AGATE: Stealthy Black-box Watermarking for Multimodal Model Copyright Protection

Jianbo Gao
Beijing Institute of Technology
China
3120235247@bit.edu.cn

Keke Gai
Beijing Institute of Technology
China
gaikeke@bit.edu.cn

Jing Yu
School of Information Engineering,
Minzu University of China.
China
jing.yu@muc.edu.cn

Liehuang Zhu
Beijing Institute of Technology
China
liehuangz@bit.edu.cn

Qi Wu
University of Adelaide, Adelaide
Australia
qi.wu01@adelaide.edu.au

Abstract

Recent advancement in large-scale Artificial Intelligence (AI) models offering multimodal services have become foundational in AI systems, making them prime targets for model theft. Existing methods select Out-of-Distribution (OoD) data as backdoor watermarks and retrain the original model for copyright protection. However, existing methods are susceptible to malicious detection and forgery by adversaries, resulting in watermark evasion. In this work, we propose Model-agnostic Black-box Backdoor Watermarking Framework (AGATE) to address stealthiness and robustness challenges in multimodal model copyright protection. Specifically, we propose an adversarial trigger generation method to generate stealthy adversarial triggers from ordinary dataset, providing visual fidelity while inducing semantic shifts. To alleviate the issue of anomaly detection among model outputs, we propose a post-transform module to correct the model output by narrowing the distance between adversarial trigger image embedding and text embedding. Subsequently, a two-phase watermark verification is proposed to judge whether the current model infringes by comparing the two results with and without the transform module. Consequently, we consistently outperform state-of-the-art methods across five datasets in the downstream tasks of multimodal image-text retrieval and image classification. Additionally, we validated the robustness of AGATE under two adversarial attack scenarios. Code is available at https://anonymous.4open.science/r/AGATE-7423.

CCS Concepts

• **Computing methodologies** → *Computer vision*.

Figure 1: Backdoor-based watermarking scheme comparison.

Keywords

Black-box Watermarking, Model Copyright Protection, Watermarking Security

1 Introduction

Multimodal models have revolutionized Artificial Intelligence by enabling cross-modal semantic alignment, e.g., Contrastive Language-Image Pretraining (CLIP) [20]. The advancement in multimodal

models underpin critical applications ranging from automated content moderation to AI-assisted solutions, making AI large models high-value Intellectual Property (IP) assets [23, 28, 30, 31, 34, 35]. However, widespread adoptions also attract malicious actors seeking to steal and redistribute proprietary models through various means, such as model extraction [26] or parameter replication [9]. Such theft not only undermines economic incentives for innovation, but also raises ethical risks, as stolen models may generate disinformation or bypass safety filters [41].

Model watermarking is deemed to be a potential alternative for protecting models' copyrights. Black-box watermarking [13, 38] enables copyright verification without access to the model, but there still exist limitations in non-trigger-based schemes [38] due to multiple causes, such as model extraction attacks [33, 37]. Trigger-based backdoor methods [6, 12, 15, 25] directly embed watermarks into model behaviors without the knowledge of model architectures or parameters for verification. However, most existing solutions rely on Out-of-Distribution (OoD) triggers [6, 15] that differ from the model's training data distribution, e.g, artificially created data [6] or irrelevant substitute data [15]. Existing trigger-based methods generally encounter two obvious issues as shown in Figure 1. For the trigger selection process, existing methods select OoD triggers that frequently exhibit statistical anomalies, e.g., model generates aligned image-text embeddings for mismatched image-text pairs, so that adversaries can identify and evade trggers through automated input sanitization [16, 32]. Moreover, OoD trigger set necessitates meticulous data selection from external sources, which is time and labor consuming. For the trigger injection and verification process, existing approaches embed triggers into models through fine-tuning strategies, which inevitably compromises the models' performance on benign data [1]. Furthermore, adversaries can exploit model fine-tuning to remove embedded backdoor watermarks [7]. Therefore, there is a critical need to investigate methods that preserve model utility while simultaneously enhancing the stealthiness of trigger selection and verification.

To address the above issues, we propose a Model-agnostic Black-box Backdoor Watermarking Framework (AGATE) that embeds copyright signatures through in-distribution adversarial triggers and verifies ownership via a two-phase cooperative mechanism. AGATE uses adversarial noises as a versatile instrument, subtly perturbing randomly chosen training samples to create triggers that are statistically indistinguishable from clean data, inducing verifiable behavioral deviations in unauthorized models. Dissimilar to conventional OoD triggers, AGATE's perturbations maintain the original data distribution, thereby avoiding adversaries identifying triggers via abnormal analysis, i.e., abnormal relation between input image-text semantics and their output embedding distance. Moreover, AGATE reduces costs of model fine-tuning, addressing the stealthiness and performance degradation issues. Specifically, we propose a two-phase cooperative watermark verification mechanism. First, adversarial triggers exacerbate semantic discrepancies between predictions of pirated models and legitimate outputs, causing identifiable anomalies. Second, we employ a lightweight transform module to rectify deviations by training on the original model's embedding space. The module serves as a semantic corrector, restoring the expected behavior exclusively when applied

to models originating from the watermarked source. By linking anomaly induction to correction capability, the two-phase design establishes an unassailable causal connection between the watermark and model provenance. Since the adversarial triggers are obtained by noise injection in original images while the multimodal model is free from fine-tuning, adversaries are difficult to identify triggers as the original image-text pairs in the dataset have indistinguishable input-output behavior on undisturbed multimodal model.

The main contributions are summarized as follows: (1) We propose a black-box backdoor watermarking framework for multimodal foundation models for the first time, which harmonizes imperceptible in-distribution triggers with a two-phase cooperative verification mechanism. The framework shifts the paradigm from manually engineered OoD artifacts to data-native adversarial perturbations, enabling stealthy watermark embedding without compromising model utility. (2) We propose a perturbation-correction cooperative verification mechanism, i.e., adversarial noise simultaneously disrupts unauthorized model behaviors and enables provable authentication through feature-space rectification. In this way, adversaries, even deceiving the first watermark verification phase if the noise injection approach leaked, can not correctly pass the second verification phase without the knowledge of transform module. (3) Extensive experiments demonstrate that our framework achieves superior performance compared to state-of-the-art approaches on five downstream datasets, ranging from +0.1% to +2.6%. Moreover, our AGATE shows strong robustness against two representative adversarial attacks according to different knowledge of adversaries.

2 Related Work

White-Box Watermarking Early research on model watermarking focused on white-box scenarios by embedding ownership signals into model parameters or activation patterns. Uchida et al. [29] tried to insert watermarks via convolutional layer weight quantization, followed by extensions exploiting attention maps [24] and batch normalization layers [3]. These methods relied on access to model internals, making them impractical for commercial black-box APIs. Adversaries can remove watermarks via parameter fine-tuning [7] or pruning [36] to reduce detection accuracy. Moreover, modifying parameters in multimodal models (e.g., CLIP) could result in disrupting cross-modal alignment and degrading downstream task performance [33, 37]. Differing from parametric dependencies, our AGATE embeds watermarks through input-output behavior mapping to eliminate reliance on model internals. The adversarial trigger inherently preserves cross-modal consistency.

Black-Box Non-Trigger Watermarking. Non-trigger black-box methods authenticated models via statistical fingerprints [2, 18], e.g., API query distributions [4] or adversarial response patterns [27]. However, adversaries evaded detection via post-processed watermarked image [10], and even minor perturbations to query frequencies reduced the accuracy of verification. In multimodal settings, attackers could bypass detections by targeting a single modality to exploit decoupled feature spaces [39]. Rather than relying on statistical correlations, we establish causal ownership evidence through adversarial triggers. Our two-phase verification protocol creates an unforgeable link between triggers and model provenance to make input/output manipulation attacks ineffective.

Black-Box Trigger-based Watermarking. Prior studies have explored trigger-based watermarking by embedding ownership signals via poisoned samples [8] or adversarial perturbations [40] to strengthen robustness; however, three limitations still exist. First, manually crafted triggers (e.g., OoD data [6]) exhibited detectable statistical anomalies [16, 32] that are easily detected and removed by adversaries. Second, retraining models on hybrid datasets incurred prohibitive costs and degraded clean-data performance [1]. Finally, modality-specific triggers disrupted cross-modal alignment so that text perturbations in corpus reduce model retrieval accuracy [17]. Differently, AGATE addresses issues above through in-distribution adversarial triggers and retraining-free embedding.

3 Methodology

3.1 Threat Model

We define the objective, knowledge, and capability of adversaries.

Adversaries' Objective. Adversaries aim to steal the multimodal models from the original owner and falsely claim legitimacy of the copyright ownership. Adversaries seek economic gains by selling or publishing stolen models.

Adversaries' Knowledge. We assume that adversaries lack knowledge of the target model's trigger generation strategy and watermark verification process, despite having replicated datasets to interact with the original model service.

Adversaries' Capability. Adversaries possess the capability to detect specific backdoor trigger sets (e.g., conducting statistical analyses on the original model's abnormal query responses) and to fabricate false triggers for bypassing backdoor-based verifications.

3.2 Framework Overview

AGATE is a black-box backdoor watermarking framework for multimodal model copyright protection. As illustrated in Figure 2, the framework comprises three main components, namely, adversarial trigger generation, transform module training, and two-phase watermark verification. Specifically, AGATE samples basic triggers from the original dataset, injecting adversarial noise to create model-specific backdoor triggers. The customized triggers train a post-hoc transform module that minimizes the embedding space distance between adversarial triggers' visual embeddings and textual embeddings, while preserving the original functionality of basic triggers. Ownership verification is accomplished by comparing the model's outputs when the transformation module is applied versus when it is not. This two-phase watermark verification process judges unauthorized model copies by identifying inherited backdoor response patterns and verifying transformation consistency. AGATE effectively decouples trigger generation from watermark validation while maintaining detection robustness.

3.3 Adversarial Trigger Generation

Our adversarial trigger generation addresses two limitations in existing backdoor watermarking techniques. (1) **High deployment costs** persist due to computational overhead and an increase in model complexity, stemming from the reliance on manually selected OoD samples for trigger constructions. (2) **Low stealthiness** weakens the integrity of watermarks as a result of adversarial detections caused by a fixed-pattern trigger.

We use an adversarial semantic perturbation method to generate a dynamic and stealthy trigger set, which consists of three major components: basic trigger randomization, adversarial semantic perturbation, and dynamic trigger number.

Basic trigger randomization. AGATE establishes a randomized sampling paradigm to address issues of over-reliance on OoD data and high construction costs in trigger generation. We construct basic triggers T_b by randomly selecting in-distribution image-text pairs $\{x, y\}$ from the ordinary dataset D. Our scheme utilizes three key properties of multimodal corpora. (1) Combinatorial randomness from free dataset selection and pair permutation exponentially expands the viable trigger space. (2) Native semantic coherence ensures trigger stealthiness through natural feature alignment. (3) Linear sampling complexity $O(1)$ eliminates manual OoD creation overhead. The emergent trigger space dimensionality satisfies Equation (1), where k denotes sampling iterations and $n^{(i)}$ represents modality-specific feature dimensions per sample. Our scheme creates super-exponential attack surface growth to impede adversaries.

$$\dim(T_b) = \binom{|D|}{k} \times \prod_{i=1}^{k} (n_{img}^{(i)} \times n_{text}^{(i)}) \tag{1}$$

Adversarial semantic perturbation. To improve adversarial detection resistance, we implement a semantic-perturbation trigger mechanism [39] that synthesizes model-specific adversarial image trigger ($\widetilde{x}$). Specifically, we sample latent vector z from a parametric noise distribution, from which generates perturbation patches $G(z)$ via adversarial generator G. Then, we fuse perturbations with basic image trigger x through controlled blending. Thus, the adversarial trigger set T_a is obtained. Equation (2) defines the operation of adding adversarial perturbation to the trigger, where $\odot$ denotes element-wise product function, and m is a positional mask matrix.

$$T_a = \{\widetilde{x}, y\} = \{(1 - m) \odot x + m \odot G(z), y\} \tag{2}$$

We use this adversarial synthesis to achieve three critical effects: (1) Model-specific dependency through generator conditioning the original model O embedding space $E(\cdot)$. (2) Visual coherence preservation via ℓ_2-norm constraints $\|\widetilde{x} - x\|_2 \leq \epsilon_1$. (3) Semantic deviation amplification measured by cosine distance or Euclidean distance $D(E_t(y)\|E_v(\widetilde{x})) \geq \delta$ between perturbed visual embedding and basic text embedding.

Dynamic trigger number. We set dynamic trigger scaling to enhance operational adaptability and adversary resistance across diverse application scenarios. This mechanism dynamically adjusts the deployment scale of trigger sets according to real-time security demands. Crucially, AGATE ensures functional isolation, where trigger quantity modifications exclusively affect the transform module's training dynamics while preserving the original model's functionality invariant.

3.4 Transform Module Training

We aim to solve the fundamental stealthiness-effectiveness trade-off dilemma in backdoor watermarking through adaptive output rectification. Existing schemes necessitate divergent outputs between triggers and normal samples to establish copyright evidence, creating detectable artifacts that adversaries exploit via inversion attacks on query response patterns. This vulnerability stems from the inherent correlation between output deviations and trigger exposure

Figure 2: Framework overview of the proposed AGATE.

risk. Our approach trains a post-hoc transform module (M) that enforces output consistency: $\forall (T_a, T_b) \in T, M(O(T_a)) = M(O(T_b))$.
Transform Module Architecture. The transform module (M) has a dual-function mechanism, i.e., obfuscating attack surfaces by decoupling observable outputs from embedded watermarks and maintaining verification capability via transform module result differential comparison. In addition, the module (M) is trained by using paired samples containing basic triggers $\{T_b^{(i)}\}$ and derived adversarial triggers $\{T_a^{(i)}\}$, which forms the training tuple $D_{train} = \{\{(x^{(i)}, y^{(i)}, \tilde{x}^{(i)})\}_{i=1}^{N}$. Positioned as a post-model processing component, M learns embedding space alignment through O' visual and text encoder.

Moreover, the module (M) employs a lightweight multi-layer perceptron architecture comprising three layers, including input, single hidden, and output layers, to enhance computational efficiency. The module maintains dimensionality similar to O so that both input and output dimensions strictly match O's encoder. The dimensional consistency creates indistinguishability between O and M's output spaces to prevent detections from architectural analysis by uninformed adversaries.
Training Loss. Multimodal contrastive alignment mechanism employs triplet relationship constraints in the joint embedding space through two collaborative objectives: (1) semantic alignment between adversarial visual embeddings $E_v(\tilde{x})$ and text embeddings $E_t(y)$, and (2) preservation of intrinsic correlations between base image embeddings $E_v(x)$ and their text embeddings $E_t(y)$. Thus, given a training set (D_{train}), contrastive loss is defined by Equation (3), where $f(\cdot)$ and $g(\cdot)$ denote learnable projection heads,

$d(u, v) = 1 - \cos(u, v)$ measures cosine dissimilarity, and hyperparameters λ and η balance the dual objectives.

$$\mathcal{L} = \sum_{i=1}^{N} \left(\underbrace{d(f(\tilde{x}^{(i)}), g(y^{(i)}))}_{\text{Adversarial Alignment}} + \lambda \cdot \underbrace{\max\left(0, d(f(x^{(i)}), g(y^{(i)})) - \eta\right)}_{\text{Intrinsic Preservation}} \right)$$
(3)

The adversarial alignment part forces $E_v(\tilde{x})$ to converge toward $E_t(y)$ in the transform module embedding space, while the preservation part maintains a minimum correlation threshold η between $E_v(x)$ and $E_t(y)$ through hinge loss regularization. Dual-constrained optimization achieves ϵ_2-alignment ($\|f(\tilde{x}) - g(y)\|_2 \leq \epsilon_2$) with provable convergence of projection heads.

3.5 Two-Phase Watermark Verification

AGATE resolves the vulnerability of attack surface exposure inherent in conventional backdoor watermarking systems, where compromised trigger sets enable adversarial circumvention of verification protocols. The proposed two-phase watermark verification mechanism is a hierarchical defense against trigger leakage threats. (1) Trigger-transform binding: Watermark verification requires simultaneous possession of both adversarial triggers T_a and the proprietary transform component M. (2) Phase-decoupled detection logic: Trigger response pattern matching using original model outputs in Phase I, and transform consistency validation through M-processed outputs. This dual requirement mechanism makes it impossible for adversaries to bypass verification even if the trigger leaks, as expressed in Equation (4) where λ denotes a security parameter, $\text{negl}(\cdot)$ is a negligible function, $\text{Verify}(\cdot)$ is a

verification function.

$$\Pr[\text{Verify}(T_a^{'}) = 1 | T_a^{'} \notin M] \leq \text{negl}(\lambda) \tag{4}$$

Two-phase verification mechanism operates on differential output analysis between processing paths. Normal samples maintain output consistency across both phases, while adversarial triggers exhibit phase-dependent divergence. Specifically, adversarial triggers produce anomalous responses differing from basic behaviors in Phase I, and transform-processed triggers restore normal responses matching basic texts in Phase II. In the case adversaries subvert verification by removing embedded triggers in Phase I, even though trigger-induced anomalies are detected. Such tampering causes the same outputs from two phases, which is against two-phase result differential requirement. It means stolen models fail to verification so that the illegitimate ownership claims are eliminated.

For any adversarial image trigger ($\widetilde{x} \in T$) input to a suspicious model (S), we obtain the result $S(\widetilde{x})$ (Result #1) in Phase I by computing semantic differences. The semantic discrepancy $||E_v(\widetilde{x}) - E_t(y)||_{H_S} \geq \sigma (\sigma > 0)$ induces erroneous retrieval outputs deviating from baseline, where $|| \cdot ||_{H_S}$ denotes the distance in the suspicious model's embedding space. Then, module M enforces output correction, resulting in $M(S(\widetilde{x}))$ (Result #2) in Phase II according to $||M(E_v(\widetilde{x})) - M(E_t(y))||_{H_M} < \tau$, where $|| \cdot ||_{H_M}$ denotes the distance in the transform's embedding space. Meanwhile, we ensure normal sample preservation to guarantee operational transparency for legitimate inputs. Finally, we conduct comparative judgment as shown in Equation (5) and the final verification result $Result$ is computed as shown in Equation (6), where $True$ and $False$ indicate whether there is infringement or not.

$$\text{Verify}(\widetilde{x}) = \begin{cases} 1 & \text{if } M(S(\widetilde{x})) = S(x) \\ 0 & \text{if } M(S(\widetilde{x})) \neq S(x) \end{cases} \tag{5}$$

$$Result = \begin{cases} True & \text{if Verify}(\widetilde{x}) = 1 \\ False & \text{if Verify}(\widetilde{x}) = 0 \end{cases} \tag{6}$$

In addition, the transform module is jointly trained using the original multimodal model encoder and independently selected trigger sets, which are unique to each original model. When replacing with the transform module of another model, this transform module cannot shorten the embedding distance in the transform embedding space and correct the output Result #2. Therefore, there will be no misjudgment of different models by the same trigger. In other words, our framework enables different versions of multimodal models designed based on the same model architecture to be uniquely determined by their unique trigger sets and transform models, which is highly universal and model-agnostic.

4 Experiments

4.1 Experiment Setup

Dateset. Performance evaluations were implemented on two representative multimodal image-text retrieval datasets (MS-COCO [14] and Flick30k [19]) and three object classification datasets (CIFAR-10 [11], CIFAR-100 [11], and VOC2007 [5]). We used Wikipedia [21] and Pascal-Sentences [22] datasets to display trigger selection.

Evaluation metrics. In image-text retrieval tasks, we adopted Recall@K (R@K) to measure the retrieval performance of text retrieval

Table 1: Performance comparison with different baselines on the MS-COCO, Flicker30k, CIFAR-10, CIFAR-100, and VOC2007. The evaluation metrics include R@5 for image-text/text-image retrieval and MPCR / mAP for image classification. △ (△ = {Method} − Origin) **represents the performance degradation compared to the original model.**

Method	Dataset	Metric	Result (%)	△ (%)
Origin	MS-COCO	R@5	58.40/76.72	0.0/0.0
	Flicker30k	R@5	85.58/96.20	0.0/0.0
	CIFAR-10	mAP	82.92	0.0
	CIFAR-100	MPCR	96.60	0.0
	VOC2007	MPCR	66.95	0.0
EmbM	MS-COCO	R@5	47.90/65.30	-10.5/-11.42
	Flicker30k	R@5	84.80/66.20	-0.78/-30.00
	CIFAR-10	mAP	80.50	-2.42
	CIFAR-100	MPCR	77.90	-18.70
	VOC2007	MPCR	66.80	-0.15
MFLO	MS-COCO	R@5	57.35/76.62	-1.05/**-0.10**
	Flicker30k	R@5	84.60/93.86	-0.98/-2.34
	CIFAR-10	mAP	74.88	-8.04
	CIFAR-100	MPCR	90.41	-6.19
	VOC2007	MPCR	66.65	-0.30
Ours	MS-COCO	R@5	58.20/76.44	**-0.20**/-0.28
	Flicker30k	R@5	85.26/95.99	**-0.32**/**-0.21**
	CIFAR-10	mAP	82.60	**-0.32**
	CIFAR-100	MPCR	90.90	**-5.70**
	VOC2007	MPCR	66.85	**-0.10**

with image queries and image retrieval with text queries. In classification tasks, we used the Mean Per Class Recall (MPCR) for image tasks and the mean Average Precision (mAP) for multi-label tasks. To demonstrate the effectiveness of the trigger, we used the cosine distance $D(cos)$ and the Euclidean distance $D(euc)$ to approximate the similarity between image embedding and text embedding in the embedding space.

Implementation Details. We chose CLIP [20], a representative model series of multimodal models, as the original model, including ViT-B-16-quickgelu (OpenAI), ViT-B-16 (laion400m_e32), and ViT-B-32 (OpenAI). We randomly selected a basic trigger from the original dataset for enhancing dynamic variability. Next, we added noise to basic triggers to obtain implicit triggers. We finally inputted implicit triggers into the CLIP model to obtain the corresponding textual triggers for each implicit adversarial trigger. Our evaluations utilized Adam to train the transform module with a learning rate of 1×10^{-3} for 1000 epochs on a single RTX 4090 GPU.

Baselines. We adopted existing multimodal backdoor watermarking methods as benchmarks: (1) **EmbMarker** [17] (EmbM) selected a set of mid-frequency words from a general text corpus to form a trigger word collection and chose one target embedding as the watermark embedded into the model. (2) **MFL-Owner** [6] (MFLO) selected a group of images from OoD data and used LLM to generate texts not related to the images to form a trigger set together.

Table 2: Performance comparison of different noise types and addition strategies (Add) for generating adversarial triggers. Metrics RMSE, PSNR, SSIM, and UQI for visual similarity, and $D(cos)$ for semantic divergence

Noise	Add	RMSE ↓	PSNR ↑	SSIM ↑	UQI ↑	$D(cos)$ ↓
GN	GNA	24.03	20.52	0.78	0.77	26.36
	LNA	10.83	27.44	0.94	0.94	26.27
	BON	7.23	30.95	0.97	0.97	25.26
	SPN	12.41	26.25	0.93	0.62	26.27
	CANA	24.01	20.53	0.78	0.77	26.44
PN	GNA	3.66	36.87	**0.99**	**0.99**	25.10
	LNA	**1.88**	**42.63**	**0.99**	**0.99**	26.00
	BON	1.89	**42.59**	**0.99**	**0.99**	25.67
	SPN	3.68	36.80	**0.99**	**0.99**	25.43
	CANA	10.52	27.69	0.95	0.95	25.94
SPN	GNA	54.20	13.45	0.42	0.41	24.59
	LNA	24.08	20.50	0.77	0.77	25.43
	BON	27.01	19.50	0.73	0.72	25.40
	SPN	32.82	17.81	0.66	0.65	24.79
	CANA	18.49	22.79	0.85	0.85	25.56
MN	GNA	93.81	8.69	0.35	0.35	26.41
	LNA	41.01	15.87	0.65	0.64	25.73
	BON	29.96	18.60	0.80	0.79	25.59
	SPN	169.17	3.56	0.11	0.10	25.70
	CANA	169.01	3.57	0.11	0.10	25.90
Adv	GAN	8.73	29.31	0.96	0.95	**24.02**

Gaussian Noise (GN), Poisson Noise (PN), Salt-and-Pepper Noise (SPN), Multiplicative Noise (MN), Adversarial Noise (Adv); Global Noise Addition (GNA), Local Noise Addition (LNA), Blended Original Noise (BON), Spatially Variant Noise (SPN), Content-Aware Noise Addition (CANA), Generative Adversarial Network (GAN); Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Universal Quality Index (UQI), and Cosine Distance ($D(cos)$).

4.2 State-of-the-art Comparison

Our evaluations compared AGATE with a few baselines to investigate the impact on the downstream image-text retrieval and image classification tasks. Table 1 depicted that AGATE achieved the closest performance to the original model across datasets. For instance, on Flicker30k text/image retrieval, AGATE only suffered a minor degradation of 0.21% (R@5=95.99%), leading to performance retention rate of 99.78%, while EmbMaker and MFL-Owner exhibited significant drops of 30.00% and 2.34%, respectively. Similarly, for image classification tasks, AGATE maintained a high mAP of 82.60% with merely a 0.32% performance gap compared to the original model on CIFAR-10, whereas MFL-Owner and EmbMaker showed larger gaps of 8.04% and 2.42%. We analyzed that additional OoD triggers trained by existing baselines caused interference to the performance of the original model. AGATE's triggers were selected from the original dataset, so that intrinsic preservation is achieved during the training. Table 1 indicated that AGATE had fewer performance degradations between data sets. Thus, the results depicted that AGATE had reliability and effectiveness in multimodal task application scenarios while ensuring model copyright protection.

4.3 Impact of Trigger Generation Strategies

We evaluated the impact of different noise types and addition strategies on trigger generation. Evaluations set up different experimental

groups by combining various noise types and addition strategies for generating adversarial triggers (see Table 2). Triggers generated from different groups had high stealthiness while ensuring that the resulting noisy images exhibited minimal perceptual differences from the original images and maintained a low semantic similarity with the original textual descriptions. We use $D(cos)$ of noisy image and text embedding to quantify the semantic divergence.

Table 2 depicted that poisson noise emerged as the most effective in achieving high perceptual fidelity. The LNA and BON strategies yielded exceptionally low RMSE values (1.88 and 1.89) and the highest PSNR values (42.63 and 42.59 dB), coupled with near-optimal SSIM and UQI scores (both 0.99). In comparison, triggers generated by other traditional noise types and addition strategies had little difference in maintaining semantic similarity, but their visual similarity is far inferior to PN. However, $D(cos)$ for PN remained within the 25.10–26.00 range, indicating slight semantic disruption.

Moreover, adversarial noise generated by a GAN outperformed other schemes for achieving a great trade-off between visual similarity and adversarial effectiveness. The GAN-based strategy achieved an RMSE of 8.73 and a PSNR of 29.31 dB while maintaining a high SSIM of 0.96, and it achieved the lowest $D(cos)$ value of 24.02, indicating the strongest semantic divergence among all tested strategies. We observed that the noise generated by GAN was optimized to utilize the correlations of different modalities while preserving visual coherence. Table 2 highlighted the potential of GAN-based adversarial triggers for effectively deceiving multimodal models while maintaining high stealthiness.

4.4 Impact of the Transform Module

To evaluate the effectiveness of the transform module in correcting adversarial triggers while preserving the original model behavior for basic triggers, we used two CLIP models to generate triggers, i.e., ViT-B-16-quickgelu (OpenAI) and ViT-B-16 (laion400m_e32), correspondingly M_A and M_B. Table 3 showed that our transform module had advantages in following aspects.

Adversarial Correction. For adversarial trigger inputs, transform module T_A, which trained on the same CLIP model M_A, successfully converted misclassified samples (*Res1=False*) to correct predictions (*Res#2=True*) with dramatic Δ_{cos} improvement of +82.67. It indicated adversarial trigger has been successfully corrected by narrowing the distance between the image and text embeddings.

Benign Preservation. Both modules T_A and T_B maintain correct classifications for basic triggers while enhancing feature distinctiveness. Both increases in $D(cos)$ and reduction in $D(euc)$ showed that transform modules effectively adjusted basic triggers' embeddings, moving them closer to the embeddings of other normal samples, while maintaining their classification results.

Model dependency. Module T_B showed limited efficacy, failing to correct adversarial triggers' classifications result, demonstrating that only transform modules related to the anterior model were effective. T_A was trained by M_A to learn more suitable feature distributions, while cross-model trained T_B led to feature distribution shift. It provided a strong guarantee for protecting the uniqueness of multimodal copyright and showed more robust performance in mitigating adversarial attacks.

Table 3: Effectiveness of the transform module. Res#1 and Res#2 represent the output results of without and with transform module, respectively, while Res. indicates the result of XOR comparison between two results.

Input	Model	$D(cos)$	$D(euc)$	Res#1	Module	$D(cos)$	$D(euc)$	Res#2	$\triangle_{cos}$	$\triangle_{euc}$	Res.
Basic	M_A	32.39	1.14	True	T_A	72.53	0.73	True	+38.14	-0.41	0
	M_A	32.39	1.14	True	T_B	55.95	0.93	True	+21.56	-0.21	0
Trigger	M_A	5.96	1.37	False	T_A	88.63	0.47	True	+82.67	-0.89	1
	M_A	5.96	1.37	False	T_B	35.52	1.13	False	+29.56	-0.24	0

(a) CLIP (b) Transform Module

Figure 3: Distributions of image-text pairs in CLIP and transform module embedding space.

Figure 3 illustrated that the distance between adversarial triggers and basic text triggers in the transform embedding space was significantly reduced after implementing the transform module. The relative positions of adversarial triggers and basic image triggers became more concentrated, which evidenced that our transform module successfully narrowed the distance between adversarial triggers and basic text triggers, while increasing their similarity to basic image triggers. Consequently, the transform model could correctly alter the output results of adversarial triggers, aligning them with outputs of basic image triggers.

4.5 Robustness

We evaluated two adversarial attack scenarios to assess the robustness of AGATE, aligning with threat models in Section 3.1.
Scenario 1: Adversary with Partial Knowledge of Trigger Generation. The adversary was aware of the existence of triggers and attempted to fabricate false triggers to bypass the watermark verification, but was unaware of the specific trigger generation strategy. We simulated the adversary by using various experimental groups in terms of types of trigger generation strategies (Figure 4). We set the strategy of adding adversarial, rectangular, and fixed-position noise patches to the basic trigger sampled from the Pascal dataset as a benchmark. Table 4 showed that the adversary failed to bypass the watermark verification by creating forged triggers with different strategies, as reflected by *Res.* = 0.

Specifically, we simulated the adversary lacked knowledge about which dataset the trigger originated from and what type of noise was added. Similar performance was obtained when other types of noise were examined, e.g., GN, PN, SPN, and MN. Results indicated a mismatch between the trigger and the text embedding, but failed to pass verification through the transform module. Results indicated a mismatch between the trigger and the text embedding, but failed to pass verification through the transform module. We analyzed

that the transform module was trained on triggers generated by benchmark and only corrected the output results of these triggers.

In addition, we examined the performance when the adversary had knowledge of the trigger dataset and the type of noise added, but did not know the shapes and positions of adversarial patches. Thus, adversaries might place different shapes at random positions on basic images. *Res*#1 = *True* indicated a large similarity between the trigger and the text embedding, evidencing the adversary's failure to forge a trigger that could bypass the watermark verification.
Scenario 2: Adversary Lacked Full Knowledge of Transform Module. The adversary remained unaware of the specific trigger and the detailed information about the transform module. To further evaluate the robustness of our framework, we considered enhancing the adversary's ability. In this case, the adversary had full knowledge of trigger generation. The experiment results in Table 5 demonstrated that despite the adversary's knowledge of the trigger generation and the transform module, they were still unable to successfully bypass the copyright verification, which highlighted the robustness of our framework against forgery attacks. First, we observed distinct changes in $D(cos)$ and $D(euc)$ when Trigger A was input into different CLIP models. However, *Res*#1 varied when different models were used. Trigger A output normally on M_B and M_C (*Res*#1 = *True*), which meant it was not a suitable backdoor trigger for these models. Thus, the adversary failed to directly apply model-specific triggers to other models. Second, we observed that *Res*#1 and *Res*#2 remained consistent when we used different transform modules T_B and T_C after M_A. Therefore, the adversary was unable to forge a transform module that could bypass the copyright verification, which meant the transform module played a critical role in the verification process. Combining these two aspects, attacks conducted by adversaries are ineffective whether they forge triggers or transform modules.

Two scenarios demonstrated that AGATE provided great robustness in preventing trigger forgery and in maintaining copyright verification against adversarial attacks. Regardless of whether the adversary was aware of the trigger's generation method, the adversary could not successfully evade watermark verification. Model-specific triggers and model-related transform modules provided a stronger guarantee for our two-phase watermark verification.

4.6 Impact of Trigger Number

We investigated the impact of the number of triggers on the effectiveness of the transform module. Distance-based metrics were employed, including $D(cos)(\times 10^{-2})$ and $D(euc)$. The evaluation focused on the difference in embedding distance after the transform module was applied, where a higher $D(cos)$ and lower $D(euc)$ value indicated a stronger effectiveness of transform module.

| (a) Basic image. | (b) Wikipedia. | (c) Gaussian noise. | (d) Poisson noise. | (e) Salt & Pepper noise. |
| (f) Adversarial noise. | (g) Multiplicative noise. | (h) Triangle patch. | (i) Circle patch. | (j) Random Position. |

Figure 4: Visualization for different types of trigger generation strategies.

Table 4: Comparison of classification task results across different trigger generation strategies in an adversarial scenario, where the adversary lacks complete knowledge of the trigger generation process.

	Type	$D(cos)$	$D(euc)$	Res#1	Module	$D(cos)$	$D(euc)$	Res#2	$\triangle_{cos}$	$\triangle_{euc}$	Res.
DataSets	Wikipedia	7.79	1.35	False	T_A	14.06	1.30	False	+6.27	-0.05	0
Noises	GN	4.70	1.38	False	T_A	37.89	1.10	False	+33.19	-0.28	0
	PN	3.29	1.39	False	T_A	11.81	1.32	False	+8.52	-0.07	0
	SPN	9.90	1.34	False	T_A	18.55	1.27	False	+8.65	-0.07	0
	MN	5.02	1.37	False	T_A	13.50	1.30	False	+8.48	-0.07	0
Shape	Triangle	32.48	1.16	True	T_A	76.40	0.67	True	+43.92	-0.49	0
	Circle	28.97	1.19	True	T_A	77.08	0.66	True	+48.11	-0.53	0
Position	Random	34.25	1.14	True	T_A	72.19	0.73	True	+37.94	-0.41	0
Pascal, Adv, Rectangle, Fixed		5.96	1.37	False	T_A	88.63	0.47	True	+82.67	-0.90	1

Table 5: Comparison of classification task results across different triggers, models, and transform modules in an adversarial scenario, where the adversary lacks knowledge of the specific trigger and transform module details.

Trigger	Model	$D(cos)$	$D(euc)$	Res#1	Module	$D(cos)$	$D(euc)$	Res#2	$\triangle_{cos}$	$\triangle_{euc}$	Res.
Trigger A	M_A	3.11	1.39	False	T_A	92.04	0.39	True	+88.93	-1.00	1
	M_A	3.11	1.39	False	T_B	37.73	1.11	False	+34.62	-0.28	0
	M_A	3.11	1.39	False	T_C	44.53	1.05	False	+41.42	-0.34	0
	M_B	34.09	1.14	True	T_A	55.95	0.93	True	+21.86	-0.21	0
	M_C	30.57	1.14	True	T_A	51.12	0.98	False	+20.55	-0.19	1
Trigger B	M_A	34.41	1.14	True	T_A	48.62	1.00	False	+14.21	-0.14	1
Trigger C	M_A	34.33	1.14	True	T_A	47.92	1.00	False	+13.59	-0.14	1

The results, presented in Figure 5, demonstrated a clear trend. Increasing the number of triggers consistently reduced the effectiveness of the transform module. For example, $D(cos)$ reached 92.04 with 16 triggers, showing a significant increase of +88.93 compared to 3.11 before connecting the module. This indicated that were forcibly bound by the transform module. As the number of triggers increased, $D(euc)$ continued to rise, stabilizing at 0.53 when 128 triggers were used. The increase in $D(euc)$ implied that reducing the number of triggers helped ensure that trigger embeddings stayed close. Results showed that a lower trigger number strengthened the transform module's ability to modify the embeddings.

Overall, the findings indicated that reducing the number of triggers enhanced the effectiveness of the transform module. However, reducing the number of triggers made it easier for adversaries to find specific triggers, greatly reducing the security of model copyright protection. This indicated that there existed a trade-off between maintaining high security and achieving maximum effectiveness according to the special requirements of application scenarios.

(a) $D(cos)$ (b) $D(euc)$

Figure 5: Performance under different trigger numbers.

5 Conclusion

In this work, we addressed critical issues about copyright protection in multimodal AI models by proposing AGATE, a novel black-box backdoor watermarking framework. AGATE simplified the process of trigger selection by generating random adversarial noise, enhancing trigger security and stealth. Proposed transform module ensured accurate copyright verification by correcting outputs against adversarial attacks. Our work demonstrated that AGATE could efficiently protect copyrights across various multimodal models, offering an economical and effective solution.

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2024. Self-Consuming Generative Models Go MAD. In *Proc. of the 12th International Conference on Learning Representations, (ICLR'24)*. OpenReview.net.

[2] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. IPGuard: Protecting Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary. In *Proc. of the 16th ACM Asia Conference on Computer and Communications Security, (AsiaCCS'21)*. ACM, 14–25.

[3] Huili Chen, Bita Darvish Rouhani, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. 2019. DeepMarks: A Secure Fingerprinting Framework for Digital Rights Management of Deep Learning Models. In *Proc. of the 2019 International Conference on Multimedia Retrieval, (ICMR'19)*. ACM, 105–113.

[4] Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. 2022. Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models. In *Proc. of the 43rd IEEE Symposium on Security and Privacy, (SP'22)*. IEEE, 824–841.

[5] Mark Everingham. 2008. The PASCAL visual object classes challenge 2008 (VOC2008) results. http://www.pascal-network.org/challenges/VOC/voc2008/year=workshop/index.html. [Accessed 21-02-2025].

[6] Keke Gai, Dongjue Wang, Jing Yu, Mohan Wang, Liehuang Zhu, and Qi Wu. 2025. MFL-Owner: Ownership Protection for Multi-modal Federted Learning via Orthogonal Transform Watermark. In *Proc. of the 39th AAAI Conference on Artificial Intelligence, (AAAI'25)*. AAAI Press, 3136–3144.

[7] Shangwei Guo, Tianwei Zhang, Han Qiu, Yi Zeng, Tao Xiang, and Yang Liu. 2021. Fine-tuning Is Not Enough: A Simple yet Effective Watermark Removal Attack for DNN Models. In *Proc. of the 30th International Joint Conference on Artificial Intelligence, (IJCAI'21)*. ijcai.org, 3635–3641.

[8] Tran Huynh, Dang Nguyen, Tung Pham, and Anh Tran. 2024. COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks. In *Proc. of the 38th AAAI Conference on Artificial Intelligence, (AAAI'24)*. AAAI Press, 2436–2444.

[9] Ting Jiang, Deqing Wang, Fuzhen Zhuang, Ruobing Xie, and Feng Xia. 2023. Pruning Pre-trained Language Models Without Fine-Tuning. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, (ACL'23)*. Association for Computational Linguistics, 594–605.

[10] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2023. Evading Watermark based Detection of AI-Generated Content. In *Proc. of the 30th ACM SIGSAC Conference on Computer and Communications Security, (CCS'23)*. ACM, 1168–1181.

[11] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. http://www.cs.toronto.edu/~kriz/cifar.html. CIFAR-10 dataset.

[12] Bowen Li, Lixin Fan, Hanlin Gu, Jie Li, and Qiang Yang. 2023. FedIPR: Ownership Verification for Federated Deep Neural Network Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 4 (2023), 4521–4536.

[13] Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. 2023. PLMmark: A Secure and Robust Black-Box Watermarking Framework for Pre-trained Language Models. In *Proc. of the 37th AAAI Conference on Artificial Intelligence, (AAAI'23)*. AAAI Press, 14991–14999.

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of the 13th European Conference on Computer Vision Part V, (ECCV'14) (Lecture Notes in Computer Science, Vol. 8693)*. Springer, 740–755.

[15] Peizhuo Lv, Chang Yue, Ruigang Liang, Yunfei Yang, Shengzhi Zhang, Hualong Ma, and Kai Chen. 2023. A Data-free Backdoor Injection Approach in Neural Networks. In *Proc. of the 32nd USENIX Security Symposium, (USENIX Security'23)*. USENIX Association, 2671–2688.

[16] Minzhou Pan, Zhenting Wang, Xin Dong, Vikash Sehwag, Lingjuan Lyu, and Xue Lin. 2024. Finding Needles in a Haystack: A Black-Box Approach to Invisible Watermark Detection. In *Proc. of the 18th European Conference on Computer Vision Part XXXIII, (ECCV'24) (Lecture Notes in Computer Science, Vol. 15091)*. Springer, 253–270.

[17] Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. 2023. Are You Copying My Model? Protecting the Copyright of Large Language Models for EaaS via Backdoor Watermark. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, (ACL'23)*. Association for Computational Linguistics, 7653–7668.

[18] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. 2022. Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR'22)*. IEEE, 13420–13429.

[19] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proc. of the 15th IEEE International Conference on Computer Vision, (ICCV'15)*. IEEE Computer Society, 2641–2649.

[20] 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. In *Proc. of the 38th International Conference on Machine Learning, (ICML'21)*. PMLR, 8748–8763.

[21] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proc. of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.

[22] Nikhil Rasiwasia, José Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proc. of the 18th ACM International Conference on Multimedia, (ACM MM'10)*. ACM, 251–260.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition,(CVPR'22)*. IEEE, 10674–10685.

[24] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. 2019. DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks. In *Proc. of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems, (ASPLOS'19)*. ACM, 485–497.

[25] Shuo Shao, Wenyuan Yang, Hanlin Gu, Zhan Qin, Lixin Fan, Qiang Yang, and Kui Ren. 2025. FedTracker: Furnishing Ownership Verification and Traceability for Federated Learning Model. *IEEE Trans. Dependable Secur. Comput.* 22, 1 (2025), 114–131.

[26] Yun Shen, Xinlei He, Yufei Han, and Yang Zhang. 2022. Model Stealing Attacks Against Inductive Graph Neural Networks. In *Proc. of the 43rd IEEE Symposium on Security and Privacy, (SP'22)*. IEEE, 1175–1192.

[27] Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N. Asokan. 2021. DAWN: Dynamic Adversarial Watermarking of Neural Networks. In *Proc. of the 29th ACM International Conference on Multimedia, (ACM MM'21)*. ACM, 4417–4425.

[28] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank K. Soong, and Tie-Yan Liu. 2024. NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 6 (2024), 4234–4245.

[29] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding Watermarks into Deep Neural Networks. In *Proc. of the 2017 International Conference on Multimedia Retrieval, (ICMR'17)*. 269–277.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proc. of the 31st Conference on Neural Information Processing Systems, (NeurIPS'17)*. 5998–6008.

[31] Jiacheng Wang, Hongyang Du, Dusit Niyato, Zehui Xiong, Jiawen Kang, Bo Ai, Zhu Han, and Dong In Kim. 2024. Generative Artificial Intelligence Assisted Wireless Sensing: Human Flow Detection in Practical Communication Environments.

IEEE J. Sel. Areas Commun. 42, 10 (2024), 2737–2753.

[32] Tong Wang, Yuan Yao, Feng Xu, Miao Xu, Shengwei An, and Ting Wang. 2024. Inspecting Prediction Confidence for Detecting Black-Box Backdoor Attacks. In *Proc. of the 38th AAAI Conference on Artificial Intelligence, (AAAI'24)*. AAAI Press, 274–282.

[33] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, Hongyang Chao, and Han Hu. 2023. TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance. In *Proc. of the 19th IEEE International Conference on Computer Vision, (ICCV'23)*. IEEE, 21913–21923.

[34] Liujie Xu, Jing Wu, Jing-Dong Zhu, and Ling Chen. 2025. Effects of AI-assisted dance skills teaching, evaluation and visual feedback on dance students' learning performance, motivation and self-efficacy. *Int. J. Hum. Comput. Stud.* 195 (2025), 103410.

[35] Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *Int. J. Hum. Comput. Stud.* 197 (2025), 103455.

[36] Yifan Yan, Xudong Pan, Mi Zhang, and Min Yang. 2023. Rethinking White-Box Watermarks on Deep Learning Models under Neural Structural Obfuscation. In *Proc. of the 32nd USENIX Security Symposium, (USENIX Security'23)*. USENIX

Association, 2347–2364.

[37] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR'24)*. IEEE, 15952–15962.

[38] Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting Language Generation Models via Invisible Watermarking. In *Proc. of the 40th International Conference on Machine Learning, (ICML'23)*. PMLR, 42187–42199.

[39] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. 2023. AdvCLIP: Downstream-agnostic Adversarial Examples in Multimodal Contrastive Learning. In *Proc. of the 31st ACM International Conference on Multimedia, (ACM MM'23)*. ACM, 6311–6320.

[40] Wenjun Zhu, Xiaoyu Ji, Yushi Cheng, Shibo Zhang, and Wenyuan Xu. 2023. TPatch: A Triggered Physical Adversarial Patch. In *Proc. of the 32nd USENIX Security Symposium, (USENIX Security'23)*. USENIX Association, 661–678.

[41] Wei Zong, Yang-Wai Chow, Willy Susilo, Joonsang Baek, Jongkil Kim, and Seyit Camtepe. 2024. IPRemover: A Generative Model Inversion Attack against Deep Neural Network Fingerprinting and Watermarking. In *Proc. of the 38th AAAI Conference on Artificial Intelligence, (AAAI'24)*. AAAI Press, 7837–7845.