Vision-Language Model IP Protection via Prompt-based Learning

Lianyu Wang[1]*, Meng Wang[2]*, Huazhu Fu[3]†, Daoqiang Zhang[1]†

[1]The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, China

[2]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore

[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Abstract

Vision-language models (VLMs) like CLIP (Contrastive Language-Image Pre-Training) have seen remarkable success in visual recognition, highlighting the increasing need to safeguard the intellectual property (IP) of well-trained models. Effective IP protection extends beyond ensuring authorized usage; it also necessitates restricting model deployment to authorized data domains, particularly when the model is fine-tuned for specific target domains. However, current IP protection methods often rely solely on the visual backbone, which may lack sufficient semantic richness. To bridge this gap, we introduce IP-CLIP, a lightweight IP protection strategy tailored to CLIP, employing a prompt-based learning approach. By leveraging the frozen visual backbone of CLIP, we extract both image style and content information, incorporating them into the learning of IP prompt. This strategy acts as a robust barrier, effectively preventing the unauthorized transfer of features from authorized domains to unauthorized ones. Additionally, we propose a style-enhancement branch that constructs feature banks for both authorized and unauthorized domains. This branch integrates self-enhanced and cross-domain features, further strengthening IP-CLIP's capability to block features from unauthorized domains. Finally, we present new three metrics designed to better balance the performance degradation of authorized and unauthorized domains. Comprehensive experiments in various scenarios demonstrate its promising potential for application in IP protection tasks for VLMs.

1. Introduction

Driven by the availability of large-scale data and powerful computing hardware, vision-language models (VLMs) like CLIP have recently achieved remarkable generalization across a wide range of downstream tasks [24, 35, 36],

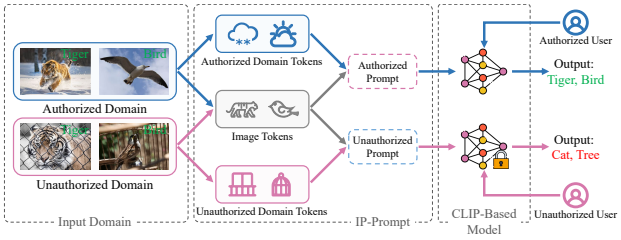

Figure 1. Illustration of model IP protection with IP-CLIP. Domain and image tokens form the IP-Prompt, which a CLIP-based model audits to verify data origin. This prevents unauthorized transfers and degrades performance in unauthorized domains. Notably, IP-Prompt is a lightweight, plug-and-play module for CLIP-based models.

leading to a surge in their commercial significance. However, developing a well-trained VLM is a resource-intensive endeavor, requiring substantial investments in time, manpower, and resources. This includes the design of specialized architectures [2, 10], access to vast amounts of high-quality data [6, 18, 31], and the use of expensive computational resources [37]. As a result, protecting these models' intellectual property (IP) has garnered significant attention [28–30, 32].

Previous research on IP protection has primarily concentrated on two aspects: ownership verification (i.e., verifying who owns the model) [3, 20, 22, 25] and usage authorization (i.e., authorizing who has the right to deploy the model) [9, 23]. Some of these approaches incorporate deep watermarks, embedding unique identifiers such as inputs, parameters, gradients, architectures, or even outputs. Others extract distinctive model characteristics, acting as "fingerprints" [21] for deep models. While these techniques provide a degree of protection, they can be easily bypassed through fine-tuning or retraining. Moreover, authorized users are often unrestricted in how they apply the model, allowing them to effortlessly transfer high-performance models to similar tasks, which can lead to implicit IP infringement. This problem stems from the fact that VLM's trained visual backbones often generalize across domains, which

*L. Wang and M. Wang contributed equally to this work.

†Corresponding author: H. Fu (hzfu@ieee.org) and D. Zhang (dqzhang@nuaa.edu.cn).

can breed model stealing, leading to illegal misuse and implicit intellectual property infringement. An intuitive solution is to refine the model's generalization boundary to focus on domain-specific features and restrict their use to authorized domains. NTL [28] achieves this by amplifying the maximum mean discrepancy (MMD) between authorized and unauthorized domains, thus narrowing the model's generalization scope. In contrast, CUTI-domain [29] introduces an intermediate domain that combines features from both domains, preventing unauthorized transfers. Although existing deep model IP protection methods can provide commendable performance in specific scenarios, they face two fundamental challenges. Firstly, they require training models from scratch or extensive fine-tuning, which is particularly demanding for VLMs due to their resource-intensive nature. To address this, some prompt tuning methods techniques, such as CoOp [36] and MaPLe [16] have shown superior performance on some specific downstream tasks. CoOp uses soft prompts to learn text prompts, while MaPLe introduces visual language prompts to enhance synergy. Secondly, some methods [28, 29] attempt to constrain model performance by generating supplementary data. However, these methods often introduce additional training steps, and the generated data typically lack adequate constraints and control, complicating practical use.

To tackle these challenges, we introduce IP-CLIP, a novel approach for IP protection in CLIP-based models. IP-CLIP utilizes a lightweight prompt-tuning technique called IP-Prompt (illustrated in Fig. 1) to distinguish between authorized and unauthorized prompts without requiring full fine-tuning of all pre-trained parameters. Our approach involves learning new prompts consisting of two types of tokens: i) Authorized/unauthorized domain token: this token captures the multi-scale style information of authorized/unauthorized domains from the CLIP visual encoder. ii) Image token: to effectively learn the visual distribution in the semantic space and obtain cue distributions for each class, we utilize multi-scale visual feature responses from various layers of the CLIP visual encoder. The downstream CLIP-based model integrates these two tokens into its decision-making process, allowing it to simultaneously identify both the Authorization and category of the input image. This enables accurate predictions for images from the authorized domain while deliberately producing incorrect results for samples from unauthorized domains. Notably, IP-Prompt functions as a lightweight, plug-and-play module that can be positioned at the front end of various CLIP-based models to provide IP protection. Additionally, we introduce a style enhancement branch with feature banks for both authorized and unauthorized domains. This branch integrates self-enhanced and cross-domain features into the model, improving its ability to recognize authorized features while excluding unauthorized ones. Finally, we design three new metrics tailored to the IP protection scenario to balance performance between authorized and unauthorized domains. The main contributions of this paper are summarized as follows:

- We propose the **IP-CLIP** framework, an innovative approach for IP protection of VLMs, with only minimal parameter updates. This framework is designed to prevent the unauthorized transfer of well-trained, large-scale VLMs from authorized to unauthorized domains.
- We design a lightweight, plug-and-play **IP-Prompt** that can be integrated into various CLIP-based models for effective IP protection of VLMs.
- Our approach includes a **style enhancement branch** that generates diverse visual features and integrates self-enhanced and cross-domain features into the model. This enables the protected model to better identify authorized features and exclude unauthorized ones.
- We introduce three **new metrics** for a comprehensive evaluation of IP protection capabilities, addressing gaps in current methods. Extensive experiments demonstrate the effectiveness of IP-CLIP on various datasets and scenarios, providing strong evidence that our method offers a robust solution for model IP protection.*

2. Related Work

2.1. Visual Language Models and Prompt Tuning

Large-scale visual language models (VLMs) integrate visual and textual inputs for a more comprehensive understanding, achieving strong performance in various computer vision tasks [13, 14, 17]. Models like CLIP [24] and VisualBERT [19] rely on pre-trained language models (e.g., BERT [7], GPT [1]) for text encoding, while visual inputs are processed via convnets or visual transformers. As these models scale up, their computational demands increase, making updates costly. To address this, parameter-efficient tuning methods are essential.

Prompt tuning is one such approach, which focuses on learning a small set of parameters while keeping the larger model frozen [15]. CoOp [36] introduced the use of soft prompts in VLMs, demonstrating that carefully crafted text prompts can enhance image recognition performance. By incorporating lightweight neural networks to dynamically generate prompts for individual images, CoCoOp [35] addresses the issue of prompt overfitting. VPT [15] achieved strong results by using a small number of visual prompts, and MaPLe [16] further combined textual and visual prompts within CLIP to improve the alignment between text and image representations. Although these parameter fine-tuning methods have demonstrated effectiveness, they offer insufficient security. Lacking robust IP protection, the critical issue of safeguarding IP in large-scale

*https://github.com/LyWang12/IP-CLIP

models has garnered growing attention and scrutiny.

2.2. Intellectual Property (IP) Protection

A comprehensive IP protection strategy should address both ownership verification and applicability authorization. Ownership verification identifies the rightful owner of the model, typically using watermarks or fingerprinting. Peng *et al*. [22] introduced a general adversarial perturbation fingerprinting method, which uses contrastive learning to match fingerprints with similarity scores. Bai *et al*. [3] proposed BadCLIP, which impacts image and text encoders using trigger-aware prompts, while. Ren *et al*. [25] adopted a poison-only backdoor approach for embedding watermarks and used hypothesis testing for remote verification. However, these methods have been proven vulnerable to certain removal and covering techniques.

Applicability authorization focuses on restricting the model's generalizability to specific domain. Wang *et al*. [28] introduced non-transfer learning (NTL), which uses an estimator with a feature kernel to highlight domain-specific differences. Zeng *et al*. [33] extended NTL to natural language processing and auxiliary domain classifiers for better domain separation. Hong *et al*. [11] further proposed H-NTL, leveraging a causal model to disentangle content and style as latent factors, thereby guiding the learning of non-transferable representations based on intrinsic causal relationships. Wang *et al*. [29] proposed an innovative compact non-transferable isolation domain (CUTI-domain) to isolate authorized and unauthorized domains, limiting performance transfer. Existing IP protection methods can be effective but often require extensive training or fine-tuning, which is resource-intensive for VLMs. Additionally, methods relying on supplementary data often lack necessary constraints and controllability, complicating their practical use.

3. Method

3.1. Problem Definition

IP protection aims to confine model performance to the authorized domain while reducing its recognition ability in the unauthorized domain. Formally, we define the IP protection task as follows [12]:

Definition 1 (IP protection): *Let* $D_a = \{x_{ai}, y_{ai}\}_{i=1}^{N_a}$ *denote the dataset for the authorized domain, and* $D_u = \{x_{ui}, y_{ui}\}_{i=1}^{N_u}$ *represent the dataset for the unauthorized domain, where* N_a *and* N_u *are the number of samples in the authorized and unauthorized domains, respectively. Data* X_a *and* X_u *from these domains are drawn from different distributions but share the same label space* Y*. In the authorized domain, the model aims to map data to labels:*

$$F(X_a) \to Y. \tag{1}$$

The challenge of the IP protection task is to achieve non-transferability to the unauthorized domain while minimally affecting performance in the authorized domain:

$$F(X_u) \perp Y \; and \; F(X_a) \perp F(X_u), \tag{2}$$

where $\perp$ *denotes statistical independence.* Current IP protection methods usually rely solely on visual backbones [12, 28, 29], which may lack sufficient semantic richness. To bridge this gap, we introduce IP-CLIP, a lightweight IP protection strategy tailored for vision-language models.

3.2. Overview of IP-CLIP

Fig. 2 (a) illustrates the details of our proposed IP-CLIP framework. The primary objective is to constrain model performance to the authorized domain by learning both image and domain-specific tokens, thereby emphasizing the unique features of the authorized domain while preventing unauthorized generalization. To accomplish this, we feed both the authorized domain data x_a and the unauthorized domain data x_u into CLIP's frozen visual encoder in parallel, producing the output features f_v^a and f_v^u, respectively. A learnable IP Projector is employed to capture multi-scale features from different layers of the visual encoder, generating authorized / unauthorized domain tokens T_a / T_u and image tokens $[V_1, V_2, \ldots, V_L]$, which are concatenated as input prompts for the frozen text encoder of CLIP, as described in Sec. 3.3. The prediction result is obtained by calculating the similarity between text feature f_t and visual feature f_v, and the label is denoted as y. The style enhancement branch (Sec. 3.4), associated with the feature banks, further improves the robustness of the features in distinguishing between authorized and unauthorized domains. The frozen layers of our proposed IP-CLIP framework are labeled with snowflakes, while the few trainable layers are marked with sparks.

3.3. Our Proposed Prompt Learning

Instead of the static prompting technique, we aim to learn prompts directly from the visual domain to efficiently encode visual distributions. Our IP protection approach has two main objectives in prompt tuning: i) introduce domain-specific tokens for authorized / unauthorized domains, and ii) generate domain-independent image tokens for visual recognition tasks, as illustrated in Fig. 2 (c). Specifically, multi-scale features $[f_v^{(1)}, f_v^{(2)}, \ldots, f_v^{(M)}]$ are extracted from the frozen visual encoder, where $f_v^{(m)}$ represents the response from the m-th layer of the encoder. To create domain-specific tokens for authorized / unauthorized domains, multi-scale style features (represented by first-order and second-order batch-wise feature statistics) are computed and combined, resulting in $[\mu^{(1)}; \sigma^{(1)}; \ldots; \mu^{(M)}; \sigma^{(M)}]$, which are then processed by the IP Projector to produce domain-specific tokens T. Additionally, the multi-scale features $[f_v^{(1)}, f_v^{(2)}, \ldots, f_v^{(M)}]$ are

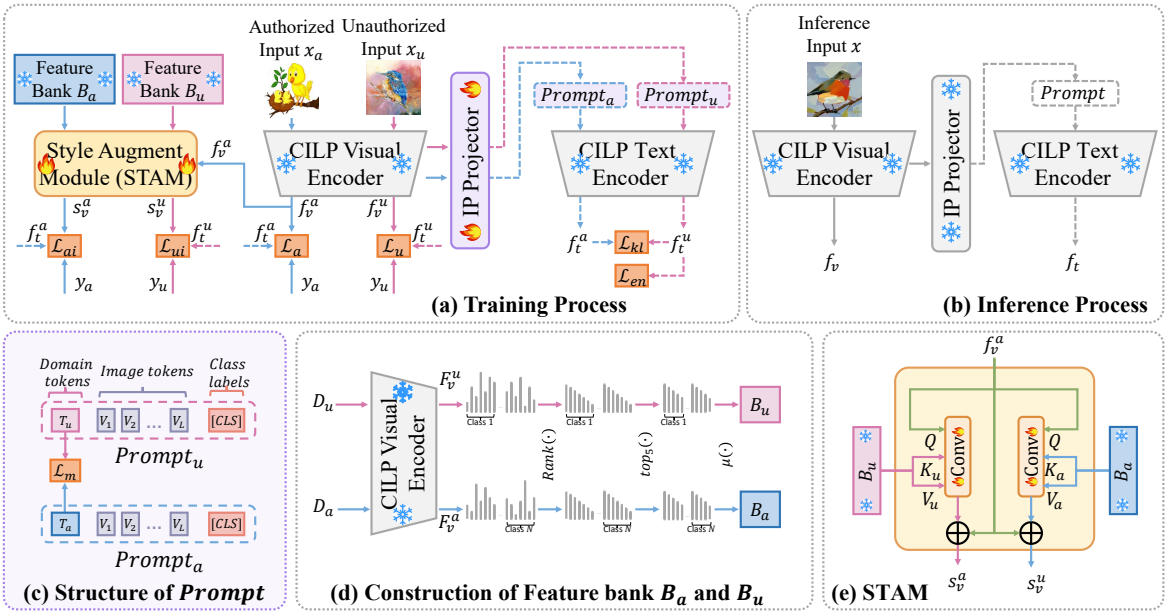

Figure 2. (a) The architecture of IP-CLIP is based on a frozen CLIP backbone, where snowflakes denote frozen layers and sparks represent trainable layers. During training, inputs from both the authorized domain x_a and unauthorized domain x_u are fed into the frozen CLIP visual encoder in parallel to generate feature vectors f_v^a and f_v^u. The IP projector extracts domain tokens and image tokens from the visual encoder, which are then used to construct prompts as inputs to the text encoder. The style enhancement branch takes the frozen feature bank and f_v^a as input, with s_v representing the enhanced visual features. The prediction result is derived by calculating the similarity between the visual feature s_v/f_v and the text feature f_t. y and $\mathcal{L}$ represent the label and loss function, respectively. (b) The Inference process of IP-CLIP. (c) Structure of $Prompt_a$ and $Prompt_u$. (d) Construction of Feature bank B_a and B_u, where D and F represent the input dataset and its corresponding visual feature set, respectively. During training, the feature banks remain frozen. (e) Structure of STAM.

passed through IP Projector to generate L image-specific tokens $[V_1, V_2, \ldots, V_L]$. Finially, the prompt for the authorized domain is denoted as:

$$Prompt_a = [T_a; V_1, V_2, \ldots, V_L; [CLS]], \tag{3}$$

while for the unauthorized domain, it is denoted as:

$$Prompt_u = [T_u; V_1, V_2, \ldots, V_L; [CLS]], \tag{4}$$

which are then input into the frozen text encoder to generate text features f_t^a and f_t^u, respectively.

3.4. Style-Enhancement Branch

For the style enhancement branch, we construct feature banks for both the authorized and unauthorized domains and introduce a style augment module (STAM) to diversify the features.

Constructing feature banks. Leveraging CLIP's zero-shot capabilities, we extract text and image features from D_a and D_u, as in Fig. 2 (d). For the authorized domain, we compute a confidence score (i.e., the maximum probability) for each image based on CLIP's predictions. Similarly, in the unauthorized domain, we calculate confidence scores and assign pseudo-labels based on the highest score. We then select the visual features with the highest confidence in each category from both domains to construct N-way K-shot feature banks, where N is the number of categories and $K = 5$ is the number of samples per category. Finally, the centroid features for each category are calculated to form the authorized domain feature bank (B_a) and the unauthorized domain feature bank (B_u), both expressed as $\mathbb{R}^{N \times C}$, where C denotes the feature dimension. Note that the feature bank is built by iterating over the data only before training, after which it is frozen during the training process.

STyle Augment Module (STAM). STAM utilizes the frozen feature banks to guide images in acquiring self-enhanced and cross-domain features, as illustrated in Fig. 2 (e). First, the query Q is calculated from the input feature f_v^a, while the key K_a and value V_a are derived from the authorized domain bank. Similarly, K_u and V_u are calculated from the unauthorized domain bank. We derive enhanced s_v^a and s_v^u by utilizing a learnable attention layer combined with a residual connection. This mechanism enables the image feature to concentrate on the features from the authorized or unauthorized domain banks. This process can be formally expressed as:

$$s_v^a = \text{Conv}(\text{softmax}\left(\frac{QK_a^T}{\sqrt{d_k}}\right)V_a) + f_v^a, \tag{5}$$

$$s_v^u = \text{Conv}(\text{softmax}\left(\frac{QK_u^T}{\sqrt{d_k}}\right)V_u) + f_v^u. \quad (6)$$

Here, $\sqrt{d_k}$ denotes the scaling factor, while T represents the transpose operation.

3.5. Training Strategy

Target-specified IP-CLIP. We begin by detailing the training process for our proposed IP-CLIP, assuming both the authorized and unauthorized domains are known. To allow the model to effectively differentiate between the authorized domain token T_a and unauthorized domain token T_u, we use mean squared error (MSE) loss to maximize their separation, as described by:

$$\mathcal{L}_m = \mathcal{L}_{MSE}(T_a, T_u). \quad (7)$$

Next, we utilize contrastive loss function $\mathcal{L}_a$ / $\mathcal{L}_v$ to optimize the image-text mapping between image feature f_v^a / f_v^u and the text feature f_t^a / f_t^u, as shown in:

$$\mathcal{L}_a = \frac{\exp(\langle f_v^a, f_t^a(y_a)\rangle/\tau)}{\sum_{k=1}^{K}\exp(\langle f_v^a, f_t^a(k)\rangle/\tau)}, \quad (8)$$

where τ denotes temperature parameter, K denotes the number of classes and $\langle\cdot,\cdot\rangle$ denotes the cosine similarity.

Similarly, the enhanced feature s_v^a / s_v^u is aligned with the text representation f_t^a / f_t^u by $\mathcal{L}_{ai}$ / $\mathcal{L}_{ui}$, which can be expressed as:

$$\mathcal{L}_{ai} = \frac{\exp(\langle s_v^a, f_t^a(y_a)\rangle/\tau)}{\sum_{k=1}^{K}\exp(\langle s_v^a, f_t^a(k)\rangle/\tau)}. \quad (9)$$

For text representations, we use Kullback-Leibler (KL) divergence loss to further separate the distances between the authorized and unauthorized domains:

$$\mathcal{L}_{kl} = KL(f_t^a, f_t^u). \quad (10)$$

Additionally, we impose constraints on the similarity distribution of the unauthorized domain's text features, ensuring they maintain low entropy through:

$$\mathcal{L}_{en} = \mathcal{L}_{entropy}(f_t^u). \quad (11)$$

Finally, our overall loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_a - \mathcal{L}_u + \mathcal{L}_{ai} - \mathcal{L}_{ui} - \mathcal{L}_{kl} - \lambda_1 \cdot \mathcal{L}_m + \lambda_2 \cdot \mathcal{L}_{en}. \quad (12)$$

Where λ_1 and λ_2 are weight factors. The overall training strategy is shown in *Supplementary Algorithm 1*.

Target-free IP-CLIP. In a restricted setting where only authorized domain data is accessible, our IP protection focuses on reducing recognition performance for potential out-of-domain (OOD) data with similar content but different styles. Unlike Wang [29]'s use of GANs for OOD data synthesis, we intervene on the style factor to achieve this. Our method enhances style [5] without changing the content (as in *Supplementary Tab. 1*). We treat all style-augmented images as unauthorized and train the model similarly to target-specific IP-CLIP. The full algorithm is detailed in *Supplementary Algorithm 2*.

Inference. During testing, as shown in Fig. 2 (b), the sample is input into visual encoder, and the trained IP Projector generates the corresponding prompt, which is then fed into text encoder. Finally, the cosine similarity between f_v and f_t is computed to produce the prediction p:

$$p = \arg\max_i \langle f_t, f_{v,i}\rangle, \quad (13)$$

where i denote the index of class.

4. Experiment

4.1. Implementation Details

We evaluated our method on three popular domain adaptation / generalization benchmarks, which feature more categories, larger numbers, and more complex content compared to the existing works [28–30]:

1. **Office-31 [26]** comprises images from three distinct domains—Amazon, Dslr, and Webcam—spanning 31 categories and containing over 4,000 samples.
2. **Office-Home-65 [27]** consists of over 15,000 images distributed across four domains—Art, Clipart, Product, and Real-World—organized into 65 distinct categories.
3. **Mini-DomainNet [34]** contains over 140,000 images across domains including Clipart, Painting, Real, and Sketch, with 126 categories.

The substantial differences in image style and quality across domains in these datasets make them ideal for evaluating the effectiveness of model IP protection algorithms in cross-domain image recognition tasks.

Our comprehensive experiments are implemented on the PyTorch platform and an NVIDIA GeForce RTX 3090 GPU with 24GB of memory. The Adam optimizer, with an initial learning rate of e^{-5}, is employed for model optimization. We utilize the pre-trained CLIP backbone architecture. Consistent with standard evaluation protocols, accuracy (%) is used as the primary performance metric for each task.

4.2. Result of Target-Specified IP-CLIP

In the target-specified scenario, we randomly select two domains from each dataset: one as the authorized domain and the other as the unauthorized domain, thereby forming a IP protection task. We first compute A_a^{SL}/A_u^{SL}, the performance of supervised learning CLIP with prompt fine-tuning (SL-CLIP) trained on the authorized domain and tested on the authorized / unauthorized domain, and A_a^{IP}/A_u^{IP}, the performance of IP-CLIP on the same domain. This process is denoted as: $A^{SL} \Rightarrow A^{IP}$, with results shown in Tab. 1.

Authorized/Unauthorized	Amazon	Dslr	Webcam	$W_{ua}\uparrow$	$D_u\uparrow$	$D_a\downarrow$
Amazon	79.4 ⇒ 79.4	87.5 ⇒ 7.5	88.8 ⇒ 8.8	63.52	80.00	0.00
Dslr	83.8 ⇒ 3.8	95.7 ⇒ 95.7	98.8 ⇒ 6.3	82.54	86.25	0.00
Webcam	80.0 ⇒ 3.8	92.5 ⇒ 2.5	94.4 ⇒ 94.4	78.45	83.10	0.00
Mean		/		74.84	83.12	0.00

Table 1. The accuracy (%) of target-specified IP-CLIP on the Office-31 [26]. The vertical/horizontal axis denotes the authorized/unauthorized domain. In each task, the left of '⇒' shows the test accuracy of supervised learning CLIP on the unauthorized domain, while the right presents the accuracy of IP-CLIP. W_{ua} represents the weighted drop, while D_u and D_a denote the drop rates for the unauthorized and authorized domains, respectively.

Given CLIP's strong feature extraction capabilities, it tends to generalize well, resulting in higher A^{SL}. However, our goal is to restrict the model to the authorized domain, leading to a lower A^{IP}. Additionally, the previous method only assessed the drop rates $D_a = A_a^{SL} - A_a^{IP}$ for the authorized and $D_u = \mu(A_u^{SL} - A_u^{IP})$ for the unauthorized domains, which is insufficient. An effective IP protection model must balance maintaining high performance in the authorized domain with degrading performance in the unauthorized domain. To address this, we define a new weighted metric, W_{ua}, as follows:

$$W_{ua} = A_a^{IP} \cdot [D_u - D_a]. \tag{14}$$

Tab. 2 present the performance comparison between the proposed IP-CLIP and SOTA methods on the Office-31 [26]. The results for CUTI [29] and NTL [28] were obtained by reproducing their original implementations. For a fair comparison, we adapted these methods into CLIP-based versions, referred to as CUTI† and NTL†, respectively. The results indicate that the CLIP-based model exhibits stronger protection capabilities compared to the CNN-based model, achieving an average W_{ua} of 74.84% for IP-CLIP, 72.48% for CUTI†, 54.98% for NTL†, 70.09% for CUTI, and 62.11% for NTL. IP-CLIP achieves the highest scores across nearly all metrics. Although CUTI slightly outperforms IP-CLIP in D_u in the "webcam" domain, its D_a is 2.5%, significantly above IP-CLIP's 0.0%. The goal of the IP protection task is to reduce performance in the unauthorized domain while preserving accuracy in the authorized domain. Thus, relying solely on D_u or D_a is insufficient for comprehensive evaluation, making a combined metric like W_{ua} essential for a balanced assessment.

Additionally, we evaluated the proposed IP-CLIP on Office-Home-65 [27] and Mini-DomainNet [34] to further verify its effectiveness and versatility. The experimental results are summarized in Tab. 2, with further details available in *Supplementary Tab. 2-16*. Across these datasets, the CLIP-based IP protection scheme consistently outperforms its CNN counterpart, with IP-CLIP demonstrating the strongest protection capabilities. Fig. 3 presents several visualization examples.

4.3. Result of Ownership Verification

To further verify model ownership, erroneous results are deliberately triggered. Specifically, a conventional backdoor watermark is applied to each authorized domain [29], with the processed data used as the corresponding unauthorized domain. For ease of observation and analysis, we computed the accuracy of the supervised convolutional neural network (SL-CNN) related to CNN-based NTL/CUTI, as well as the supervised CLIP (SL-CLIP) according to CLIP-based NTL†/CUTI†/IP-CLIP. After computing A_a and A_u, a new weighted metric is introduced based on these values:

$$O_{ua} = A_u^{SL} \cdot [A_a^{Method} - A_u^{Method}]. \tag{15}$$

As presented in Tab. 3, the difference in accuracy between SL-CNN/SL-CLIP with a watermark (A_a^{SL}) and without a watermark (A_u^{SL}) is minimal, indicating low sensitivity to the watermark. In contrast, IP-CLIP shows a significant reduction in accuracy on unauthorized domains with embedded watermarks (A_u^{IP}). This disparity in performance serves as an effective measure for verifying model ownership. Furthermore, the performance comparison between IP-CLIP and other state-of-the-art methods reveals that, compared to CNN-based models, CLIP-based models show stronger model protection capabilities. Notably, O_{ua} of IP-CLIP is 71.3%, outperforming CUTI† and NTL† by approximately 5.6% and 18.7%, respectively, with statistically significant differences ($p < 0.05$ [4, 8]).

4.4. Result of Target-Free IP-CLIP

In a more rigorous setting, i.e., the target-free scenario, we generate unauthorized domains for each authorized domain, as described in Sec. 3.5. Specifically, to assess the performance of target-free IP-CLIP on the Office-31 [26] dataset, we conduct three transfer tasks. For each task, one domain is selected as the authorized domain, with unauthorized domains generated accordingly, while the remaining unknown domains are used for testing. The experimental results are presented in Tab. 4 and Tab. 5.

Similarly, we constructed tasks using more datasets and compared the results with the SOTA method, as shown in Tab. 5 (with additional details provided in *Supplementary Tab. 17-31*). After analyzing the results, we found that IP-CLIP consistently achieved the highest W_{ua} across all three datasets. This demonstrates its ability to effectively reduce recognition accuracy for unauthorized domains while maintaining strong recognition performance for authorized domains, even in tasks of varying complexity, thus proving its effectiveness in the restricted model IP protection task.

4.5. Result of Applicability Authorization

In the applicability authorization scenario, we assess the model's effectiveness by limiting its generalization ability

Datasets	Authorized Domain	$W_{ua}\uparrow$					$D_u\uparrow$					$D_a\downarrow$				
		NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP
Office-31 [26]	Amazon	41.37	60.94	56.34	62.06	**63.52**	55.50	74.40	75.80	79.35	**80.00**	3.10	0.80	1.80	0.60	**0.00**
	Dslr	70.94	75.33	76.09	80.13	**82.54**	74.20	81.90	77.35	85.05	**86.25**	1.55	0.80	1.30	0.70	**0.00**
	Webcam	74.02	74.02	32.50	75.24	**78.45**	75.80	38.70	75.80	**84.38**	83.10	**0.00**	**0.00**	3.10	2.50	**0.00**
	Mean	62.11	70.09	54.98	72.48	**74.84***	68.50	76.32	65.00	82.93	**83.12**	1.55	0.53	2.07	1.27	**0.00***
Office-Home-65 [27]	Art	27.53	35.62	13.44	41.58	**52.00**	37.27	47.16	15.83	53.40	**61.33**	0.80	0.30	**0.10**	3.00	0.30
	Clipart	43.23	45.67	48.83	53.37	**56.45**	54.31	57.35	65.67	72.40	**75.47**	0.20	0.20	0.30	0.63	**0.10**
	Product	41.31	41.78	39.90	56.82	**58.71**	45.01	45.82	43.00	61.83	**63.77**	0.30	0.50	**0.00**	0.37	0.30
	RealWorld	22.93	35.87	28.87	49.41	**53.25**	30.37	42.95	34.67	57.33	**59.33**	2.40	0.30	1.90	1.50	**0.10**
	Mean	33.75	39.73	32.76	50.29	**55.10***	41.74	48.32	39.79	61.24	**64.98***	0.43	0.33	0.57	1.38	**0.20**
Mini-DomainNet [34]	Clipart	25.63	30.29	38.62	50.26	**51.47**	36.60	40.87	46.30	59.40	**61.00**	2.10	0.80	0.60	**0.20**	0.30
	Painting	19.53	19.88	41.66	46.88	**53.85**	32.37	33.23	53.80	66.90	**67.07**	**0.50**	0.70	1.60	5.30	**0.50**
	Real	29.26	31.52	52.29	54.77	**58.82**	35.87	38.40	59.03	62.30	**65.27**	1.20	1.10	0.80	1.10	**0.20**
	Sketch	29.37	30.18	33.78	51.09	**54.59**	45.77	46.90	42.77	64.57	**68.57**	1.00	0.96	0.60	0.70	**0.50**
	Mean	25.95	27.97	41.59	50.75	**54.68***	37.65	39.85	50.48	63.29	**65.48***	1.27	0.87	1.00	2.20	**0.33***

Table 2. W_{ua}, D_u, and D_a of target-specified IP-CLIP, CUTI†, NTL†, CUTI and NTL. W_{ua} represents the proposed weighted drop, while D_u and D_a denote the drop rates for the unauthorized and authorized domains, respectively. The best performance is indicated by the numbers in bold. Statistical significance (p-value < 0.05 [4, 8]) is denoted with: *(IP-CLIP vs. others).

Datasets	Authorized with / without Patch	CNN-Based Models					CLIP-Based Modesl					
		SL-CNN	NTL [28]		CUTI [29]		SL-CLIP [24]	NTL† [28]		CUTI† [29]		IP-CLIP	
		A_u/A_a	A_u/A_a	$O_{ua}\uparrow$	A_u/A_a	$O_{ua}\uparrow$	A_u/A_a	A_u/A_a	$O_{ua}\uparrow$	A_u/A_a	$O_{ua}\uparrow$	A_u/A_a	$O_{ua}\uparrow$
Office-31 [26]	Amazon	59.4 / 78.1	3.1 / 67.2	38.1	1.6 / 78.1	45.4	80.0 / 81.3	15.0 / 77.5	50.0	3.8 / 80.0	61.0	3.8 / 81.3	**62.0**
	Dslr	50.0 / 98.4	0.0 / 92.2	46.1	4.7 / 93.8	44.6	97.5 / 98.8	5.0 / 95.0	87.8	2.5 / 95.0	90.2	3.8 / 97.5	**91.4**
	Webcam	62.5 / 95.3	1.6 / 93.8	57.6	4.7 / 92.2	54.7	95.0 / 97.5	2.5 / 93.8	86.7	7.5 / 95.0	83.1	1.3 / 96.3	**90.3**
Office-Home-65 [27]	Art	54.7 / 76.8	1.6 / 45.6	24.1	1.6 / 76.0	40.7	83.5 / 85.5	16.5 / 87.3	59.1	6.0 / 87.0	67.6	5.0 / 87.5	**68.9**
	Clipart	70.8 / 78.1	1.6 / 54.9	37.7	3.1 / 69.0	46.7	73.8 / 74.3	5.5 / 73.5	50.2	17.0 / 73.3	41.5	5.5 / 73.5	**50.2**
	Product	65.9 / 92.2	2.3 / 69.8	44.5	2.6 / 91.1	58.3	90.5 / 94.0	60.5 / 92.5	29.0	31.0 / 93.0	56.1	2.0 / 92.8	**82.2**
	RealWorld	61.2 / 82.6	1.8 / 77.3	46.2	0.3 / 83.6	51.0	87.5 / 88.5	17.5 / 87.8	61.5	5.0 / 86.3	71.1	6.5 / 92.0	**74.8**
Mini-DomainNet [34]	Clipart	50.3 / 65.5	0.8 / 37.8	18.6	1.6 / 67.8	33.3	84.0 / 85.1	57.1 / 86.4	24.6	13.7 / 85.2	60.1	5.6 / 85.4	**67.0**
	Painting	39.6 / 57.6	0.8 / 46.1	17.9	1.0 / 56.9	22.1	79.5 / 81.9	31.1 / 80.0	38.9	4.1 / 78.8	59.4	4.1 / 81.1	**61.2**
	Real	50.2 / 82.6	0.0 / 40.3	20.2	0.5 / 83.2	41.5	88.9 / 89.4	26.2 / 91.9	58.4	11.4 / 92.1	71.7	5.9 / 89.7	**74.5**
	Sketch	57.6 / 63.5	0.3 / 57.4	32.9	0.7 / 61.3	34.9	81.0 / 81.0	39.7 / 79.7	32.4	4.8 / 79.7	60.7	2.5 / 79.1	**62.0**
Mean		/	/	34.9	/	43.0	/	/	52.6	/	65.7	/	**71.3***

Table 3. The results of ownership verification by SL-CNN [29], NTL [28], CUTI [29], NTL†, CUTI†, and IP-CLIP. O_{ua} represents the proposed weighted drop, while A_u and A_a denote the accuarcy for the domain with and without patch, respectively. The best performance is indicated by the numbers in bold. Statistical significance (p-value < 0.05 [4, 8]) is denoted with: *(IP-CLIP vs. others).

Authorized/Test	Amazon	Dslr	Webcam	$W_{ua}\uparrow$	$D_u\uparrow$	$D_a\downarrow$
Amazon	79.4 ⇒ 79.0	87.5 ⇒ 9.8	88.8 ⇒ 38.3	50.32	64.10	0.40
Dslr	83.8 ⇒ 23.3	95.7 ⇒ 95.3	98.8 ⇒ 64.3	44.89	47.50	0.40
Webcam	80.0 ⇒ 17.8	92.5 ⇒ 10.0	94.4 ⇒ 92.5	65.17	72.35	1.90
Mean		/		53.46	61.32	0.90

Table 4. The accuracy (%) of target-free IP-CLIP on the Office-31 [26]. The vertical/horizontal axis denotes the authorized/test domain.

to the authorized domain. Specifically, following the approach outlined in Sec. 4.3, we designate one domain as the original domain, to which we apply a specific watermark, resulting in the processed data being classified as the authorized domain. The unauthorized domain set is then formed by mixing the original domain, the domain generated from the original domain, and the generated domain with the watermark. During testing, the original domain and other unknown domains are used as the test set.

Tab. 6 and Tab. 7 present the experimental results of IP-CLIP and SOTA methods on the Office-31 [26], while results from additional datasets are shown in Tab. 6 (see *Supplementary Tab. 32-46* for further details). An interesting pattern emerges from the Tab. 7: in some domains, the A_u of NTL and CUTI outperform that of IP-CLIP, while their A_a is lower than that of IP-CLIP, and even in extreme cases is only one-third; Conversely, in certain cases, the A_a performance of NTL, CUTI, and IP-CLIP is comparable, but their A_u performance is worse. This demonstrates that relying on a single indicator (i.e., A_u and A_a) to assess IP protection is inadequate, highlighting the need for a comprehensive weighted metric $D_{ua} = A_a \cdot [A_a - A_u]$. As expected, IP-CLIP consistently achieves the highest D_{ua} across various domains, confirming that its generalization is effectively constrained to the authorized domain.

5. Conclusion

Protecting the intellectual property (IP) of visual language models (VLMs) like CLIP is a significant challenge in ar-

Datasets	Authorized Domain	$W_{ua}\uparrow$					$D_u\uparrow$					$D_a\downarrow$				
		NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP
Office-31 [26]	Amazon	0.56	4.69	11.90	25.60	**50.32**	7.80	13.30	17.25	36.65	**64.10**	7.05	7.05	1.90	3.10	**0.40**
	Dslr	6.88	6.83	36.72	38.83	**44.89**	9.40	9.35	43.90	43.30	**47.50**	2.30	2.30	3.90	1.90	**0.40**
	Webcam	2.90	2.95	45.80	30.95	**65.17**	8.60	5.45	50.95	33.60	**72.35**	5.45	2.35	1.60	**0.60**	1.90
	Mean	3.45	4.82	31.47	31.80	**53.46***	8.60	9.37	37.37	37.85	**61.32***	4.93	3.90	2.47	1.87	**0.90**
Office-Home-65 [27]	Art	0.10	-0.19	-0.71	-0.65	**4.82**	1.93	6.53	2.83	3.40	**12.07**	**1.80**	6.80	3.70	4.20	6.00
	Clipart	0.75	1.36	0.30	5.19	**14.88**	1.34	8.24	0.90	8.23	**19.83**	0.40	6.40	0.50	1.20	**0.00**
	Product	3.13	4.21	14.08	12.57	**23.67**	6.08	13.08	19.03	18.50	**30.40**	**2.60**	8.10	3.30	4.30	3.80
	RealWorld	2.39	3.72	13.07	3.82	**20.41**	2.83	8.83	17.67	5.50	**22.93**	**0.00**	4.20	2.70	1.20	0.20
	Mean	1.59	2.28	6.68	5.23	**15.95***	3.05	9.17	10.11	8.91	**21.31***	**1.20**	6.38	2.55	2.73	2.50
Mini-DomainNet [34]	Clipart	-3.25	-1.85	-0.89	2.24	**2.95**	11.80	5.30	3.50	7.07	**7.63**	17.30	8.00	4.60	4.30	**4.00**
	Painting	-0.52	0.27	0.39	0.21	**0.97**	7.53	3.87	**4.40**	3.57	3.93	8.50	3.40	3.90	3.30	**2.70**
	Real	2.60	2.05	4.46	5.86	**13.77**	5.73	6.00	9.37	8.93	**18.13**	2.60	3.50	4.20	**2.30**	2.50
	Sketch	2.44	-1.63	3.07	1.29	**3.74**	**14.53**	6.70	7.37	5.17	8.23	10.20	9.56	**3.40**	3.50	**3.40**
	Mean	0.32	-0.29	1.76	2.40	**5.36***	**9.90**	5.47	6.16	6.18	9.48	9.47	4.97	4.23	3.30	**3.07***

Table 5. W_{ua}, D_u, and D_a of target-free IP-CLIP, CUTI†, NTL†, CUTI and NTL. W_{ua} represents the proposed weighted drop, while D_u and D_a denote the drop rates for the unauthorized and authorized domains, respectively. The best performance is indicated by the numbers in bold. Statistical significance (p-value < 0.05 [4, 8]) is denoted with: *(IP-CLIP vs. others).

Dataset	Authorized Domain	$D_{ua}\uparrow$					$A_u\downarrow$					$A_a\uparrow$				
		NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP	NTL [28]	CUTI [29]	NTL† [28]	CUTI† [29]	IP-CLIP
Office-31 [26]	Amazon	1.63	27.95	15.67	29.26	**37.46**	5.21	**0.52**	37.43	20.83	3.53	15.63	53.13	62.50	**65.50**	63.00
	Dslr	9.23	72.92	39.25	54.47	**82.42**	4.69	**4.17**	50.50	36.53	9.77	32.81	87.50	92.80	94.30	**95.80**
	Webcam	11.82	40.01	54.59	40.56	**56.45**	**0.00**	37.00	21.30	30.60	15.53	34.38	84.40	**85.30**	80.80	83.30
	Mean	7.56	46.96	36.50	41.43	**58.78***	**3.30**	13.90	36.41	29.32	9.61*	27.60	75.01	80.20	80.20	**80.70**
Office-Home-65 [27]	Art	8.75	35.25	49.47	54.95	**60.12**	63.93	**1.04**	20.45	10.38	3.88	75.52	59.90	**81.30**	79.50	**79.50**
	Clipart	4.98	14.78	9.74	16.86	**26.52**	50.39	**0.72**	27.70	20.88	10.48	**58.85**	38.80	48.00	52.80	57.00
	Product	17.49	33.27	44.44	39.53	**57.74**	58.40	**0.78**	27.48	35.38	8.40	80.21	58.07	81.80	**83.00**	80.30
	RealWorld	15.83	3.15	51.50	62.87	**71.17**	64.97	31.32	19.20	7.83	**5.20**	83.85	39.32	82.00	83.30	**87.00**
	Mean	11.76	21.61	38.79	43.55	**53.89***	59.42	8.46	23.71	18.61	**6.99***	74.61	49.02	73.28	74.65	**75.95***
Mini-DomainNet [34]	Clipart	11.96	13.75	38.45	22.77	**50.88**	58.06	60.53	17.54	44.08	**7.35**	74.18	**78.13**	71.40	74.60	75.10
	Painting	7.47	6.47	32.78	24.48	**40.33**	58.26	45.15	24.18	34.73	**10.93**	69.08	56.58	**70.60**	69.80	69.20
	Real	21.08	22.62	35.66	33.56	**54.06**	57.03	58.43	37.90	34.83	**18.40**	82.57	**85.03**	81.60	77.90	83.30
	Sketch	7.72	7.00	38.66	48.18	**48.27**	58.47	57.24	16.55	9.45	**8.20**	69.57	67.60	71.00	**74.30**	73.70
	Mean	12.06	12.46	36.39	32.25	**48.39**	57.96	55.34	24.04	30.77	**11.22**	73.85	71.83	73.65	74.15	**75.33**

Table 6. D_{ua}, A_u, and A_a of authorization application IP-CLIP, CUTI†, NTL†, CUTI and NTL on the Office-31 [26]. D_{ua} represents the proposed weighted drop, while A_u and A_a denote the accuracy for the unauthorized and authorized domains, respectively. The best performance is indicated by the numbers in bold. Statistical significance (p-value < 0.05 [4, 8]) is denoted with: *(IP-CLIP vs. others).

Authorized/Test	Amazon	Dslr	Webcam	$D_{ua}\uparrow$	$A_u\downarrow$	$A_a\uparrow$
Amazon	4.5	3.3	2.8	37.46	3.53	63.00
Dslr	27.3	1.5	0.5	82.42	9.77	95.80
Webcam	31.0	4.3	11.3	56.45	15.53	83.30
Mean		/		58.78	9.61	80.70

Table 7. D_{ua}, A_u, and A_a of authorization application IP-CLIP on the Office-31 [26]. The vertical/horizontal axis denotes the authorized/test domain. D_{ua} represents the proposed weighted drop, while A_u and A_u denote the accuarcy of the unauthorized and test domains, respectively.

Figure 3. Several visualization examples of CLIP and IP-CLIP prediction results. Correct predictions are highlighted in green, while incorrect predictions are shown in red.

tificial intelligence. To address this, we propose IP-CLIP, a lightweight, prompt-based strategy that extracts image style and content for domain verification while preventing unauthorized feature transfers. Extensive experiments on cross-domain datasets demonstrate the effectiveness of our lightweight and easy-to-deploy IP-CLIP. Though designed for classification tasks, IP-CLIP can be extended to applications such as detection and image description. Future work will focus on enhancing generalization and adapting IP protection strategies to diverse model architectures. We believe our work will advance research in model IP protection and underscore its practical importance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 62136004, 62276130), the Key Research and Development Plan of Jiangsu Province (No. BE2022842), and H. Fu's A*STAR Central Research Fund.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[3] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *CVPR*, pages 24239–24250, 2024. 1, 3

[4] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006. 6, 7, 8

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, pages 702–703, 2020. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[7] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[8] Laurence Gillick and Stephen J Cox. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 532–535. IEEE, 1989. 6, 7, 8

[9] Jiyang Guan, Jian Liang, and Ran He. Are you stealing my model? sample correlation for fingerprinting deep neural networks. *Advances in Neural Information Processing Systems*, 35:36571–36584, 2022. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[11] Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *ICLR*, 2024. 3

[12] Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[13] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 2

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 2

[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, pages 4015–4026, 2023. 1

[19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[20] Junpei Liao, Liang Yi, Wenxin Shi, Wenyuan Yang, Yanmei Fang, and Xin Yang. Imperceptible backdoor watermarks for speech recognition model copyright protection. *Visual Intelligence*, 2(1):23, 2024. 1

[21] Erwan Le Merrer, Patrick Perez, and Gilles Tredan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32(13):9233–9244, 2020. 1

[22] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *CVPR*, pages 13430–13439, 2022. 1, 3

[23] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue. Fingerprinting deep neural networks globally via universal adversarial perturbations. In *CVPR*, pages 13430–13439, 2022. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 7

[25] Huali Ren, Anli Yan, Chong-zhi Gao, Hongyang Yan, Zhenxin Zhang, and Jin Li. Are you copying my prompt? protecting the copyright of vision prompt for vpaas via watermark. *arXiv preprint arXiv:2405.15161*, 2024. 1, 3

[26] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer, 2010. 5, 6, 7, 8

[27] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 5, 6, 7, 8

[28] Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. *arXiv preprint arXiv:2106.06916*, 2021. 1, 2, 3, 5, 6, 7, 8

[29] Lianyu Wang, Meng Wang, Daoqiang Zhang, and Huazhu Fu. Model barrier: A compact un-transferable isolation domain for model intellectual property protection. In *CVPR*, pages 20475–20484, 2023. 2, 3, 5, 6, 7, 8

[30] Lianyu Wang, Meng Wang, Huazhu Fu, and Daoqaing Zhang. Say no to freeloader: Protecting intellectual property of your deep model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 5

[31] Meng Wang, Tian Lin, Lianyu Wang, Aidi Lin, Ke Zou, Xinxing Xu, Yi Zhou, Yuanyuan Peng, Qingquan Meng, Yiming Qian, et al. Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications*, 14(1):6757, 2023. 1

[32] Mingfu Xue, Yushu Zhang, Jian Wang, and Weiqiang Liu. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2021. 1

[33] Guangtao Zeng and Wei Lu. Unsupervised non-transferable text classification. *arXiv preprint arXiv:2210.12651*, 2022. 3

[34] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE TIP*, 30:8008–8018, 2021. 5, 6, 7, 8

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 2

[37] B Zoph. Neural architecture search with reinforcement learning. *ICLR*, 2016. 1