TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery

Li Zhang[1,2] Shruti Agarwal[1] John Collomosse[1,3] Pengtao Xie[2] Vishal Asnani[1]
[1]Adobe Research, [2]University of California, San Diego, [3]DECaDE, University of Surrey
{liz042, p1xie}@ucsd.edu {shragarw, collomos, vasnani}@adobe.com

Abstract

Generative AI models pose a significant challenge to intellectual property (IP), as they can replicate unique artistic styles and concepts without attribution. While watermarking offers a potential solution, existing methods often fail in complex scenarios where multiple concepts (e.g., an object and an artistic style) are composed within a single image. These methods struggle to disentangle and attribute each concept individually. In this work, we introduce TokenTrace, a novel proactive watermarking framework for robust, multi-concept attribution. Our method embeds secret signatures into the semantic domain by simultaneously perturbing the text prompt embedding and the initial latent noise that guide the diffusion model's generation process. For retrieval, we propose a query-based TokenTrace module that takes the generated image and a textual query specifying which concepts need to be retrieved (e.g., a specific object or style) as inputs. This query-based mechanism allows the module to disentangle and independently verify the presence of multiple concepts from a single generated image. Extensive experiments show that our method achieves state-of-the-art performance on both single-concept (object and style) and multi-concept attribution tasks, significantly outperforming existing baselines while maintaining high visual quality and robustness to common transformations.

1. Introduction

Text-to-image foundation models [5, 20, 29], propelled by large-scale diffusion architectures [11, 28], have demonstrated unprecedented success in generating high-fidelity, complex visual content from natural language descriptions. This progress has brought generative AI to the forefront of creative industries [18]. However, this power introduces a critical challenge in proactive attribution: the need to embed imperceptible, robust watermarks into generated content to verify ownership and trace provenance [13, 53]. This task is fundamental to protecting the intellectual property of artists and creators whose unique styles and concepts are used to train these models. Recent state-of-the-art methods, such as ProMark [3], have pioneered a "proactive" approach, embedding imperceptible watermarks directly into the pixel space of images. These methods train the diffusion model to preserve these spatial watermarks, allowing a decoder to later detect them in generated images and establish a causal connection to the original training examples that contribute to the generations. Such invention is effective for the general concept attribution task.

Despite progress in this area, existing watermarking methods face notable limitations. Approaches that operate in the pixel domain are often fragile, failing to survive common transformations like compression or cropping [3, 32, 53]. While more recent methods embedding watermarks into the latent space offer greater robustness [4, 27, 43], they fail to address a significant challenge posed by the compositional nature of generative AI. These models frequently generate images by combining multiple, distinct concepts (e.g., a specific character rendered in a unique artistic style). Existing attribution methods, which embed a single holistic watermark, are not designed for this scenario; they cannot disentangle and independently attribute multiple concepts when their visual representations overlap in the final image. This failure to attribute composed concepts is a major gap in protecting creator IP. While some recent works attempt multi-concept attribution [38], they struggle with signal interference and lack a precise mechanism for targeted retrieval.

In response to the above challenges, we introduce TokenTrace, a proactive watermarking framework for robust, multi-concept attribution. We hypothesize that a watermark's robustness and specificity can be dramatically improved by associating it directly with the textual semantics of the concept it represents, a strategy inspired by the success of prompt-tuning in foundation models [19, 51]. This association is the key to solving the multi-concept challenge: instead of embedding competing signals into the limited pixel space, we link each concept's secret to its unique representation within the text embedding. This approach fundamentally avoids the problem of spatial overlap, as the signatures are separated in the textual semantic domain before generation begins. TokenTrace achieves this via a novel

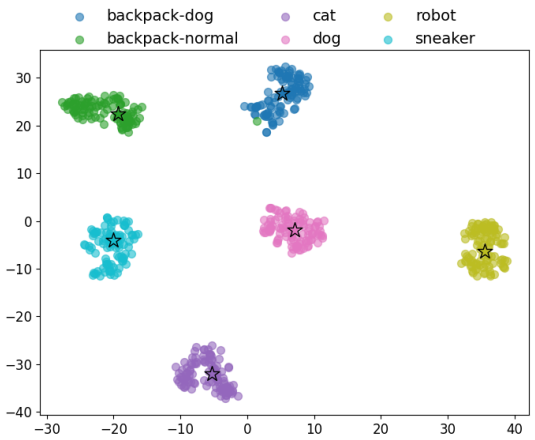

Figure 1. t-SNE visualization of predicted concept embeddings. The embeddings retrieved by TokenTrace module from DreamBooth-generated images (dots) form distinct, well-separated clusters around their ground-truth embeddings (stars), validating its ability to perform concept attribution.

dual-conditioning strategy. A concept's secret is processed by two parallel networks: a concept encoder that perturbs the text prompt embedding and a secret mapper that perturbs the initial latent noise. By embedding the signature in both the semantic and latent domains, the watermark is deeply integrated into the generative process, ensuring both robustness and this crucial concept-level separation. To retrieve the watermark, we propose a query-based TokenTrace module designed to reverse this watermark encoding, which takes the watermarked image and a query prompt as inputs to predict the corresponding concept secrets. The feasibility of this retrieval approach was validated in a preliminary experiment [1]; as shown in a t-SNE visualization (Figure 1), TokenTrace module successfully maps generated images back to distinct, well-separated clusters for each concept.

Our work makes three key contributions:

- We propose TokenTrace, a novel framework that embeds watermarks into both the text prompt and latent noise domains, intrinsically tying the watermark to the concept's semantics.
- We introduce a query-based module that can effectively disentangle and independently attribute multiple, overlapping concepts (such as objects and styles) from a single generated image.
- Our extensive experiments on both object and style concepts demonstrate that TokenTrace significantly outperforms state-of-the-art attribution baselines in both single-concept and multi-concept scenarios, all while maintaining high visual fidelity.

[1]Detailed settings for this preliminary experiment can be found in the supplementary material.

2. Related Works

Generative Customization The advent of large-scale text-to-image diffusion models has enabled remarkable image synthesis [28, 29, 31, 42, 46]. A significant and parallel line of research is generative customization, which focuses on personalizing these foundation models with novel, user-provided concepts [8, 21, 22, 44, 48]. Techniques such as Textual Inversion [14] and DreamBooth [29] allow models to learn a new concept from just a few example images, often by learning a new pseudo-token in the model's text-embedding space or leaning a new set of model parameters for a customized concept. While this customization empowers creators to inject their unique intellectual property (IP) into the generative process, it also exposes that IP to misuse [9, 36, 37]. These learned concepts can be extracted and repurposed without permission, which introduces a critical need for robust attribution methods designed specifically for these personalized concepts.

Watermarking and Concept Attribution for Generative Models To address provenance and attribution, various watermarking techniques for generative models have been proposed [13, 17, 23, 43, 45], which are broadly categorized as passive or proactive. Passive watermarking applies a signature after generation [15, 33, 52]. This approach, while simple, adds computational cost, is easily defeated by common transformations, and is thus insufficient for reliable attribution. In contrast, proactive watermarking, our focus, integrates the signature during the generative process [3, 17, 39], making it inherently more robust to transformations and the necessary choice for causal attribution. Existing proactive methods embed signatures in the pixel-domain [3, 32] or latent-domain [4, 45]. However, these traditional methods typically embed a single, content-agnostic watermark, failing to link it to specific semantic concepts, which leads to the challenges we address next.

Proactive Schemes Proactive schemes embed signals into the generative process for tasks like deepfake detection [40, 50], manipulated content localization [1, 2, 47], and concept attribution [3, 4]. While pixel-domain methods are fragile and fail when multiple concept spatially overlap, and latent-domain methods are robust but content-agnostic, a new category operates in the semantic domain [24], linking watermarks to textual representations. For instance, "Guarding Textual Inversion" [12] is designed for single-concept tracing and does not address compositional attribution. CustomMark [4] also customizes models via text prompts for multi-concept attribution but struggles to disentangle concepts that are spatially overlapping. Thus, a robust, query-based mechanism for disentangling composed concepts remains a significant and unaddressed challenge.

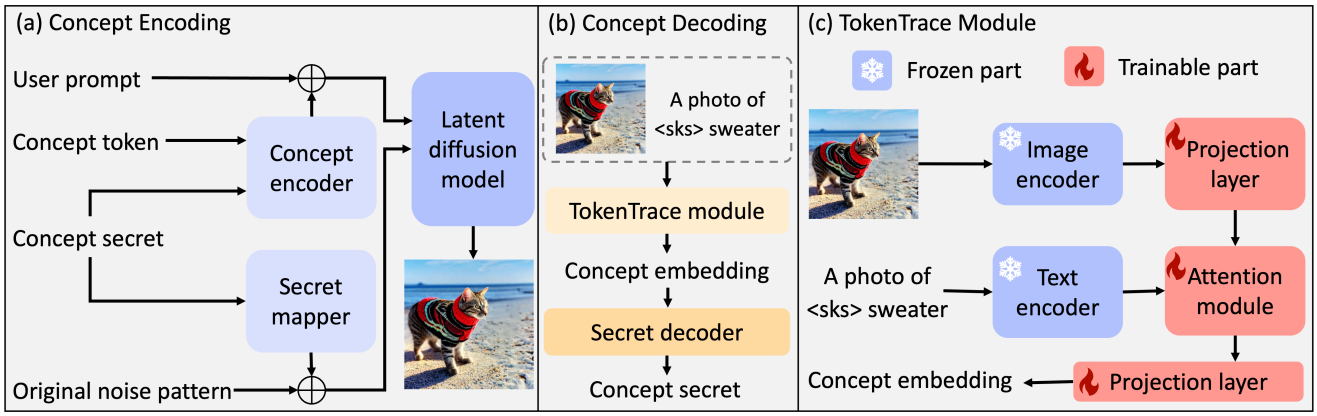

Figure 2. Overview of TokenTrace. (a) Concept encoding: A concept secret is fed into a concept encoder to perturb the targeted concept token and a secret mapper to perturb the initial noise, enabling a dual-conditioning of the latent diffusion model. (b) Concept decoding: The generated image and a query prompt are fed into the TokenTrace module to predict a concept embedding, which a secret decoder then translates back into the original concept secret for verification. (c) TokenTrace module: We use frozen image and text encoders, whose features are fused by trainable projection and attention layers to predict the final concept embedding.

3. Method

In this section, we present our method, TokenTrace, a proactive watermarking framework designed for robust, multi-concept attribution that can be used as a controlled-generation provenance tool. As illustrated in Figure 2, our framework is composed of two stages: (a) a concept encoding stage that embeds concept secrets into the generative process, (b) a concept decoding stage that retrieves the concept secrets that is indicated by a query prompt, and (c) the detailed architecture of our TokenTrace module.

3.1. Concept Encoding

Our encoding process employs a novel dual-conditioning strategy that embeds a watermark by simultaneously perturbing the generative inputs in both the semantic domain (textual embeddings) and the latent domain (initial noise). As shown in Figure 2(a), this stage involves two parallel networks that process a concept secret ($\mathcal{S}$):

- The concept encoder (f_{enc}) targets the specific concept token to be watermarked. A user prompt embedding E_{prompt} is a collection of token embeddings $\{e_1, \ldots, e_{\text{c}}, \ldots, e_{\text{k}}\}$, where e_{c} is the target concept token embedding. The encoder f_{enc} takes the concept secret ($\mathcal{S}$) and the corresponding concept token embedding (e_{c}) as input to generate a perturbation. This perturbation is fused (via element-wise addition) only with the target token embedding e_{c} to create a perturbed token embedding $\hat{e}_{\text{c}}$. The final perturbed prompt embedding, $\hat{E}_{\text{prompt}}$, is calculated by replacing the original token:

$$\begin{cases} \hat{e}_{\text{c}} = e_{\text{c}} + f_{\text{enc}}(e_{\text{c}}, \mathcal{S}) \\ \hat{E}_{\text{prompt}} = \{e_1, \ldots, \hat{e}_{\text{c}}, \ldots, e_{\text{k}}\} \end{cases} \tag{1}$$

- The secret mapper (f_{map}) takes only the concept secret ($\mathcal{S}$) as input and generates a structured perturb gaussian pattern. This pattern is fused with the original noise pattern (z_{T}) to create the perturbed initial noise, $\hat{z}_{\text{T}}$.

$$\hat{z}_{\text{T}} = z_{\text{T}} + f_{\text{map}}(\mathcal{S}) \tag{2}$$

Finally, the diffusion model (DM) is conditioned on both of these inputs, the prompt embedding with perturbed concept token embeddings ($\hat{E}_{\text{prompt}}$) and the perturbed noise ($\hat{z}_{\text{T}}$), to generate the final watermarked image (I_{wm}).

$$I_{\text{wm}} = DM(\hat{z}_{\text{T}}, \hat{E}_{\text{prompt}}) \tag{3}$$

By applying this dual-conditioning strategy, our method achieves deep integration. Specifically, by embedding the signature in both the semantic and latent domains, the watermark is fundamentally woven into the image's structure, making it far more robust to transformations than methods that only modify pixel space.

3.2. Concept Decoding

As shown in Figure 2(b), the decoding stage is a two-step pipeline designed to reliably retrieve the embedded watermark and confirm its origin. First, the watermarked image (I_{wm}) and a textual query (P_{query}) are fed into the TokenTrace module (f_{tt}). This textual query is a simple prompt, selected from a pre-defined set, that specifies which concept to retrieve. For example, to retrieve the <sks-object> concept, the query P_{query} would be "a photo of <sks-object>". The TokenTrace module then uses both inputs to predict the concept embedding ($\tilde{e}_{\text{c}}$).

$$\tilde{e}_{\text{c}} = f_{\text{tt}}(I_{\text{wm}}, P_{\text{query}}) \tag{4}$$

This predicted embedding is then passed to the secret decoder (f_{dec}), which we implement as a simple linear network. This decoder translates the high-dimensional embedding back into the original bit-secret, $\tilde{\mathcal{S}}$.

$$\tilde{\mathcal{S}} = f_{\text{dec}}(\tilde{e}_{\text{c}}) \tag{5}$$

This query-based pipeline provides a direct solution to the multi-concept challenge. By requiring a specific text query (P_{query}) to initiate retrieval, the system can selectively isolate and verify a single concept's signature. This mechanism allows it to successfully disentangle and attribute multiple, overlapping concepts (e.g., an object and a style) from a single image, a task that pixel-based methods cannot perform.

Importantly, the core of our retrieval mechanism is the TokenTrace module, which is architected to effectively fuse multi-modal information and predict a concept's original embedding from a generated image. As detailed in Figure 2(c), the module is designed for parameter-efficient finetuning by leveraging powerful frozen pre-trained encoders from CLIP [26] for robust feature extraction. The data flow is as follows:

1. The watermarked image I_{wm} is passed through the frozen Image encoder ($f_{\text{img\_enc}}$). Its output features are then fed into a trainable projection layer ($f_{\text{proj\_1}}$) to align their dimensionality with the text features.
2. The query prompt P_{query} is passed through the frozen Text encoder ($f_{\text{text\_enc}}$).
3. The aligned image features and the text features are fed into a trainable attention module (f_{attn}), which produces a fused, context-aware representation.
4. This fused representation is passed through a final trainable projection layer ($f_{\text{proj\_2}}$) to generate the predicted concept embedding, $\tilde{e}_{\text{c}}$.

This entire retrieval process is summarized by:

$$\begin{cases} F_{\text{img}} = f_{\text{proj\_1}}(f_{\text{img\_enc}}(I_{\text{wm}})) \\ F_{\text{text}} = f_{\text{text\_enc}}(P_{\text{query}}) \\ F_{\text{fused}} = f_{\text{attn}}(F_{\text{img}}, F_{\text{text}}) \\ \tilde{e}_{\text{c}} = f_{\text{proj\_2}}(F_{\text{fused}}) \end{cases} \tag{6}$$

The TokenTrace module is designed for high parameter efficiency by leveraging powerful, frozen CLIP encoders for feature extraction. This design, inspired by adapter-based finetuning [16], requires only a few lightweight projection and attention layers to be trainable. It allows the module to be adapted to new concepts quickly and scalably, avoiding the high computational cost and potential catastrophic forgetting associated with fully finetuning large-scale models.

3.3. Training Objective

Our training objective is to jointly optimize all trainable components: the concept encoder, secret mapper, the trainable projection and attention layers of the TokenTrace module, and the secret decoder. The model is trained using a composite loss function, $\mathcal{L}_{\text{total}}$, which is designed to simultaneously satisfy two competing goals: watermark retrieval accuracy and visual fidelity. This final loss is a weighted sum of four individual constraints:

- A cross-entropy loss between the original and predicted secrets to ensure the watermark can be accurately retrieved:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\mathcal{S}, \sigma(\hat{\mathcal{S}}_{\text{logits}})), \tag{7}$$

  where σ is the sigmoid function, $\mathcal{S}$ is the ground-truth, and $\hat{\mathcal{S}}_{\text{logits}}$ is the output logits from the secret decoder.
- A contrastive style descriptor (CSD) loss [34] to maintain high-level semantic consistency between the watermarked and clean images:

$$\mathcal{L}_{\text{CSD}} = 1 - \frac{\phi(I_{\text{clean}}) \cdot \phi(I_{\text{wm}})}{\|\phi(I_{\text{clean}})\| \|\phi(I_{\text{wm}})\|}, \tag{8}$$

  where ϕ represents the feature extractor, I_{clean} is the clean image, and I_{wm} is the watermarked image.
- A standard L2 loss to minimize perceptible, pixel-level differences between the watermarked and clean images, ensuring imperceptibility:

$$\mathcal{L}_{\text{L2}} = \|I_{\text{clean}} - I_{\text{wm}}\|_2^2. \tag{9}$$

- A regularization loss between the predicted and original concept embeddings to ensure the TokenTrace module's output is accurate:

$$\mathcal{L}_{\text{reg}} = \|e_{\text{c}} - \tilde{e}_{\text{c}}\|_2^2, \tag{10}$$

  where e_{c} and $\tilde{e}_{\text{c}}$ are the ground truth and predicted concept embeddings.

During training, the final loss is the weighted sum of these individual loss constraints, which are jointly optimized to balance watermark retrieval accuracy with visual fidelity:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{CSD}} + \lambda_3 \mathcal{L}_{\text{L2}} + \lambda_4 \mathcal{L}_{\text{reg}}, \tag{11}$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are the trade-off parameters to balance different kind of constrains. By using this composite loss function, we can jointly optimize for two competing objectives: retrieval accuracy (via $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{reg}}$) and visual fidelity (via $\mathcal{L}_{\text{L2}}$ and $\mathcal{L}_{\text{CSD}}$). This ensures the resulting watermark is both robustly detectable and imperceptible.

The end-to-end training procedure for TokenTrace is detailed in Algorithm 1 in the supplementary material. In summary, we jointly optimize all trainable parameters. In each training iteration, we sample a batch of data (including a clean image, prompts, secret, and ground-truth embedding) and perform the concept encoding step to generate a watermarked image. This image is then passed through the decoding pipeline to retrieve the secret. Finally, we compute the total composite loss $\mathcal{L}_{\text{total}}$ and update all trainable parameters via gradient descent. The inference (verification) process is a simple forward pass of the trained decoding modules (f_{tt} and f_{dec}), as described in Section 3.2.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed TokenTrace framework. We evaluate its effectiveness, scalability, and unique capability for multi-concept attribution across a diverse set of tasks: single style concept attribution, large-scale single object attribution, and compositional attribution using both customized and general concepts. We compare our framework against several state-of-the-art passive and proactive baselines and conduct a series of ablation studies to validate our key design choices.

4.1. Datasets

For the single style concept attribution task, we employ 23 distinct styles sourced from the WikiArt dataset [35], which is a collection of digital artworks sourced from the online art encyclopedia [2]. It is one of the most popular datasets for art-related computer vision tasks. For the single object concept attribution task, we use 1000 object classes from the ImageNet dataset [10], which is a foundational, large-scale dataset in computer vision. For our primary task of customized multi-concept attribution, we source pre-trained concept embeddings from the Stable Diffusion Textual Inversion Embeddings library [3]. This allows us to construct challenging prompts that combine multiple concepts, such as a specific object and a specific style, to test our model's disentanglement capabilities. We also evaluate general multi-concept attribution, as composing multiple customized textual inversion embeddings in a single prompt often degrades image quality. To isolate our attribution mechanism's performance from this generation artifact, we use standard, non-customized concepts (e.g., "a cat and a dog"). We construct a new evaluation benchmark by using ChatGPT [7] to generate a diverse set of prompts that involve the composition of multiple concepts. For each prompt, we also generate a list of target concepts that are presented in the prompt [4].

4.2. Experimental Settings

Baselines. We compare our method against a comprehensive suite of baselines, which are divided into two categories. The first, passive attribution, includes methods like ALADIN [30], CLIP [26], Abc [41], SSCD [25], and EKILA [6], which attempt to attribute a generated image after its creation by measuring visual correlation or similarity to known training data. The second, proactive attribution, includes our primary competitors, which embed a causal signature into the generative process itself. This category is represented by ProMark [3], a state-of-the-art method that embeds watermarks into the pixel domain, and CustomMark [4], a highly relevant baseline that, like our method, operates in the semantic domain by modifying text prompts.

[2]https://www.wikiart.org/

[3]https://cyberes.github.io/stable-diffusion-textual-inversion-models/

[4]Specific examples can be found in supplementary material.

Table 1. Comparison of attribution performance on the WikiArt and ImageNet datasets. TokenTrace achieves the highest accuracy across both datasets, outperforming all passive and proactive baselines. "Bit" represents bit accuracy (%) and "Att." represents attribution accuracy (%). Passive baselines do not have Bit Accuracy as they do not retrieve a secret bit-string.

Method	Type	WikiArt		ImageNet	
		Bit ↑	Att ↑	Bit ↑	Att ↑
ALADIN [30]	Passive	-	18.58	-	5.55
CLIP [26]	Passive	-	52.60	-	42.61
Abc [41]	Passive	-	56.03	-	53.51
SSCD [25]	Passive	-	45.34	-	25.50
EKILA [6]	Passive	-	43.03	-	30.98
ProMark [3]	Proactive	93.14	87.19	90.56	87.30
CustomMark [4]	Proactive	95.59	89.25	93.11	87.12
TokenTrace	**Proactive**	**98.33**	**91.67**	**95.82**	**90.43**

Metrics. To evaluate our framework, we use two categories of metrics. For attribution assessment, we measure attribution accuracy, the percentage of times the predicted bit-secret exactly matches the correct ground-truth secret, and bit accuracy, the average percentage of individual bits correctly predicted. For image quality assessment, we measure the visual fidelity between the clean and watermarked images using two embedding-space metrics: the CSD Score, which is the cosine similarity between CSD descriptors to assess style preservation, and the CLIP Score, which is the cosine similarity between CLIP image embeddings to assess overall semantic similarity.

Hyper-parameters. We implement TokenTrace in PyTorch. For the text-to-image latent diffusion model, we use the pre-trained Stable Diffusion 1.5 [28]. For the TokenTrace module, we utilize the pre-trained ViT-L/14 encoders from CLIP [26], which remain frozen during training. Unless stated otherwise, the watermark secret $\mathcal{S}$ is a randomly generated 16-bit binary vector. All training processes are conducted on 8 A100 NVIDIA GPUs with a batch size of 6 per GPU. All trainable parameters are optimized using the Adam optimizer with an initial learning rate of $1e-4$ and betas set to $(0.9, 0.999)$. We employ a step-based learning rate schedule with a gamma of 0.95. For our composite loss function (Eq. 11), the trade-off parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ were set to be $\{5, 5, 1, 1\}$, respectively. The model is trained for $10,000$ iterations. For the final evaluation, we use the checkpoint from the last training epoch.

4.3. Results and Analysis

Single Style Concept Attribution. We first evaluate our method's ability to attribute abstract concepts, such as artistic styles on the WikiArt [35] dataset. We compared TokenTrace against a series of both passive and proactive attribu-

(a) Qualitative Comparison for Artistic Style Concepts.

(b) Qualitative Comparison for Single Object Concepts.

Figure 3. Qualitative analysis of visual fidelity for watermarked images. (a) Results on abstract artistic style concepts from the WikiArt dataset. (b) Results on diverse object concepts from the ImageNet dataset. In both cases, the "Watermarked" images are visually indistinguishable from the "Clean" originals.

Table 2. Performance comparison for multi-concept attribution. The left plot shows results for attributing 2 customized concepts (one object and one style), while the right plot shows results for 4 general concepts. TokenTrace, consistently outperforms the CustomMark baseline in both bit accuracy, represented as "Bit", and attribution accuracy, represented as "Attrbution". Performance is further improved by applying prompt weighting (TokenTraceP). All accuracy metrics are in percent (%).

	Customized		General	
Method	Bit ↑	Attribution ↑	Bit ↑	Attribution ↑
CustomMark	92.47	85.14	78.93	72.78
TokenTrace	94.15	88.62	85.41	81.57
TokenTraceP	**96.83**	**90.53**	**90.33**	**86.08**

Prompt	A photo of <sks-fox>, art by <sks-Painting-style> style.		a photo of <sks-wasp> object, art by <sks-Cinematography-style> style.	
Image				
Concept	<sks-fox>	<sks-Painting-style>	<sks-wasp>	<sks-Cinematography-style>
Bit Acc.	100%	100%	100%	100%
Att. Acc.	100%	100%	100%	100%

Figure 4. Qualitative example of multi-customized concept prediction. We generate two images using a single prompt containing multiple watermarked concepts, and report the average bit accuracy and average attribution accuracy for each prompt.

tion baselines. The attribution accuracy and the bit prediction accuracy was calculated based on 100 generated images. As shown in Table 1, the results clearly demonstrate the superiority of our semantic watermarking approach for this complex task.

- Firstly, TokenTrace achieves a state-of-the-art Attribution Accuracy of 91.67%, significantly outperforming all other baselines.
- Secondly, TokenTrace surpasses all passive attribution methods by a large margin. For instance, it is substantially more accurate than standard CLIP retrieval (52.60%), confirming that simple similarity is insufficient for robust attribution.
- Thirdly, our method also outperforms other proactive watermarking techniques. It shows a clear improvement over both the pixel-based ProMark (87.19%) and the latent-based CustomMark (89.25%).

The qualitative results in Figure 3a further validate our method's performance. The "Watermarked" images successfully retain the intricate details and overall aesthetic of the original artistic styles. The visual fidelity between the "Clean" and "Watermarked" pairs is high. This confirms that watermarks inserted by TokenTrace is imperceptible, a crucial requirement for any practical attribution system.

Single Object Concept Attribution. We further assess evaluate the scalability and performance of TokenTrace on discrete objects, we conducted a large-scale experiment on 1000 object concepts sourced from the ImageNet [10] dataset. Similar from the style attribution task, the results presented in Table 1 clearly demonstrates the superiority of our approach. TokenTrace achieves a state-of-the-art attribution accuracy of 90.43% and a bit accuracy of 95.82%. It significantly outperforms all passive methods and shows a clear performance gain over the other proactive baselines, ProMark (87.30%) and CustomMark (87.12%). The qualitative results in Figure 3b further confirm the imperceptibility of our watermark. Across diverse concepts, the "Watermarked" images are visually indistinguishable from their "Clean" images, demonstrating that TokenTrace achieves robust attribution while maintaining high visual fidelity.

Multi-Concept Attribution. A key advantage of our framework is attributing multiple concepts within a single image. We first evaluated this on two customized concepts on the self-created concept pool of 20 concepts (10 objects and 10 artistic styles). As shown in Table 2, TokenTrace (88.62%) outperformed the CustomMark baseline (85.14%) upon the comparison of attribution accuracy, and an enhanced version, TokenTraceP (with prompt weighting) [5], further boosted this accuracy to 90.53%. We then tested a more complex four-concept scenario. Because composing multiple customized concepts degrades image quality, we constructed a new benchmark using ChatGPT [7] to generate a diverse set of prompts involving four distinguished concepts. The results in Table 2 show

[5]Details about TokenTraceP can be found in supplementary material.

Prompt	a detailed cat wearing a sweater, sitting in a forest, backlit by radiant sunlight, glowing rays.			
Image				
Concept	detailed	cat	sweater	rays
Bit Acc.	90.63%	100%	100%	84.38%
Att. Acc.	75%	100%	100%	75%

Figure 5. Qualitative example of multi-general concept prediction. We generate four images using a single prompt containing multiple watermarked concepts, and report the average bit accuracy and average attribution accuracy.

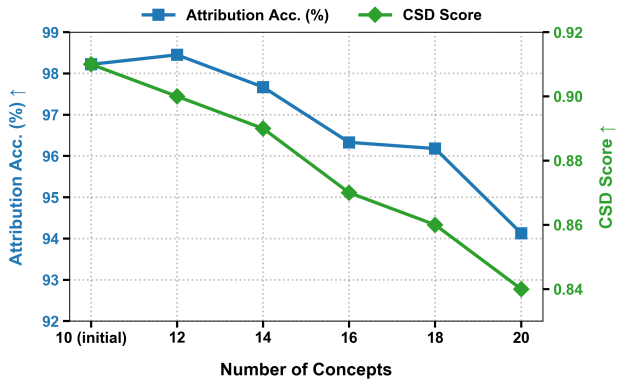

Figure 6. Performance on incremental concept learning. The plot shows the attribution accuracy (left y-axis, blue squares) and CSD score (right y-axis, green diamonds) as new concepts are sequentially added to a pre-trained model. The results demonstrate that both attribution accuracy and visual fidelity remain high, showing the method's efficiency in real-world applications.

our advantage becomes even more pronounced: TokenTrace (81.57%) dramatically outperformed CustomMark (72.78%) upon attribution accuracy, while TokenTraceP again achieved the highest accuracy (86.08%).

Finally, Figure 4 and Figure 5 provides qualitative proof of this disentanglement. Specifically, from a single image generated with four watermarked concepts (e.g., "detailed cat wearing a sweater... glowing rays"), our query-based module successfully retrieves the correct, independent secret for each concept, retrieving 'cat' and 'sweater' with 100% accuracy and 'detailed' and 'rays' with high accuracy. This confirms our query-based mechanism solves concept overlap and scales to complex prompts. More qualitative examples can be found in the supplementary material.

Sequential Learning. To evaluate our framework's ability to add new concepts without full retraining, we perform a sequential learning experiment on the WikiArt dataset. We first train an initial TokenTrace model on 10 style concepts, then sequentially introduce 2 new concepts at a time up to 20. Critically, we do not retrain from scratch but instead finetune the existing model with only 10% additional iterations per step, reflecting a real-world scenario. The results in Figure 6 demonstrate high efficiency and strong resistance to catastrophic forgetting. Both Attribution Accuracy (starting at 98.22% and ending at 94.13%) and CSD Score (starting at 0.91 and ending at 0.84) remain high, confirming that our parameter-efficient design can be dynamically updated, making it a scalable and practical solution.

Table 3. Ablation study for the objective function on the WikiArt dataset. Our full model ("All") significantly outperforms variants where individual loss components (CSD, Latent L2, Image L2) are removed, demonstrating that each component is necessary for achieving both high attribution accuracy and visual fidelity.

Attribution Performance (%)		
Method	**Bit Acc. ↑**	**Attribution Acc. ↑**
No CSD	91.81	83.75
No L2 (latent)	96.03	88.52
No L2 (Image)	93.65	86.37
All	**98.33**	**91.67**
Image Quality		
Method	**CLIP Score ↑**	**CSD Score ↑**
No CSD	0.73	0.65
No L2 (latent)	0.82	0.76
No L2 (Image)	0.81	0.73
All	**0.87**	**0.82**

4.4. Ablation Studies

Ablation on Loss Components. Our composite loss function ($\mathcal{L}_{\text{total}}$), as defined in Eq. 11, is designed to jointly optimize for retrieval accuracy and visual fidelity. To demonstrate the contribution of each loss term, we train our model from scratch with different components of the objective function removed. The full model ("All") is trained with all loss components, including $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{CSD}}$, $\mathcal{L}_{\text{L2}}$, and $\mathcal{L}_{\text{reg}}$. As shown in Table 3, our complete model ("All") achieves the best performance. Removing the CSD loss ("No CSD") causes the most significant performance drop. This finding confirms that the $\mathcal{L}_{\text{CSD}}$ term is critical for maintaining the high-level semantic and style consistency required for accurate retrieval. Removing the image-space L2 loss ("No L2 (Image)") also leads to a notable decrease in performance, with attribution accuracy dropping to 86.37% and the CSD Score falling to 0.73. The latent-space L2 ("No L2 (Latent)") also contributes, with its removal dropping accuracy to 88.52%.

Ablation on Bit Secret Length. We next investigate the impact of the length of the bit secret $\mathcal{S}$, on both attribution accuracy and visual fidelity. We trained separate models on the WikiArt dataset using secret lengths of 5, 8, 16 (our default), 32, and 64 bits. As shown in Figure 7, we observe a clear and expected trade-off between information capacity and performance. Specifically, The model achieves the highest attribution accuracy (94.38%) with a 5-bit secret, as this smaller signal is easiest to embed and retrieve. As the bit length increases, all metrics show a graceful degradation. When moving from our 16-bit default to a 64-bit secret, the

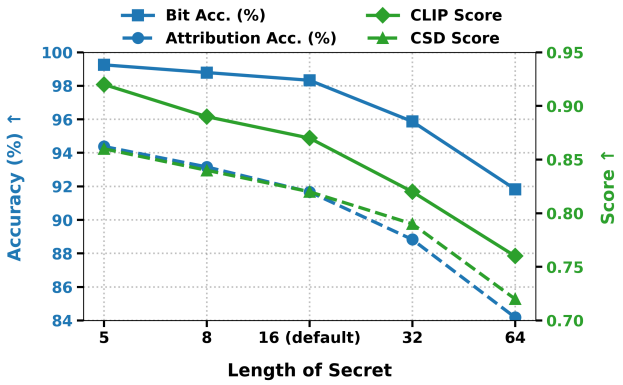

Figure 7. Ablation study for the length of the bit secret on the WikiArt dataset. Performance shows a graceful trade-off between capacity and accuracy. Our 16-bit default provides the best balance of high attribution accuracy and visual fidelity.

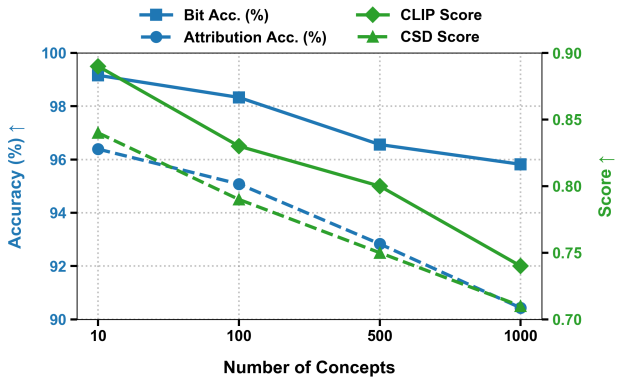

Figure 8. Ablation study for the number of concepts on the ImageNet dataset. Performance remains high and degrades gracefully as the number of concepts scales from 10 to 1000, demonstrating the high scalability of our method.

attribution accuracy drops from 91.67% to 84.18%, and the CSD score falls from 0.82 to 0.72. This trend is logical, as embedding a larger and more complex signal (64 bits) into the image is an inherently harder optimization problem and is more likely to interfere with the image's original semantic content, thus slightly reducing visual fidelity. These findings validate our choice of 16 bits as the default, as it provides an balance, offering a sufficiently large space for unique signatures while maintaining both high attribution accuracy and high visual quality.

Ablation on the Number of Concepts. We next evaluate our framework's scalability by training models on the ImageNet dataset with concept pools of increasing size: 10, 100, 500, and 1000 concepts. The results in Figure 8 demonstrate a very high level of performance and graceful degradation. As the concept library scales from 10 to 1000 concepts, the attribution accuracy remains exceptionally high, dropping only 6 percentage points (96.39% $\rightarrow$ 90.43%). Similarly, bit accuracy shows strong resilience (99.16% $\rightarrow$ 95.82%), and visual fidelity scales well, with the CSD score dropping only from 0.84 to 0.71. These findings demonstrate that TokenTrace is highly scalable and can robustly handle a large library of distinct concepts, confirming its practicality for real-world applications.

Robustness to Image Distortions. Finally, to validate our watermark's resilience, we conducted a robustness analysis by applying common image distortions to watermarked WikiArt images before retrieval, simulating real-world re-encoding and editing. The results in Table 4 demonstrate that our dual-conditioning strategy creates a highly robust watermark. The method shows exceptional resilience, maintaining over 90.04% attribution accuracy against Rotation and over 87% against both JPEG compression and a targeted Adversarial Attack. Performance also degrades gracefully for more severe transformations, with CropAndResize (86.57%), GaussianBlur (84.81%), ColorJitter (83.22%), and GaussianNoise (82.94%) all retaining over 82% accuracy. These results confirm that by embedding the signature in both the semantic and latent domains, our watermark is a deeply integrated signal, allowing for reliable attribution even after significant image alterations.

Table 4. Ablation study for different types of distortions on the WikiArt dataset. Our method demonstrates high robustness, maintaining strong attribution accuracy. The detailed description of "Adversarial Attack" can be found in supplementary material.

Method	Bit Acc. ↑	Attribution Acc. ↑
No distortion	98.33	91.67
JPEG	94.68	88.20
Rotation	96.21	90.04
CropAndResize	93.28	86.57
GaussianBlur	91.32	84.81
GaussianNoise	89.18	82.94
ColorJitter	89.62	83.22
Sharpness	92.71	86.54
Adversarial Attack [49]	94.08	87.17

5. Conclusion

In this paper, we introduced TokenTrace, a novel proactive watermarking framework that solves the challenging task of multi-concept attribution. By embedding a secret into both the textual semantic and latent domains, our method creates a robust signature that avoids the spatial overlap problem plaguing pixel-based methods. Our key contribution is the query-based TokenTrace module, which can successfully disentangle and independently retrieve signatures for multiple, composed concepts from a single image. Extensive experiments demonstrate that TokenTrace significantly outperforms state-of-the-art baselines on single-concept (style and object) and multi-concept tasks, all while maintaining high visual fidelity and robustness to distortions.

References

[1] Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. Proactive image manipulation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15386–15395, 2022. 2

[2] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12343–12352, 2023. 2

[3] Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. ProMark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10802–10811, 2024. 1, 2, 5

[4] Vishal Asnani, John Collomosse, Xiaoming Liu, and Shruti Agarwal. Custommark: Customization of diffusion models for proactive attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1512–1522, 2025. 1, 2, 5

[5] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025. 1

[6] Kar Balan, Shruti Agarwal, Simon Jenni, Andy Parsons, Andrew Gilbert, and John Collomosse. EKILA: Synthetic media provenance and attribution for generative art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 913–922, 2023. 5

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5, 6

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 2

[9] Simon Chesterman. Good models borrow, great models steal: intellectual property rights and generative ai. *Policy and Society*, 44(1):23–37, 2025. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 1

[12] Weitao Feng, Jiyan He, Jie Zhang, Tianwei Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Catch you everything everywhere: Guarding textual inversion via concept watermarking. *arXiv preprint arXiv:2309.05940*, 2023. 2

[13] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 1, 2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[15] Emir Ganic and Ahmet M Eskicioglu. Robust DWT-SVD domain image watermarking: embedding data in all frequencies. In *Proceedings of the 2004 Workshop on Multimedia and Security*, pages 166–174, 2004. 2

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 4

[17] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37:3937–3963, 2024. 2

[18] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 47(6):4409–4437, 2025. 1

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 1

[20] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature medicine*, 30(4):1154–1165, 2024. 1

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[22] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 2

[23] Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20937–20946, 2025. 2

[24] Hong-Hanh Nguyen-Le, Van-Tuan Tran, Thuc Nguyen, and Nhien-An Le-Khac. A survey on proactive deepfake defense: Disruption and watermarking. *ACM Computing Surveys*, 58 (5):1–37, 2025. 2

[25] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 5

[26] 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. 4, 5

[27] Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pages 118–136. Springer, 2024. 1

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 5

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 2

[30] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. ALADIN: All layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11926–11935, 2021. 5

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[32] Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2

[33] Amit Kumar Singh, Nomit Sharma, Mayank Dave, and Anand Mohan. A novel technique for digital image watermarking in spatial domain. In *2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing*, pages 497–501. IEEE, 2012. 2

[34] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 4

[35] Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2018. 5

[36] Saliltorn Thongmeensuk. Rethinking copyright exceptions in the era of generative ai: Balancing innovation and intellectual property protection. *The Journal of World Intellectual Property*, 27(2):278–295, 2024. 2

[37] Kalpana Tyagi. Copyright, text & data mining and the innovation dimension of generative ai. *Journal of Intellectual Property Law & Practice*, 19(7):557–570, 2024. 2

[38] Guanjie Wang, Zehua Ma, Chang Liu, Xi Yang, Han Fang, Weiming Zhang, and Nenghai Yu. Must: Robust image watermarking for multi-source tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5364–5371, 2024. 1

[39] Hanyi Wang, Han Fang, Shi-Lin Wang, and Ee-Chien Chang. ROAR: Reducing inversion error in generative image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19742–19751, 2025. 2

[40] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3546–3555, 2021. 2

[41] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7192–7203, 2023. 5

[42] Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20906–20915, 2025. 2

[43] Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8213–8224, 2025. 1, 2

[44] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2

[45] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024. 2

[46] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024. 2

[47] Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3008–3018, 2025. 2

[48] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524, 2025. 2

[49] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024. 8

[50] Yuan Zhao, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, and Xin Yu. Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4602–4611, 2023. 2

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1

[52] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. 2

[53] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded adversarial examples for copyright protection against diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24420–24430, 2024. 1