# TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery

Supplementary Material

## 1. Detailed Algorithm for TokenTrace

The end-to-end training procedure for TokenTrace is summarized in Algorithm 1. We jointly optimize all trainable parameters, denoted by $\{\theta_{\text{enc}}, \theta_{\text{map}}, \theta_{\text{pb}}, \theta_{\text{dec}}\}$. In each training iteration, we sample a batch of data containing a clean image ($I_{\text{clean}}$), its corresponding prompts ($P_{\text{user}}$) that contains the concept prompt ($P_{\text{concept}}$), a secret ($\mathcal{S}$), and the ground-truth concept embedding ($E_{\text{concept}}$). We first perform the concept encoding step to generate the perturbed inputs and produce a watermarked image. Then, we pass this image 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.

**Algorithm 1** TokenTrace Training

**Require:** Trainable parameters $\theta_{\text{enc}}, \theta_{\text{map}}, \theta_{\text{pb}}, \theta_{\text{dec}}$; Loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$

**for** each training step **do**
  **1. Concept Encoding & Generation**
  Sample batch ($I_{\text{clean}}$, $P_{\text{user}}$, $P_{\text{concept}}$, $\mathcal{S}$, $E_{\text{concept}}$)
  $E_{\text{prompt}} \leftarrow \text{Embed}(P_{\text{user}})$
  $\hat{E}_{\text{prompt}} \leftarrow E_{\text{prompt}} + f_{\text{enc}}(E_{\text{concept}}, \mathcal{S})$
  Sample initial noise $z_{\text{T}} \sim \mathcal{N}(0, 1)$
  $\hat{z}_{\text{T}} \leftarrow z_{\text{T}} + f_{\text{map}}(\mathcal{S})$
  $I_{\text{wm}} \leftarrow DM(\hat{z}_{\text{T}}, \hat{E}_{\text{prompt}})$
  **2. Concept Decoding & Verification**
  $P_{\text{query}} \leftarrow P_{\text{concept}}$ ▷ or $P_{\text{user}}$ as query
  $\tilde{E}_{\text{concept}} \leftarrow f_{\text{pb}}(I_{\text{wm}}, P_{\text{query}})$
  $\tilde{s}_{\text{logits}} \leftarrow f_{\text{dec}}(\tilde{E}_{\text{concept}})$
  **3. Loss Calculation**
  $\mathcal{L}_{\text{BCE}} \leftarrow \text{BCE}(\mathcal{S}, \sigma(\tilde{s}_{\text{logits}}))$
  $\mathcal{L}_{\text{L2}} \leftarrow \|I_{\text{clean}} - I_{\text{wm}}\|_2^2$
  $\mathcal{L}_{\text{CSD}} \leftarrow 1 - \text{sim}(\phi(I_{\text{clean}}), \phi(I_{\text{wm}}))$
  $\mathcal{L}_{\text{reg}} \leftarrow \|E_{\text{concept}} - \tilde{E}_{\text{concept}}\|_2^2$
  $\mathcal{L}_{\text{total}} \leftarrow \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{CSD}} + \lambda_3 \mathcal{L}_{\text{L2}} + \lambda_4 \mathcal{L}_{\text{reg}}$
  **4. Optimization**
  Update $\theta_{\text{enc}}, \theta_{\text{map}}, \theta_{\text{pb}}, \theta_{\text{dec}}$ using $\nabla_\theta \mathcal{L}_{\text{total}}$
**end for**

For the inference there are two processes: watermark generation, as shown in Algorithm 2, and concept attribution, as shown in Algorithm 3. The watermarking process begins during image generation using a dual-conditioning strategy. A unique binary secret $\mathcal{S}$ is assigned to the target concept. This secret is processed by two parallel networks: a concept encoder, which generates a perturbation embedding that is fused with the user's text prompt embedding, and a secret mapper, which generates a structured noise pattern that is fused with the initial gaussian noise. The latent diffusion model is then conditioned on these two perturbed inputs, modifying both the textual semantic and latent domains, to synthesize the final watermarked image $I_{\text{wm}}$, deeply integrating the signature into the content. Concept attribution is performed as a targeted retrieval operation. Given a suspect image $I_{\text{wm}}$ and a specific concept to verify, we construct a corresponding textual query $P_{\text{query}}$ (e.g., "a photo of <sks-object>"). These inputs are fed into the trained TokenTrace module, where frozen CLIP encoders extract multi-modal features that are aligned and fused by learned projection and attention layers to predict the target concept embedding, $\tilde{E}_{\text{concept}}$. This predicted embedding is passed through the secret decoder to recover the binary secret $\tilde{s}$. Attribution is confirmed by comparing this retrieved sequence against the ground-truth secret; a high bit-match rate provides cryptographic-like proof that the specific concept was used during generation.

**Algorithm 2** TokenTrace Watermarking (Generation)

**Require:** Trained parameters $\theta_{\text{enc}}, \theta_{\text{map}}$
**Require:** User prompt $P_{\text{user}}$, Concept prompt $P_{\text{concept}}$, Secret $\mathcal{S}$
**Ensure:** Watermarked Image $I_{\text{wm}}$

**1. Semantic Perturbation**
$E_{\text{prompt}} \leftarrow \text{Embed}(P_{\text{user}})$ ▷ Get initial prompt embeddings
$\Delta E \leftarrow f_{\text{enc}}(E_{\text{concept}}, \mathcal{S})$ ▷ Generate concept perturbation
$\hat{E}_{\text{prompt}} \leftarrow E_{\text{prompt}} + \Delta E$ ▷ Fuse perturbation with target token
**2. Latent Perturbation**
Sample initial noise $z_{\text{T}} \sim \mathcal{N}(0, 1)$
$\Delta z \leftarrow f_{\text{map}}(\mathcal{S})$ ▷ Generate noise perturbation
$\hat{z}_{\text{T}} \leftarrow z_{\text{T}} + \Delta z$ ▷ Fuse perturbation with initial noise
**3. Generation**
$I_{\text{wm}} \leftarrow DM(\hat{z}_{\text{T}}, \hat{E}_{\text{prompt}})$ ▷ Generate image using diffusion model
**return** $I_{\text{wm}}$

## 2. Details of TokenTraceP

This section provides the technical details for our enhanced model variant, TokenTraceP, which is featured in our multi-concept attribution experiments.

In our multi-concept experiments, we evaluate the model's ability to attribute multiple concepts (e.g., an object and a style) composed in a single prompt. A well-known

**Algorithm 3** TokenTrace Inference (Attribution)

**Require:** Trained parameters $\theta_{\text{tt}}, \theta_{\text{dec}}$
**Require:** Input image $I_{\text{wm}}$, Query prompt $P_{\text{query}}$, Ground-truth secret $\mathcal{S}_{\text{GT}}$
**Ensure:** Retrieved Secret $\tilde{s}$, Attribution Decision

**1. Concept Decoding**
$\tilde{E}_{\text{concept}} \leftarrow f_{\text{tt}}(I_{\text{wm}}, P_{\text{query}})$ ▷ Predict concept embedding using query
$\tilde{s}_{\text{logits}} \leftarrow f_{\text{dec}}(\tilde{E}_{\text{concept}})$ ▷ Predict raw logits
**2. Secret Recovery**
$\tilde{s}_{\text{prob}} \leftarrow \sigma(\tilde{s}_{\text{logits}})$ ▷ Apply Sigmoid function
$\tilde{s} \leftarrow (\tilde{s}_{\text{prob}} > 0.5)$ ▷ Binarize to obtain retrieved bit-secret
**3. Verification**
$Score \leftarrow \text{BitAccuracy}(\tilde{s}, \mathcal{S}_{\text{GT}})$ ▷ Compare with ground truth

challenge in generative customization is "concept overpowering" [2, 6], where one concept in a prompt (typically the object) visually dominates the generation, while the other (typically the style) is rendered less faithfully. This weak rendering of the style concept is not just a visual quality issue; it directly impacts attribution. A poorly-rendered style provides a weaker signal for our TokenTrace module to detect, leading to lower retrieval accuracy for that concept's secret.

To address this, our enhanced TokenTraceP variant integrates prompt weighting [1], a common technique in the diffusers library to "increase or decrease the scale of the text embedding vector" for a specific concept. This method is applied only during the generation of the watermarked image ($I_{\text{wm}}$). In our standard TokenTrace encoding, the final watermarked text embedding $\hat{E}_{\text{prompt}}$ is a collection of tokens, including the perturbed object token $\hat{e}_{\text{object}}$ and the perturbed style token $\hat{e}_{\text{style}}$:

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

For the TokenTraceP variant, we apply an additional scaling factor $\alpha > 1.0$ (we set $\alpha$ equals to 1.1 by default in the experiments conducted in main paper) specifically to the style token's embedding after our perturbation. This new, weighted style token $\hat{e}_{\text{style\_P}}$ is calculated as:

$$\hat{e}_{\text{style\_P}} = \alpha \times \hat{e}_{\text{style}} \quad (2)$$

The final prompt embedding $\hat{E}_{\text{prompt\_P}}$ sent to the diffusion model then uses this amplified style token:

$$\hat{E}_{\text{prompt\_P}} = \{e_1, \ldots, \hat{e}_{\text{object}}, \ldots, \hat{e}_{\text{style\_P}}, \ldots, e_{\text{k}}\} \quad (3)$$

This re-weighting forces the diffusion model's cross-attention layers to focus more on the style concept, ensuring it is rendered more robustly and faithfully alongside the dominant object.

[1] *https://huggingface.co/docs/diffusers/v0.21.0/en/using-diffusers/weighted_prompts*

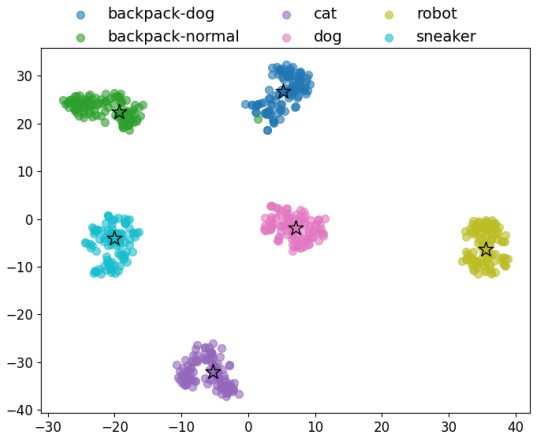



Figure 1. t-SNE visualization of predicted concept embeddings. The plot shows that embeddings retrieved by our TokenTrace module from DreamBooth-generated images (dots) form distinct, well-separated clusters around the corresponding ground-truth embeddings (stars), validating its ability to perform concept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

## 3. Preliminary explorations of TraceToken module

Before training our full end-to-end framework, we conducted a preliminary experiment to validate the core hypothesis of our method: whether the TokenTrace module can learn to extract distinct, coherent concept embeddings from generated images. To isolate the module's retrieval capability, we first trained the TokenTrace module using real images from the DreamBooth dataset [2]. This dataset contains sets of images for various unique concepts (e.g., "cat," "dog," "sneaker"). After this initial training, we used the official DreamBooth [6] model itself to generate 100 novel images for each of these concepts. We then fed these generated images into our trained TokenTrace module to predict their corresponding concept embeddings. Finally, to visualize the high-dimensional embedding space, we applied t-SNE [4] to reduce the dimension of the predicted embeddings.

The t-SNE visualization of the predicted embeddings is shown in Figure 1. The results clearly demonstrate that the TokenTrace module is highly effective at this retrieval task. As can be observed:

[2] https://github.com/google/dreambooth

- Embeddings predicted from images of the same concept (dots) form tight, well-defined clusters around their corresponding ground-truth embeddings (stars, derived from real images).
- Embeddings from different concepts (e.g., 'sneaker' vs. 'robot') are well-separated in the feature space, forming distinct and non-overlapping distributions.

This successful separation confirms that our TokenTrace module can effectively learn to map generated images back to their underlying conceptual representations. This preliminary result validated our hypothesis that using a predicted concept embedding as the basis for secret retrieval is a feasible and promising approach for concept attribution.

## 4. Comparison with Zero-Shot CLIP Baseline

We conducted an ablation study to compare with zero-shot CLIP [5] baseline. Specifically, this zero-shot CLIP baseline utilizes a standard, non-watermarked generative process followed by a passive semantic attribution stage. First, an image is synthesized using an original, frozen text-to-image diffusion model driven by a user prompt that contains specific concepts without any proactive semantic or latent perturbations. In the attribution phase, the generated image is processed by a pre-trained CLIP model to extract a high-dimensional image embedding. This embedding is then compared against a candidate pool of text embeddings representing potential concepts within the library, using cosine similarity to identify the most likely top-k matches. "K" represents the number of concepts. We conducted this ablation study on ImageNet dataset for 2 customized concepts attribution, and on our self-built 4 general concepts attribution benchmark.

As shown in Table 1, there is a significant performance gap between the original CLIP baseline and TokenTrace on the multi-concept attribution scenario. To further investigate this gap, we analyzed the separate results of CLIP on the customized datasets. CLIP achieved 42.50% attribution accuracy for the object concepts, while performance dropped further to 25.54% for the style concepts. The experiments confirm that while models like CLIP are useful for general classification, they are insufficient for the precise task of concept attribution.

Table 1. Attribution accuracy comparison against CLIP baseline.

| Method | Custom Acc (%) | General Acc (%) |
|---|---|---|
| Zero-Shot CLIP | 38.21 | 35.86 |
| TokenTrace | 88.62 | 81.57 |

## 5. Ablation on the Query Sensitivity

To quantify robustness to query phrasing, we first conducted a sensitivity analysis on our general concept benchmark by replacing the original ground-truth tokens with GPT-generated synonyms (e.g., querying "bunny" instead of "rabbit") during the retrieval phase. And, we also evaluated the system's specificity using a customized dataset where images were generated with personalized identifiers (e.g., "sks-dog"). During inference, we queried these specific instances with the broad class token (e.g., "dog"). From the results shown in Table 2, we observed that the, in the general concept dataset, method maintains comparable high attribution accuracy for these close synonyms. This stability arises because semantically similar terms map to adjacent points within the CLIP latent space, ensuring that the synonym's embedding remains within the "success cluster" of the original concept's manifold. However, we observed a significant and expected drop in the customized concept dataset. This degradation is a desirable safety feature rather than a failure; it demonstrates that the method correctly distinguishes between a specific user's personalized concept and a general category. By failing to attribute the generic query to the specific instance, TokenTrace prevents false accusations in copyright scenarios, ensuring that a watermark bound to a unique subject is not inadvertently triggered by broad, unrelated queries.

Table 2. Attribution robustness to synonym query variations.

| Input Token | Custom Acc (%) | General Acc (%) |
|---|---|---|
| Synonym | 50.18 | 80.63 |
| Original | 88.62 | 81.57 |

## 6. General Multi-Concept Attribution Dataset

In our main experiments, we evaluate multi-concept attribution on customized concepts (e.g., specific objects and styles) learned via textual inversion. A well-known limitation of this approach is that image quality and concept coherence degrade significantly when more than two customized embeddings are composed in a single prompt. To test the scalability of our attribution mechanism on more complex prompts (e.g., four distinct concepts) without this confounding generation-quality artifact, we created a new benchmark using general, non-customized concepts. This dataset allows us to evaluate the core hypothesis: can TokenTrace disentangle and retrieve multiple, independent secrets from a single, coherently-generated image?

We constructed the benchmark using ChatGPT with a strict template designed to produce "naturalistic scene descriptions" rather than simple keyword lists. The model was instructed to generate prompts containing exactly distinct concepts from diverse semantic categories to ensure the benchmark covers a wide range of visual manifolds rather than being biased toward a specific domain. Several examples are presented in the supplementary document. To ensure concepts are unambiguously defined under tokenization, we enforced two rigorous constraints during generation:

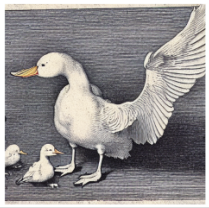
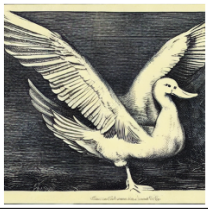
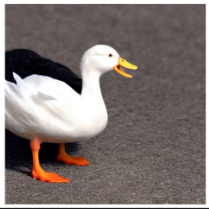
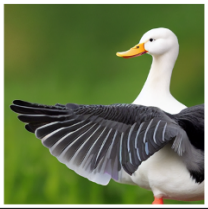
| Prompt | a photo of <sks-Aflac-duck> object, art by <sks-durer-style> style. | | a photo of <sks-Aflac-duck> object, art by <sks-hanfu-anime-style> style. | |
|---|---|---|---|---|
| Image |  |  |  |  |
| Concept | <sks-Aflac-duck> | <sks-durer-style> | <sks-Aflac-duck> | <sks-hanfu-anime-style> |
| Bit Acc. | 100% | 100% | 100% | 100% |
| Att. Acc. | 100% | 100% | 100% | 100% |

Figure 2. Qualitative example of multi-customized concept prediction. We generate two images for each prompt containing multiple watermarked concepts, and report the average bit accuracy and average attribution accuracy.

| Prompt | a photo of <sks-cheburashka> object, art by <sks-hanfu-anime-style> style. | | a photo of <sks-babau> object, art by <sks-hanfu-anime-style> style. | |
|---|---|---|---|---|
| Image |  |  |  |  |
| Concept | <sks-cheburashka> | <sks-hanfu-anime-style> | <sks-babau> | <sks-hanfu-anime-style> |
| Bit Acc. | 100% | 100% | 100% | 100% |
| Att. Acc. | 100% | 100% | 100% | 100% |

Figure 3. Qualitative example of multi-customized concept prediction. We generate two images for each prompt containing multiple watermarked concepts, and report the average bit accuracy and average attribution accuracy.

- Single-Word Constraint: We restricted every target concept to be exactly one word. This constraint eliminates ambiguity regarding multi-token boundaries or subword embedding averaging, ensuring a deterministic 1-to-1 mapping between the concept and its CLIP token embedding.
- Semantic Distinctiveness: We mandated that the concepts within a single prompt must be semantically distinctive (non-overlapping). This filtering rule prevents attribution collisions caused by synonyms (e.g., avoiding "woods" and "forest" in the same prompt), ensuring that successful recovery is due to precise disentanglement rather than semantic leakage.

These constraints ensure the benchmark is both reproducible and challenging. As evidenced by our error analysis (Hamming distance histograms), the benchmark is not "easy"; the model exhibits a bimodal performance where it either perfectly retrieves the concept or fails completely due to the difficulty of disentangling these specific single-word anchors from the complex scene context. Below are several representative examples from the dataset:

```
[
    {
        "prompt": "a cat wearing a sweater sitting on a sunny beach",
        "targets": ["cat", "sweater", "sunny", "beach"]
    },
    {
        "prompt": "a bird flying above snowy mountains under a blue sky",
        "targets": ["bird", "snowy", "mountains", "sky"]
    },
    {
        "prompt": "a musician playing piano in a hall with a dog beside",
        "targets": ["musician", "piano", "hall", "dog"]
    },
```

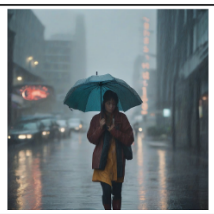
| Prompt | a woman tightly holding an umbrella while walking through the heavy rain. | | | |
|---|---|---|---|---|
| Image |  |  |  |  |
| **Concept** | women | umbrella | heavy | rain |
| **Bit Acc.** | 100% | 100% | 78.13% | 90.63% |
| **Att. Acc.** | 100% | 100% | 50% | 75% |

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

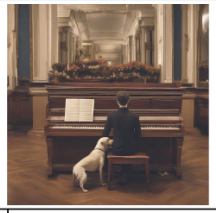
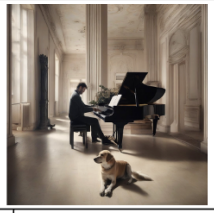
| Prompt | a musician playing piano in a hall with a dog beside. | | | |
|---|---|---|---|---|
| Image |  |  |  |  |
| **Concept** | musician | piano | hall | dog |
| **Bit Acc.** | 100% | 100% | 100% | 100% |
| **Att. Acc.** | 100% | 100% | 100% | 100% |

Figure 5. Qualitative example of multi-customized concept prediction. We generate four images for each prompt containing multiple watermarked concepts, and report the average bit accuracy and average attribution accuracy.

```
    {
        "prompt": "a fox sitting beside a glowing campfire at night",
        "targets": ["fox", "glowing", "campfire", "night"]
    },
    {
        "prompt": "a woman tightly holding an umbrella while walking through the heavy rain",
        "targets": ["women", "umbrella", "heavy", "rain"]
    },
    ...
]
```

Listing 1. Example entries from our self-constructed general multi-concept dataset. Each entry contains a generation prompt and a list of target concepts for retrieval.

In our experiment, each general concept in the target vocabulary (e.g., “cat,” “dog,” “sweater,” “beach”) is pre-associated with its own unique, learnable secret. This dataset allows us to generate a single image using the full prompt and then iteratively query the TokenTrace module with each of the four target words to verify that all four distinct secrets can be successfully and independently retrieved.

## 7. Comprehensive Attribution and Security Test

To provide a final, rigorous evaluation of our method’s overall performance and security, we conduct a comprehensive attribution test that jointly evaluates the True Positive Rate (TPR) and the False Positive Rate (FPR). We construct a large test set ($X_{\text{test}}$) composed equally of watermarked images ($I_{\text{wm}}$) and zero-watermark clean images ($I_{\text{clean}}$), and query every image for the same known secret ($\mathcal{S}_{\text{GT}}$). This test measures the following key metrics:

- Effectiveness (TPR): The percentage of times the system correctly attributes the secret to a watermarked image ($I_{\text{wm}}$).
- Security (FPR): The percentage of times the system incor-

| Prompt | A peaceful and breathtaking landscape painting in the signature style of Dali, illustrating rolling green hills, a tranquil lake reflecting the sky, distant mountains softened by mist, and multiple trees. | | | | |
|---|---|---|---|---|---|
| Image |  | | | | |
| Concept | Dali | hills | lake | mountains | trees |
| Bit Acc. | 100% | 76.25% | 87.5% | 100% | 100% |
| Att. Acc. | 100% | 60% | 80% | 100% | 100% |

Figure 6. Disentanglement of Concrete Concepts and Attributes. Per-concept retrieval accuracy for the complex road scene prompt. Prominent concepts (dog, sunglasses) are recovered at 100% accuracy, while the abstract attribute 'sunny' shows the lowest retrieval rate (60%), confirming performance dependency on visual clarity.

| Prompt | on a sunny road beside the ocean, a dog wearing sunglasses stands confidently as the sea breeze blows past. | | | | |
|---|---|---|---|---|---|
| Image |  | | | | |
| Concept | sunny | road | ocean | dog | sunglass |
| Bit Acc. | 77.5% | 100% | 90% | 100% | 100% |
| Att. Acc. | 60% | 100% | 80% | 100% | 100% |

Figure 7. Disentanglement of Abstract Styles and Environmental Concepts. Per-concept retrieval accuracy for the complex landscape prompt. Primary concepts ('Dali' style, mountains) are recovered perfectly, while the general environmental concept 'hills' shows the lowest retrieval rate (60%), illustrating the performance boundary for general, non-dominant concepts.

rectly attributes the secret to a clean image ($I_{\text{clean}}$). For a robust system, this value must be near zero.

- Overall Performance ($F_1$ Score): A balanced harmonic mean of precision and recall, demonstrating the system's overall viability.

The results confirm that TokenTrace achieves excellent performance and perfect security on this comprehensive test. We measured the True Positive Rate (TPR) at 92.75% and the False Positive Rate (FPR) at 0.00%. The resulting $F_1$ Score is 96.20%. This zero-FPR indicates that the model is highly specific, learning the unique, embedded signature and not general noise, confirming the integrity and high security of our causal attribution mechanism.

## 8. Qualitative Examples on Multi-Concept Attribution

In this section, we provide additional qualitative examples to further demonstrate the effectiveness and robustness of our TokenTrace framework. These examples supplement the multi-concept attribution results presented in the main paper, showing our method's ability to disentangle and independently retrieve secrets for both customized and general concepts.

Figure 2 and Figure 3 show two additional examples of compositional attribution for customized concepts sourced from the Textual Inversion library. In each case, an image is generated from a prompt that combines a specific object concept and a specific style concept. As the results demonstrate, when the single composite image is queried with a prompt for the object, it correctly retrieves the object's secret. When the same image is queried with a prompt for the style, it correctly retrieves the style's secret. This confirms our method's ability to disentangle customized, overlapping concepts.

Figure 4 and Figure 5 show two additional examples of attribution for general concepts. These images were generated from complex prompts containing multiple watermarked keywords (e.g., "cat," "sweater," "beach"). The

results provide further evidence that our query-based TokenTrace module can successfully isolate and retrieve the correct, distinct secret for each general concept, even when they are composed in a single image.

Figure 6 and Figure 7 provide additional qualitative analysis, demonstrating both the strong disentanglement capability of the TokenTrace module and the performance boundaries related to concept abstractness. The experiments show that retrieval success is contingent on the faithfulness of the concept's visual representation by the underlying generative model.

In both complex prompts, all primary, high-priority concepts, such as the stylistic attribution 'Dali' (100% Att. Acc.) or the object 'dog' and accessory 'sunglasses' (both 100% Att. Acc.), are retrieved perfectly, confirming the module's core disentanglement strength. However, both figures illustrate a clear performance drop for non-dominant, abstract, or atmospheric concepts:

- In the road scene (Figure 6), the abstract attribute 'sunny' has the lowest retrieval rate (60% Att. Acc.).
- In the landscape scene (Figure 7), the general environmental concept 'hills' also achieves the lowest retrieval rate (60% Att. Acc.).

This pattern confirms that when the diffusion model struggles to robustly ground an abstract concept in the visual output, the corresponding visual features are weak, which in turn diminishes the semantic signal for the TokenTrace module's frozen CLIP encoders to capture. This highlights a boundary condition of our method: while highly effective, successful retrieval remains contingent on the visual clarity of the concept rendered by the generative model.

## 9. Qualitative Analysis on Complex Prompts

This section presents a qualitative analysis of TokenTrace's performance on images generated from complex, descriptive prompts, utilizing the WikiArt dataset. We compare the visual fidelity of "clean" images (left in each pair) against their "watermarked" counterparts (right in each pair). Our primary goal with this analysis is to visually confirm that our dual-conditioning embedding strategy introduces minimal perceptible degradation to the generated images, even when the prompts are intricate and aim for specific artistic styles or detailed scenes. Each example highlights a prompt designed to elicit a rich visual output, varying in subject matter, artistic style, and atmospheric conditions.

As evident from the Figure 8, the watermarked images consistently maintain high visual quality, faithfully representing the aesthetic and thematic intentions of the original prompts. Details such as rolling hills, towering waves, mystical landscapes, and historical figures are preserved with remarkable consistency. The subtle perturbations introduced by TokenTrace, which are embedded in both the textual semantic and latent domains, do not lead to noticeable artifacts, blurring, or distortion, thus validating that our method achieves robust watermarking without compromising the generative model's artistic capabilities. This visual assessment reinforces the quantitative metrics presented in the main paper regarding image quality (e.g., CSD scores).

## 10. Image Fidelity Analysis

To evaluate the impact of watermarking on the overall generation quality, we measure the CLIP score, CSD score, and Fréchet Inception Distance (FID) [3] on the ImageNet dataset. A lower FID score indicates that the generated images are closer to the distribution of real images.

From the results show in Table 3, we can see that TokenTrace achieved a higher visual quality across all metrics than ProMark, and achieved comparable performance to CustomMark. We attribute this superiority to our additive semantic perturbation as mentioned in the novelty part. Such a strategy can ensure the watermark within the concept's original semantic manifold without finetuning the original diffusion model. Unlike replacement-based methods that may cause a 'semantic shift' away from the target concept, our additive approach applies a minimal, learned perturbation that preserves the fundamental stylistic and structural features of the concept.

Table 3. Visual quality comparison on CLIP, CSD, and FID.

| Method | CLIP ↑ | CSD ↑ | FID ↓ |
|---|---|---|---|
| ProMark | 0.68 | 0.65 | 17.63 |
| CustomMark | 0.72 | 0.70 | 14.73 |
| TokenTrace | 0.74 | 0.71 | 14.98 |

## 11. Description about the "Adversarial Attack"

The "Adversarial Attack" present in the Table 4 of main paper represents a high-capability, black-box adversary that erases hidden signals without any prior knowledge of our encoder and decoder, then uses a diffusion-based denoising process to reconstruct the image. Technically, this attack first corrupts the image's latent representation by adding gaussian noise to its latent representation, and then reconstructs it using a generative model (e.g. diffusion model) to eliminate watermarked signals. While its authors prove that standard pixel-based watermarks are vulnerable to this attack, our results demonstrate that by embedding the watermark within the semantic manifold of the concept, TokenTrace maintains significant attribution accuracy even after generative purification.

## 12. Quantitative Disentanglement Evidence

To verify our disentanglement claims, we performed a quantitative analysis using a controlled dataset selected


Figure 8. Qualitative comparison of images generated from complex prompts on the WikiArt dataset. For each pair, the left image is the clean (unwatermarked) version, and the right image is the watermarked version generated by TokenTrace. Our method consistently maintains high visual fidelity and adherence to the prompt, even for intricate artistic scenes.

from our general concept attribution dataset, specifically, it contains 10 concepts (5 Objects, 5 Styles) for better visualization. To generate the prompts, we employed GPT with the same constraints as we used previously. Finally, we construct a new subset that contains 25 prompts, each prompt contains exactly 4 target concepts (2 Objects and 2 Styles), and across the entire dataset, every single target concept appears exactly 10 times. Figure 9 presents the resulting concept-level confusion matrix. We can observe that the matrix exhibits a perfect block-diagonal structure. The top-right quadrant (object $\rightarrow$ style confusion) and bottom-left quadrant (style $\rightarrow$ object confusion) contain only zeros. This quantitatively proves that the model perfectly disentangles the two domains: an object concept is never misidentified as a style, and vice versa. Despite the high "overlapping difficulty" introduced in the prompts, the object concept retrieval achieved perfect performance. This demonstrates that the model successfully isolates the object concept regardless of the context. We also analyzed failure cases where concepts interfere. We observed confusion between semantically overlapping concepts, such as "cozy" and "quiet". Crucially, such interference is contained strictly within the style domain. The semantic ambiguity of the style concepts does not propagate to or degrade the identification of the object objects.

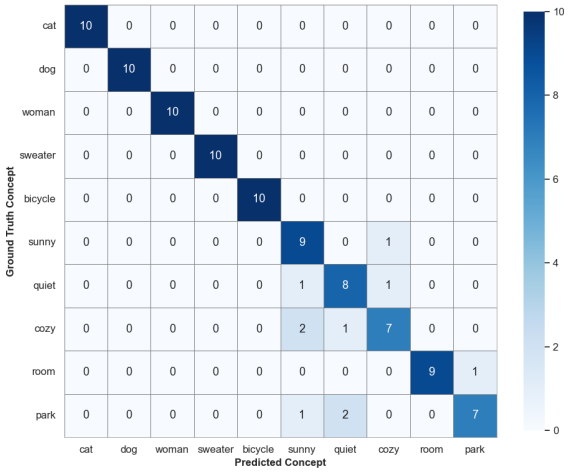



Figure 9. Confusion Matrix. Strict block-diagonal structure proves TokenTrace disentangles Objects from Styles.

## 13. Metric Analysis & Error Distribution

We observed the close proximity between reported Bit Accuracy and Attribution Accuracy (e.g., 90% vs 86%), as it hints that bit errors might be highly correlated rather than independent. To explore that, we performed a detailed error analysis to verify this hypothesis. We plotted the distribution of Hamming distances between the retrieved secrets and the ground truth. As shown in Figure 10, the error distribution is strongly bimodal. The vast majority of samples cluster at a Hamming distance of 0 (perfect retrieval), with

a "clean gap" separating them from the failure cases (Hamming distance $> 8$). This confirms that errors are indeed correlated: the watermark signal tends to survive intact or be lost entirely, rather than suffering from independent random bit flips in this dataset. And, We further investigated the relationship between attribution success and the semantic quality of the retrieved concept. As shown in Figure 11, attribution failures are characterized by high Euclidean distances in the CLIP latent space, which directly correlates with the clustered bit errors. The similar bimodal behavior demonstrates that our decoder is highly stable; errors only occur when the semantic anchor is lost due to the high complexity of multi-concept prompts.

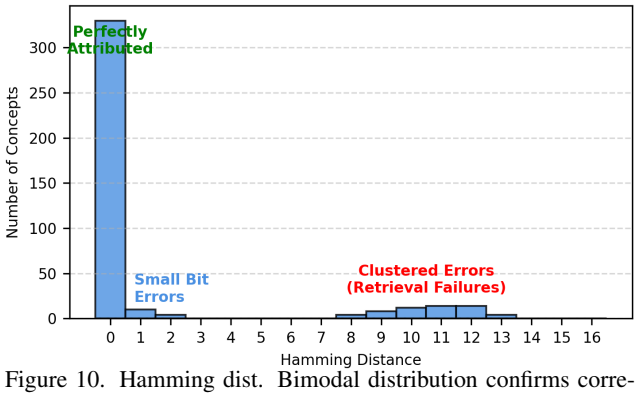



Figure 10. Hamming dist. Bimodal distribution confirms correlated errors.

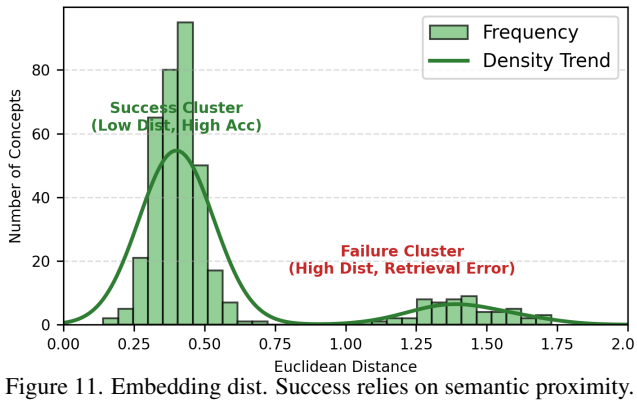



Figure 11. Embedding dist. Success relies on semantic proximity.

## 14. Strategies for Practical Scaling of Applications

when transitioning to larger concept inventories, a strategy beyond simply increasing secret length is required, as our experiments showed performance degradation when the bit-length exceeds 16 bits. After some investigation, we find that there are several practical scaling strategies:

- Instead of a flat 16-bit space, concepts can be organized into a hierarchy (e.g., Category -¿ Sub-category -¿ Specific Concept). A 16-bit secret could be used at each level of the hierarchy, allowing for a massive total concept space while keeping the per-concept bit-extraction task within the high-performance 16-bit regime.
- Integrating standard error-correcting codes (e.g. reed-solomon code, BCH code, etc.), which could help mitigate the bit-errors for the "small bit errors" (Hamming distance 1–2). Such strategies can potentially allow for slightly longer secrets without a large loss of attribution performance.

Regarding verification costs when the target concept is unknown, since our verification is query-based, if a verifier does not know which of N concepts was used, a brute-force search would require N extraction passes. In such a scenario, the hierarchical identifiers that mentioned above can help reduce the computational cost greatly. We will include a "Practical Deployment" section in the revision to discuss these scaling and cost considerations explicitly.

## 15. Prompt Details

This section details the standardized prompt templates used in our single-concept and multi-concept experiments, following the approach for training customized embeddings (e.g., CustomMark [1]).

Following is the complete list of standardized prompt templates used in our single-concept attribution experiments:

1. Prompts used during training for single style concept attribution. The [name] placeholder is replaced by the specific watermarked tokens during the experiment.
   - *a painting, art by [name]*
   - *a rendering, art by [name]*
   - *a cropped painting, art by [name]*
   - *the painting, art by [name]*
   - *a clean painting, art by [name]*
   - *a dirty painting, art by [name]*
   - *a dark painting, art by [name]*
   - *a picture, art by [name]*
   - *a cool painting, art by [name]*
   - *a close-up painting, art by [name]*
   - *a bright painting, art by [name]*
   - *a rendition, art by [name]*
   - *a nice painting, art by [name]*
   - *a small painting, art by [name]*
   - *a weird painting, art by [name]*
   - *a large painting, art by [name]*
   - *a serene landscape painting in the style of [name]*
   - *a bustling cityscape in the style of [name]*
   - *a painting of a cozy cottage in the woods in the style of [name]*
   - *a vibrant underwater scene in the style of [name]*
   - *a whimsical painting of a flying elephant in the style of [name]*
   - *a still life painting featuring fruit and flowers in the style of [name]*
   - *a portrait of a famous historical figure in the style of [name]*

- *a painting of a dreamy night sky in the style of [name]*
- *a colorful abstract painting in the style of [name]*
- *a street scene from Paris in the style of [name]*
- *a depiction of a beautiful sunset over the ocean in the style of [name]*
- *a painting of a peaceful mountain village in the style of [name]*
- *an energetic painting of dancers in motion in the style of [name]*
- *a painting of a snow-covered winter scene in the style of [name]*
- *a painting of a tropical paradise in the style of [name]*
- *a painting of a magical forest filled with fantastical creatures in the style of [name]*
- *a painting of a dramatic stormy seascape in the style of [name]*
- *a portrait of a majestic lion in the style of [name]*
- *a painting of a romantic scene between two lovers in the style of [name]*
- *a painting of a serene Japanese garden in the style of [name]*
- *a painting of a bustling marketplace in the style of [name]*
- *a painting of a tranquil river scene in the style of [name]*
- *a painting of a fiery volcano eruption in the style of [name]*
- *a painting of a futuristic cityscape in the style of [name]*
- *a painting of a whimsical circus scene in the style of [name]*
- *a painting of a mysterious moonlit forest in the style of [name]*
- *a painting of a dramatic desert landscape in the style of [name]*
- *a portrait of a regal peacock in the style of [name]*
- *a painting of a mystical island in the style of [name]*
- *a painting of a lively carnival scene in the style of [name]*

2. Prompts used during training for single object concept attribution. The [name] placeholder is replaced by the specific watermarked tokens during the experiment.
   - *a photo of a [name]*
   - *a rendering of a [name]*
   - *a cropped photo of the [name]*
   - *the photo of a [name]*
   - *a photo of a clean [name]*
   - *a photo of a dirty [name]*
   - *a dark photo of the [name]*
   - *a photo of my [name]*
   - *a photo of the cool [name]*
   - *a close-up photo of a [name]*
   - *a bright photo of the [name]*
   - *a cropped photo of a [name]*
   - *a photo of the [name]*
   - *a good photo of the [name]*
   - *a photo of one [name]*
   - *a close-up photo of the [name]*
   - *a rendition of the [name]*
   - *a photo of the clean [name]*
   - *a rendition of a [name]*
   - *a photo of a nice [name]*
   - *a good photo of a [name]*
   - *a photo of the nice [name]*
   - *a photo of the small [name]*
   - *a photo of the weird [name]*
   - *a photo of the large [name]*
   - *a photo of a cool [name]*
   - *a photo of a small [name]*
   - *a photo of a [name] playing sports*
   - *a rendering of a [name] at a concert*
   - *a cropped photo of the [name] cooking dinner*
   - *the photo of a [name] at the beach*
   - *a photo of a clean [name] participating in a marathon*
   - *a photo of a dirty [name] after a mud run*
   - *a dark photo of the [name] exploring a cave*
   - *a photo of my [name] at graduation*
   - *a photo of the cool [name] performing on stage*
   - *a close-up photo of a [name] reading a book*
   - *a bright photo of the [name] at a theme park*
   - *a cropped photo of a [name] hiking in the mountains*
   - *a photo of the [name] painting a mural*
   - *a good photo of the [name] at a party*
   - *a photo of one [name] playing an instrument*
   - *a close-up photo of the [name] giving a speech*
   - *a rendition of the [name] during a workout*
   - *a photo of the clean [name] gardening*
   - *a rendition of a [name] dancing in the rain*
   - *a photo of a nice [name] volunteering at a charity event*
   - *a photo of a [name] surfing a giant wave*
   - *a rendering of a [name] skydiving over a scenic landscape*
   - *a cropped photo of the [name] riding a rollercoaster*
   - *the photo of a [name] rock climbing a steep cliff*
   - *a photo of a clean [name] practicing yoga in a peaceful garden*
   - *a photo of a dirty [name] participating in a paintball match*
   - *a dark photo of the [name] stargazing at a remote location*
   - *a photo of my [name] crossing the finish line at a race*
   - *a photo of the cool [name] breakdancing in a crowded street*
   - *a close-up photo of a [name] blowing out candles on a birthday cake*

- *a bright photo of the [name] flying a kite on a sunny day*
- *a cropped photo of a [name] ice-skating in a winter wonderland*
- *a photo of the [name] directing a short film*
- *a good photo of the [name] participating in a flash mob*
- *a photo of one [name] skateboarding in an urban park*
- *a close-up photo of the [name] solving a Rubik's cube*
- *a rendition of the [name] fire dancing at a beach party*
- *a photo of the clean [name] planting a tree in a community park*
- *a rendition of a [name] performing a magic trick on stage*
- *a photo of a nice [name] rescuing a kitten from a tree*

3. Prompts used during inference for multiple customized concepts attribution. The [name_object] and [name_style] placeholders are replaced by the specific watermarked tokens during the experiment.
   - *a photo of a [name_object] in the style of [name_style] with a clear background.*
   - *a rendering of a [name_object] in the style of [name_style] with a clear background.*
   - *a cropped photo of the [name_object] in the style of [name_style] with a clear background.*
   - *the photo of a [name_object] in the style of [name_style] with a clear background.*
   - *a photo of a clean [name_object] in the style of [name_style] with a clear background.*
   - *a photo of a dirty [name_object] in the style of [name_style] with a clear background.*
   - *a dark photo of the [name_object] in the style of [name_style] with a clear background.*
   - *a photo of my [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the cool [name_object] in the style of [name_style] with a clear background.*
   - *a close-up photo of a [name_object] in the style of [name_style] with a clear background.*
   - *a bright photo of the [name_object] in the style of [name_style] with a clear background.*
   - *a cropped photo of a [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the [name_object] in the style of [name_style] with a clear background.*
   - *a good photo of the [name_object] in the style of [name_style] with a clear background.*
   - *a photo of one [name_object] in the style of [name_style] with a clear background.*
   - *a close-up photo of the [name_object] in the style of [name_style] with a clear background.*
   - *a rendition of the [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the clean [name_object] in the style of [name_style] with a clear background.*
   - *a rendition of a [name_object] in the style of [name_style] with a clear background.*
   - *a photo of a nice [name_object] in the style of [name_style] with a clear background.*
   - *a good photo of a [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the nice [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the small [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the weird [name_object] in the style of [name_style] with a clear background.*
   - *a photo of the large [name_object] in the style of [name_style] with a clear background.*
   - *a photo of a cool [name_object] in the style of [name_style] with a clear background.*
   - *a photo of a small [name_object] in the style of [name_style] with a clear background.*

## References


[1] 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. 9

[2] Anh Bui, Trang Vu, Trung Le, Junae Kim, Tamas Abraham, Rollin Omari, Amar Kaur, and Dinh Phung. Mitigating semantic collapse in generative personalization with a surprisingly simple test-time embedding adjustment. *arXiv preprint arXiv:2506.22685*, 2025. 2

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017. 7

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 2

[5] 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. 3

[6] 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. 2