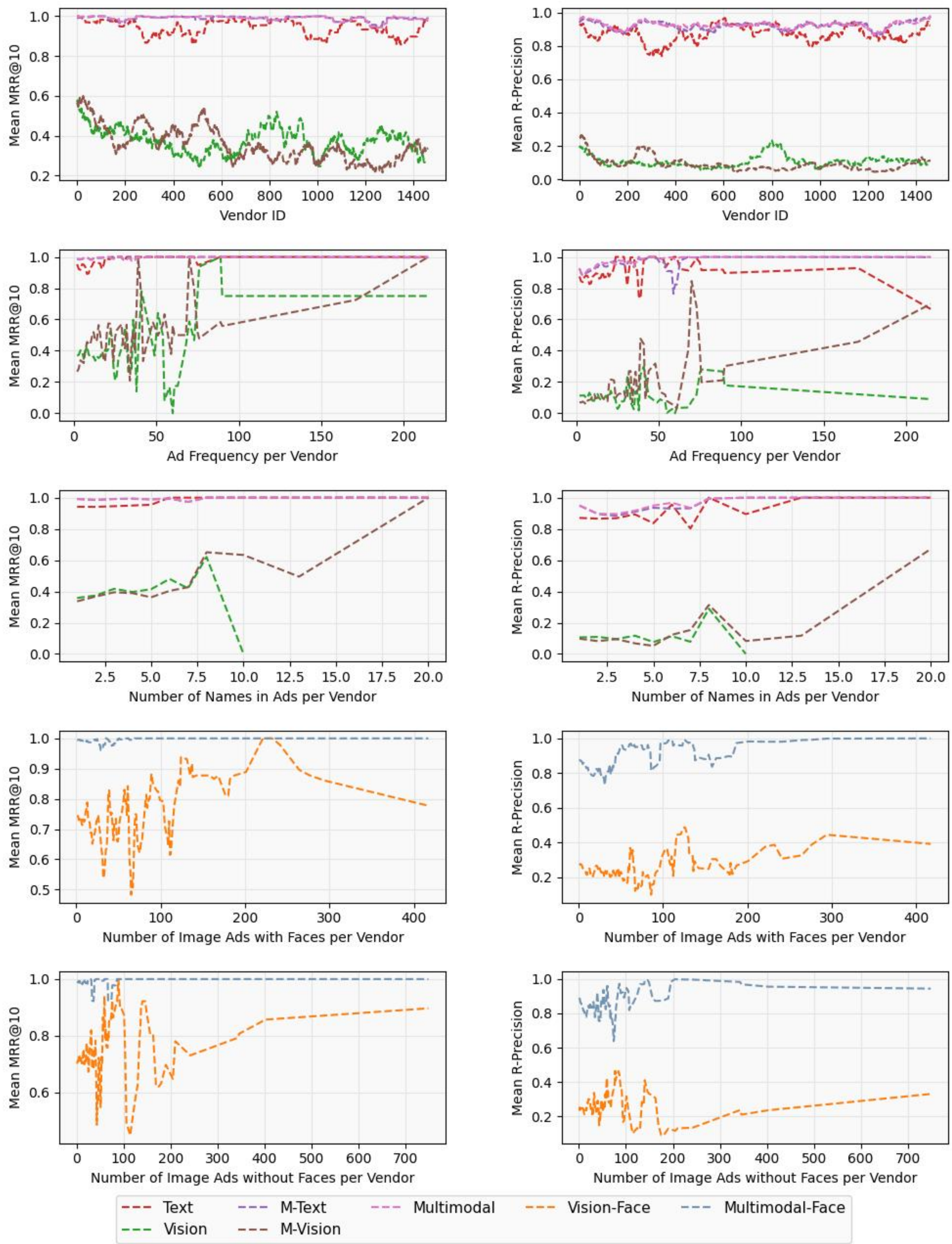


Figure 4: Comparison of retrieval performance on the South region test datasets. Text, vision, and multimodal baselines (DeCLUTR-small, ViT-base-patch16-224, and DeCLUTR-ViT, respectively) are trained end-to-end for vendor identification using the joint CE+SupCon objective on the South region dataset. M-Text and M-Vision represent text-only and image-only embeddings from the multimodal system. Vision-Face and Multimodal-Face denote evaluations of escort images with and without faces.

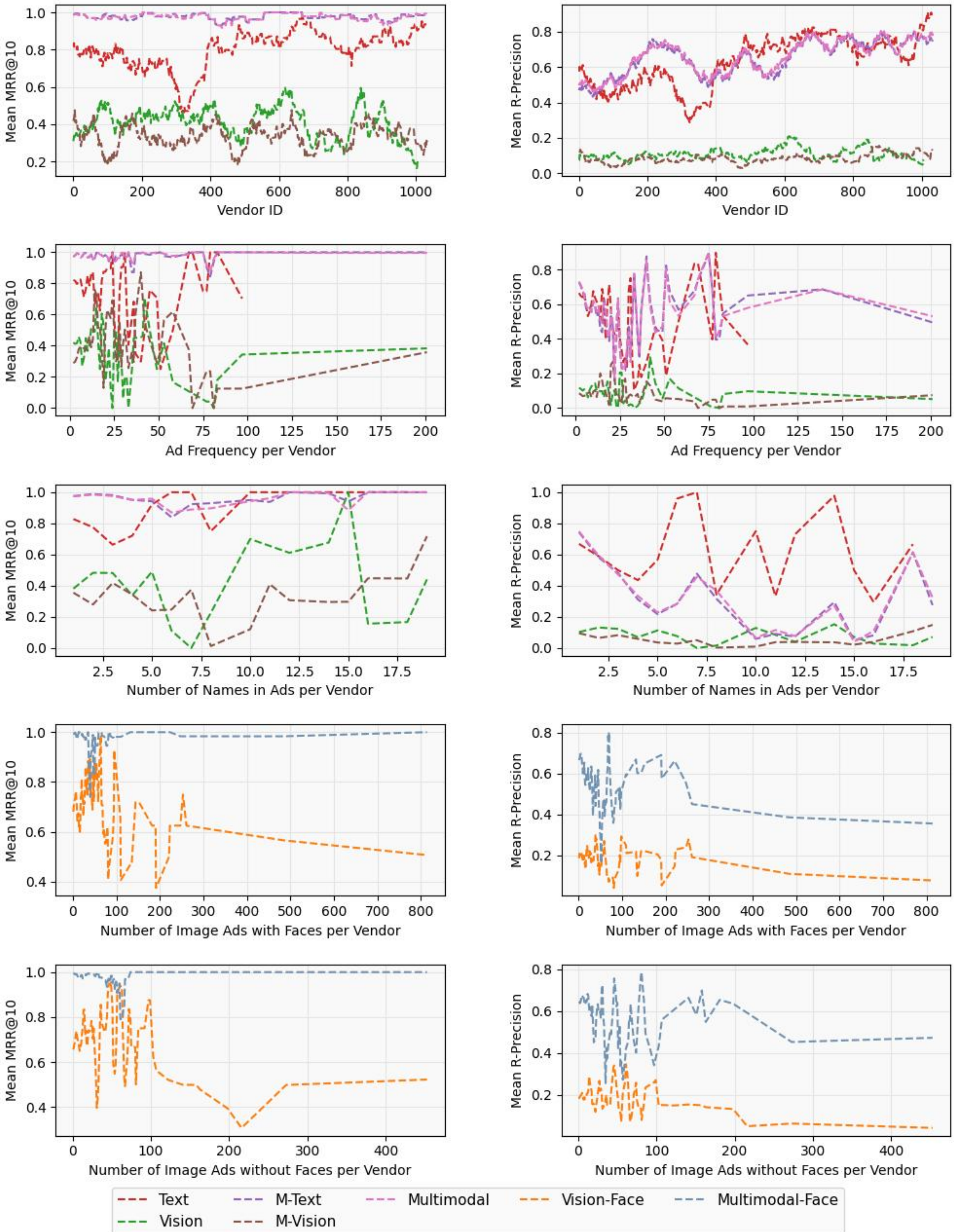


Figure 5: Comparison of retrieval performance on the Midwest region test datasets. Text, vision, and multimodal baselines (DeCLUTR-small, ViT-base-patch16-224, and DeCLUTR-ViT, respectively) are trained end-to-end for vendor identification using the joint CE+SupCon objective on the South region dataset. M-Text and M-Vision represent text-only and image-only embeddings from the multimodal system. Vision-Face and Multimodal-Face denote evaluations of escort images with and without faces.

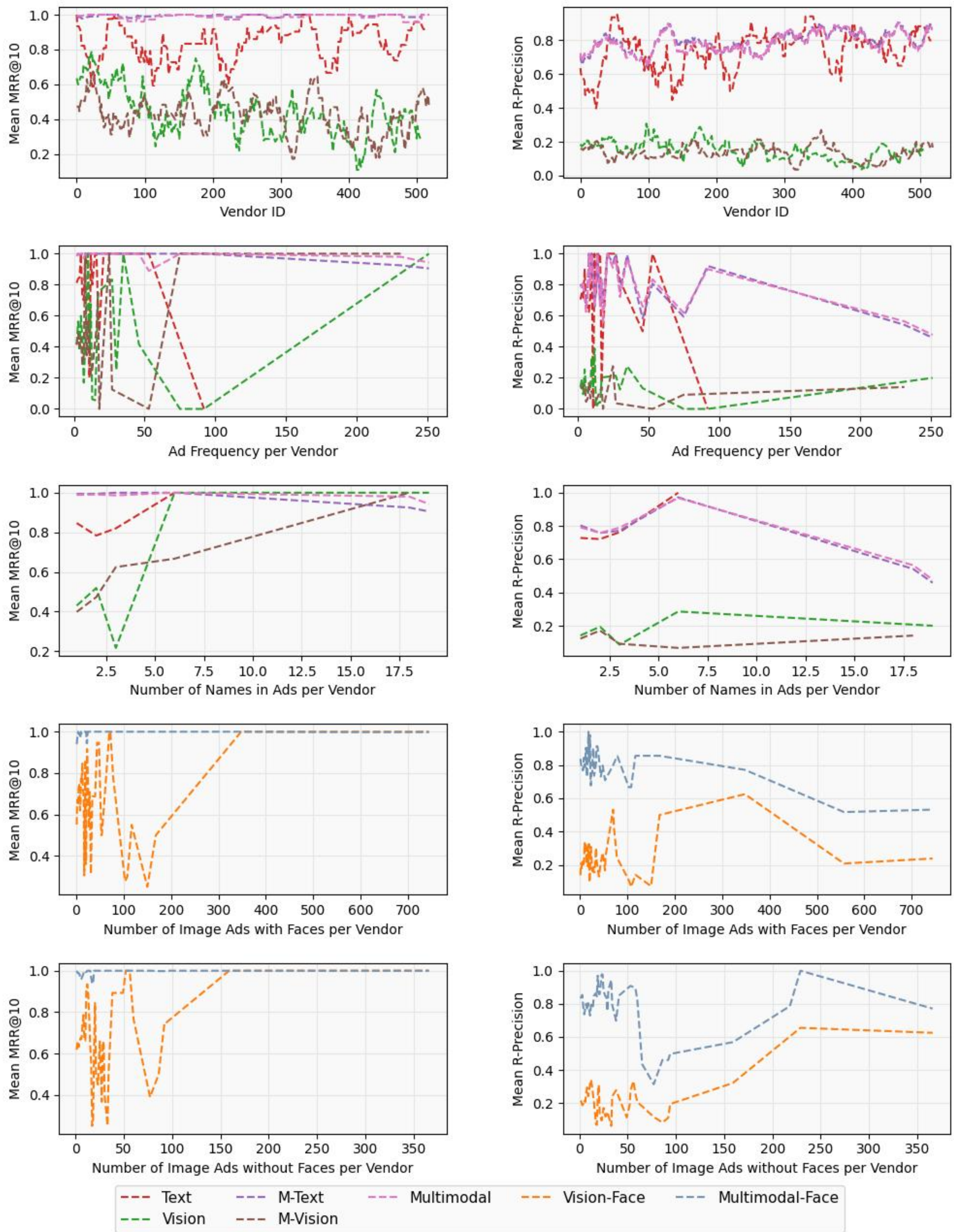


Figure 6: Comparison of retrieval performance on the West region test datasets. Text, vision, and multimodal baselines (DeCLUTR-small, ViT-base-patch16-224, and DeCLUTR-ViT, respectively) are trained end-to-end for vendor identification using the joint CE+SupCon objective on the South region dataset. M-Text and M-Vision represent text-only and image-only embeddings from the multimodal system. Vision-Face and Multimodal-Face denote evaluations of escort images with and without faces.

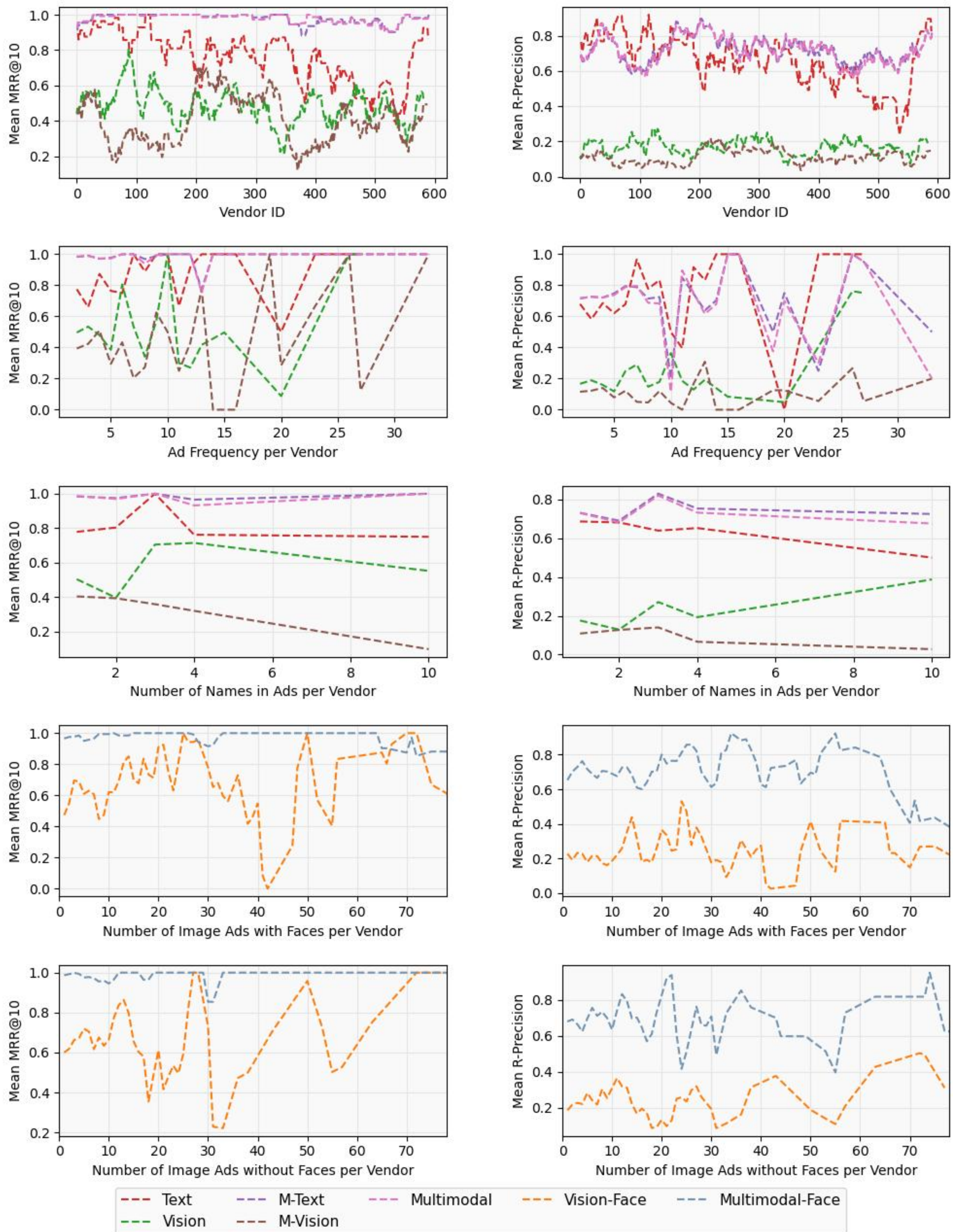


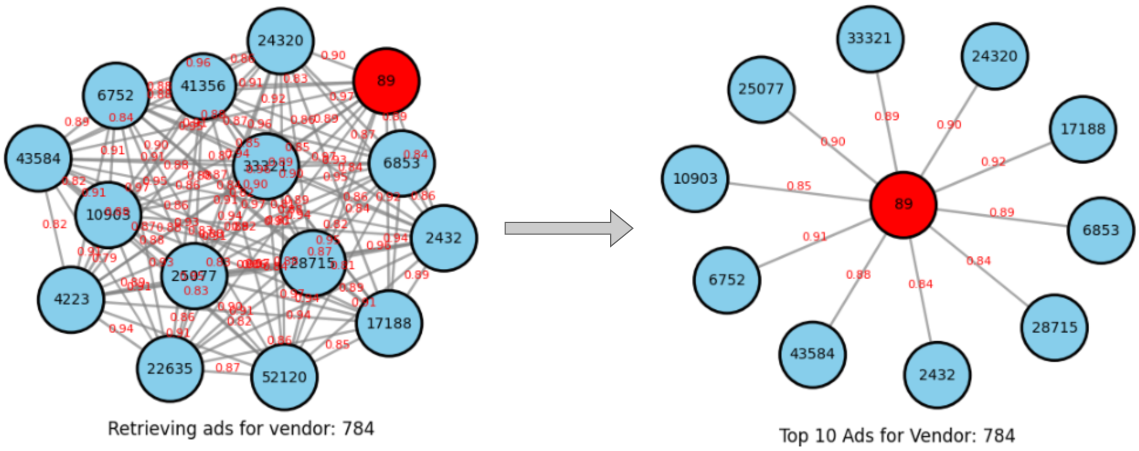
Figure 7: Comparison of retrieval performance on the Northeast region test datasets. Text, vision, and multimodal baselines (DeCLUTR-small, ViT-base-patch16-224, and DeCLUTR-ViT, respectively) are trained end-to-end for vendor identification using the joint CE+SupCon objective on the South region dataset. M-Text and M-Vision represent text-only and image-only embeddings from the multimodal system. Vision-Face and Multimodal-Face denote evaluations of escort images with and without faces.

A.7 Practical Utility

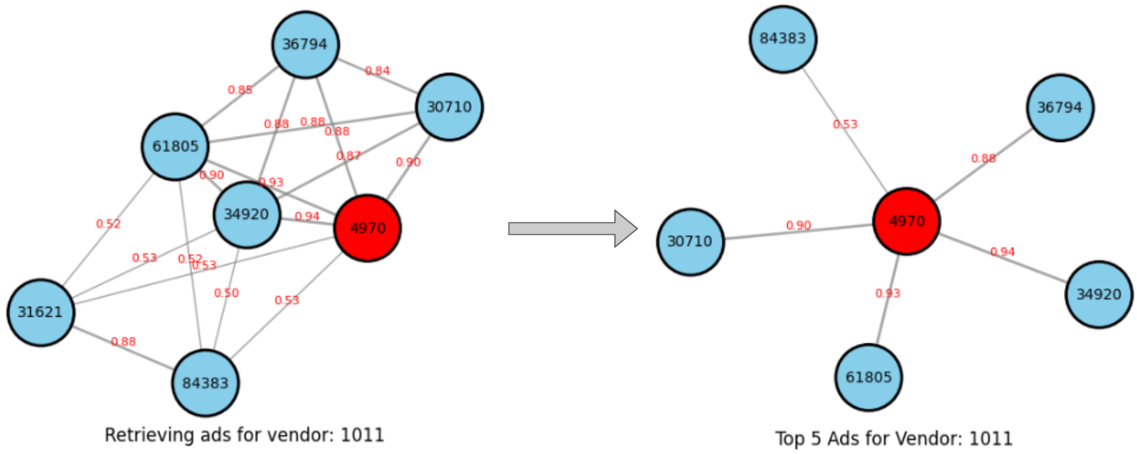
To demonstrate the practical utility of our research, we employ the multimodal DeCLUTR-ViT model, trained with the CE+SupCon objective on the South region dataset, to create knowledge graphs using retrieval-based methods. The choice of representations for constructing these graphs is informed by the retrieval performance of text, vision, and multimodal embeddings on R-Precision and MRR@10 metrics. Since text-only representations from the multimodal baseline exhibit superior retrieval performance across both metrics for our dataset, we utilize them to perform our retrieval analysis.

Figures 8a and 8b illustrate knowledge graphs generated for vendor labels 784 and 1101 from the South region datasets, respectively. To construct these graphs, we begin with a query advertisement (highlighted in red) and retrieve all relevant ads from the training dataset based on R-Precision performance. Each advertisement is represented as a node in the graph and labeled with its unique ID. Notably, these IDs serve as anonymous identifiers, as all personally identifiable information in the dataset has been removed using comprehensive masking techniques. Edges in the graph encode the similarity scores between connected nodes and the query advertisement, providing a quantifiable measure of relatedness. The graphs on the left of the figures depict all retrieved ads for a given query, visualizing the comprehensive network of connected advertisements for a specific vendor. To provide flexibility for researchers, investigators, and law enforcement agencies (LEAs), we propose an alternative approach using MRR@K. This allows stakeholders to retrieve the top-K most relevant ads based on similarity, enabling focused analysis depending on investigative confidence or manual verification thresholds. The resulting knowledge graphs, visualized on the right side of the figures, present a filtered view, facilitating efficient examination of high-confidence matches.

By leveraging these knowledge graphs, stakeholders can visualize vendor activity across advertisements, identify patterns, and establish connections, using it to initiate investigations into identifying HT identifiers.



(a) Vendor 784



(b) Vendor 1101

Figure 8: Knowledge graph representation generated using AA retrieval for Vendor labels 784 and 1101 from the South region dataset. The left graph utilizes R-Precision metrics to link all relevant ads for a query ad (highlighted in red), while the right graph applies (a) MRR@10 and (b) MRR@5 to identify the top-10 most likely relevant ads. Nodes represent advertisement IDs, and edges denote the similarity between ads, both in relation to each other and the query ad, showcasing the effectiveness of AA retrieval in constructing relational insights.