Toward LoRA Copyright Protection with an Authorized Dual-Watermarking Framework

Zhipeng Yin[1], **Zichong Wang**[1], **Ruijun Chen**[1], **Xin Ning**[1], **Xingyu Zhang**[2] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, USA
[2]University of Pittsburgh, PA, USA

Abstract

Text-to-Image (T2I) diffusion models have been widely adopted due to their strong generative capabilities, while Low-Rank Adaptation (LoRA) has emerged as an efficient mechanism for customizing these models for diverse creative and commercial applications. This trend has fostered LoRA-centric service platforms that that enable the customization and commercial distribution of LoRA modules according to user requirements. However, the growing prevalence of LoRA and its critical role in customized AI services have raised urgent concerns about LoRA copyright protection. To address this gap, we propose $LoRA^2D$, an authorized dual-watermarking framework specifically designed to protect LoRA modules in T2I diffusion models. $LoRA^2D$ integrates license-based authorization control with explicit watermarks as visible deterrents for unauthorized or trial usage, which can be removed upon valid authorization, while persistently embedding an implicit watermark for robust black-box ownership verification. Extensive experiments on multiple image-generation datasets demonstrate the effectiveness and practicality of $LoRA^2D$ for securing copyrights in LoRA-adapted T2I diffusion models.

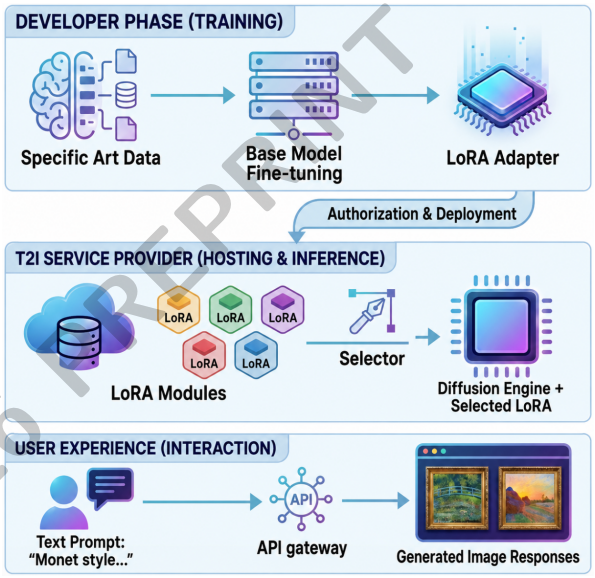

Figure 1: Illustration of the LoRA-Service (LS) paradigm, where authorized LoRA modules are curated by service providers and dynamically applied to pre-trained backbone models, where users can select authorized LoRA modules based on their specified queries for personalized T2I generation.

1 Introduction

Text-to-Image (T2I) diffusion models have made remarkable advances in recent years, exemplified by prominent models such as Stable Diffusion [Borji, 2022], DALL·E [Zhou and Nabus, 2023], and Midjourney [Kwon, 2024], driving significant progress in visual content generation across various industries. However, training these large-scale generative models demands considerable computational resources and massive volumes of high-quality visual datasets, rendering full-parameter fine-tuning economically and practically prohibitive for many researchers and developers [Parthasarathy *et al.*, 2024; Tu *et al.*, 2024]. Consequently, practitioners commonly adopt the strategy of leveraging pre-trained diffusion models and efficiently fine-tuning them for specific downstream tasks [Ma *et al.*, 2025]. To mitigate the prohibitive costs associated with traditional full-model tuning, Low-Rank Adaptation (LoRA) has emerged as a prevalent and highly efficient solution [Hu *et al.*, 2022]. LoRA modules function as lightweight plug-in components, containing substantially fewer parameters relative to the large backbone diffusion models they augment [Shen *et al.*, 2025]. These modules are typically trained on specialized, task-specific image datasets and can subsequently be seamlessly integrated into pre-trained diffusion architectures. This paradigm enables a modular and cost-effective approach, in which the pre-trained diffusion model provides generalized visual synthesis capabilities, while LoRA modules contribute precise and customized knowledge crucial for targeted image-generation tasks [Huan and Shun, 2025]. The flexibility, efficiency, and effectiveness of LoRA for T2I diffusion models have given rise to a burgeoning market, fostering the commercial dis-

*Corresponding author.

tribution of professionally customized LoRA modules. This emerging paradigm, termed *LoRA-Service* (LoRAS), involves three primary stakeholders: the LoRA developer, the T2I service provider, and the end user. An illustrative overview of the LoRAS workflow is shown in Figure 1. By clearly delineating responsibilities and incentives among stakeholders, LoRAS also promotes specialized LoRA contributions and establishes a thriving ecosystem centered around the provision of specialized visual generative services [Ma *et al.*, 2025; Cao *et al.*, 2024].

With the increasing reliance on LoRA modules for specialized AI solutions, protecting the intellectual property rights associated with these modules has become a critical concern for commercial LoRA providers [Aldhaheri *et al.*, 2024; Sundaram *et al.*, 2019]. The unauthorized or illicit use of LoRA modules poses significant threats in multiple dimensions. From an economic perspective, unauthorized use directly undermines the benefits and incentives that LoRA developers gain from investing significant effort and resources into the development and optimization of modules [Andrade and Yoo, 2019]. Furthermore, the datasets used to train LoRA often contain sensitive information, making unauthorized reuse susceptible to severe privacy risks, including potential data leaks and inference attacks. Additionally, without robust protection mechanisms, LoRA can be manipulated, destroyed, and redistributed, resulting in unpredictable or harmful model behavior. [Liu *et al.*, 2024; Luo *et al.*, 2025] Despite these obvious risks, there remains a notable research gap in effective copyright protection strategies for LoRA modules.

Ensuring ownership of artificial intelligence models is typically achieved through watermarking technology, which has become a widely accepted and effective method of copyright protection within the industry [Zhang *et al.*, 2018; Ray and Roy, 2020]. Watermarking methods embed specific identification information or unique signatures into either the parameters or outputs of AI models. Even if the model undergoes various intentional modifications, such as fine-tuning or parameter adjustments, the embedded information can still be extracted and verified. However, conventional watermarking methods originally designed for large-scale machine learning models are not directly applicable to LoRA modules. This limitation stems primarily from the fact that LoRA modules have far fewer trainable parameters than large foundational models, resulting in extremely low redundancy available for embedding watermarks. Consequently, imposing traditional watermarking methods might negatively affect LoRA's ability to achieve the intended performance in downstream tasks.

Motivated by the necessity to address these specific challenges and ensure reliable protection of LoRA modules, we introduce a novel framework $LoRA^2D$: an authorized dual-watermarking approach for LoRA copyright protection. Our method is specifically designed for image generation tasks, with Stable Diffusion serving as the backbone architecture. Unlike previous methods, the $LoRA^2D$ framework integrates two complementary watermarking mechanisms into the LoRA module: explicit (visible) watermarks and implicit (invisible) watermarks. Our main contributions are summarized as follow:

- We conducted a systematic study on LoRA copyright protection in the context of LoRA plugin services, analyzed the risks associated with unauthorized use of LoRA in platform-based artificial intelligence ecosystems, and identified the technical standards required to achieve LoRA copyright protection.
- We propose the first authorization-controlled dual watermarking framework for LoRA copyright protection. Our method combines license-based explicit watermarks with persistent implicit watermarks, enabling flexible management of LoRA distribution and reliable ownership verification in both authorized and unauthorized scenarios.
- Extensive experiments on benchmark datasets demonstrate that our dual watermarking method effectively prevents unauthorized use and achieves robust watermark detection, thereby meeting the practical requirements of copyright enforcement.

2 Related Works

Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] has become a popular parameter-efficient fine-tuning technique for large machine learning models. By inserting low-rank trainable matrices into specific layers of the pre-trained model, the LoRA module can efficiently adapt to downstream tasks without modifying the original model parameters. This plug-and-play capability has been demonstrated across various applications in both language and vision domains, including text-to-image diffusion models such as Stable Diffusion [Yang *et al.*, 2024]. The success and widespread adoption of LoRA have driven the development of advanced variants to enhance its efficiency and applicability. For example, rsLoRA [Kalajdzievski, 2023] introduces a rank-stabilizing scaling factor to address gradient collapse in higher ranks and unlock the performance potential of high-rank adapters without incurring inference overhead. AdaLoRA [Zhang *et al.*, 2023a] proposes an adaptive budget allocation strategy that evaluates the relative importance of LoRA components and selectively removes less informative singular values during training, allowing the model to concentrate its limited parameter capacity on the most influential updates and achieve effective fine-tuning under strict parameter budgets. These improvements reflect ongoing efforts to extend LoRA's utility, making it increasingly suitable for deployment in resource-constrained environments. Given the growing commercial value and importance of LoRA, the necessity of implementing robust copyright protection measures to prevent unauthorized use has become increasingly evident.

Watermarks for Diffusion Models. With the widespread adoption of diffusion models in commercial image generation services, protecting their intellectual property from unauthorized exploitation has become increasingly critical. Existing watermarking methods for diffusion models generally fall into two categories: *white-box* and *black-box*. White-box techniques typically embed watermarks directly into model parameters via retraining or fine-tuning. For instance, AquaLoRA [Feng *et al.*, 2024] incorporates watermark signals through parameter-efficient adaptations in the U-Net

backbone, enabling reliable watermark verification when internal weights are accessible. Chou et al. [Chou *et al.*, 2023] implant implicit watermarking through modified diffusion dynamics, triggered by predefined signals. Conversely, black-box methods embed watermark triggers into the model's generation behaviors. Zhao et al. [Zhao *et al.*, 2023] and Liu et al. [Liu *et al.*, 2023] employ trigger-based prompt strategies for watermark encoding. However, these trigger-dependent techniques remain susceptible to adversarial detection or removal. Additionally, existing methods largely neglect watermark protection tailored explicitly for standalone LoRA modules. In contrast, our proposed framework addresses this limitation by introducing a dual-watermark embedding strategy specifically designed for LoRA-based diffusion models. This approach introduces robust implicit watermarking for black-box verification and explicit watermarking as a visual deterrent, specifically designed to protect LoRA modules.

3 Background

Text-to-Image Diffusion Models. Text-to-Image (T2I) diffusion models generate images from textual prompts through a probabilistic diffusion process operating in a latent representation space [Rombach *et al.*, 2022]. Typically, these models integrate three essential modules: i) A text encoder, ii) A latent diffusion module, and iii) An autoencoder, typically a variational autoencoder. To effectively model complex image distributions, T2I diffusion models first introduce noise into structured image data (forward diffusion) and then learn to reverse this noise-adding process to reconstruct images from noise (reverse diffusion).

Forward Diffusion Process. During this stage, structured image data is gradually corrupted into noise through a Markovian sequence. Formally, forward diffusion is defined by the conditional distribution $p(x_t \mid x_{t-1})$, incrementally introducing Gaussian noise into latent variables at each discrete time step t, represented as: $x_t = \sqrt{1-\beta_t}\, x_{t-1} + \beta_t\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$, where β_t represents a variance schedule dependent on time. Equivalently, the cumulative diffusion from the initial latent representation x_0 to a given noisy state x_t can be simplified through the cumulative parameters $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$: $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$.

Reverse Diffusion Process. In this process, noise is progressively transformed back into structured image data. Given a noisy latent state z_t and a embedding $\tau(y)$ derived from the textual prompt y, the diffusion model estimates the original noise through the prediction function $\epsilon_\theta(z_t, t, \tau(y))$. Accurate prediction facilitates the estimation of the posterior distribution $q(z_{t-1} \mid z_t)$, reversing the forward diffusion dynamics and progressively reconstructing coherent images from noise.

Low-Rank Adaptation. Fine-tuning large pretrained models, particularly large language models (LLMs) and diffusion models, often involves significant computational costs and storage requirements due to their enormous parameter sizes. Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] efficiently addresses this by restricting updates to low-rank matrices, significantly reducing resource demands. Specifically, given a pretrained weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA modifies it as: $W' = W + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. Here, r denotes the rank of the adaptation, typically much smaller than the original dimensions d and k. During fine-tuning, only matrices A and B are updated, drastically reducing memory usage and computational complexity. LoRA has demonstrated effectiveness in various domains, including natural language processing, vision, and generative modeling, making it widely adopted for practical, scalable fine-tuning.

4 Methodology

4.1 The proposed framework – $LoRA^2D$

In this section, we introduce a novel framework, $LoRA^2D$, an authorized dual-watermarking system composed of complementary components: explicit watermark embedding (B_{vis}) and implicit watermark embedding (B_{imp}). Both watermarks, together with the primary task module (B_{task}), are integrated into a multi-head LoRA structure, where each component occupies an independent low-rank head to minimize interference and ensure parameter efficiency. Specifically, the explicit watermark provides visual authorization deterrence via a perceptible logo, while the implicit watermark enables continuous black-box verification for copyright enforcement. A license-based gating mechanism dynamically controls the visibility of the explicit watermark head, ensuring unobtrusive authorized usage. During copyright verification, suspected services are queried, and the presence of the implicit watermark is validated using a pretrained watermark decoder. We detail each component in the following sections.

4.2 Dual Watermarks Embedding

We begin by introducing the dual-watermark pre-training mechanism of our proposed $LoRA^2D$, which employs two complementary embedding components: **(i) implicit watermark embedding**, encoding invisible watermark signals into the latent space through a dual-encoder fusion mechanism for black-box verification; and **(ii) explicit watermark embedding**, generating a perceptible visual logo within designated regions of output images as a clear deterrent against unauthorized use. In the following, we detail implicit watermark embedding component firstly.

Implicit watermark embedding. In this section, we detail the implicit watermark embedding component of $LoRA^2D$. This module aims to encode a robust and imperceptible ownership signal into the generative behavior of a LoRA, ensuring generated images implicitly carry a recoverable watermark while preserving original generative fidelity. Specifically, the implicit watermark is embedded in the latent representation of the image, rather than directly in the pixel space. Such implicit watermarks enable reliable black-box copyright verification after deployment, even if explicit visual marks are removed or suppressed. Given that LoRA modules have a limited number of trainable parameters unsuitable for embedding high-capacity watermarks directly, we propose a two-stage embedding strategy operating in latent diffusion space, tightly integrated with LoRA fine-tuning. Our approach utilizes a dual-encoder fusion structure consisting of a watermark encoder E_w and an image encoder E_I. The use of two

separate encoders allows the watermark and image-content distributions to be modeled independently, which is important because the binary watermark signal and the continuous image latent features have substantially different statistical properties. Such a probabilistic dual-encoder design enhances adaptability by explicitly capturing the uncertainty and modality-specific characteristics of both watermark and image representations, while also reducing mutual interference between them. Consequently, this structure improves watermark robustness while preserving the original image quality and generative fidelity. Instead of directly producing deterministic embeddings, the two encoders parameterize latent distributions for the watermark message and the image content, respectively. Formally, given a binary watermark message $\mathbf{w} \in \{0,1\}^k$, where k denotes the predefined watermark length, The watermark encoder produces a latent distribution:

$$h_w = E_w\left(\mu_w(\mathbf{w}), \mathrm{diag}(\sigma_w^2(\mathbf{w}))\right) \tag{1}$$

where $\mu_w(\cdot)$ represent the mean parameterized by the watermark encoder, $\mathrm{diag}(\sigma_w^2(\cdot))$ denotes a diagonal covariance matrix and h_w denotes the latent embeddings of the watermark.

Given a clean cover image I_c, the image encoder parameterizes the latent distribution of the image-content:

$$h_I = E_I\left(\mu_I(I_c), \mathrm{diag}(\sigma_I^2(I_c))\right) \tag{2}$$

where $\mu_I(\cdot)$ are the corresponding mean parameterized by the image encoder, and similarly $\mathrm{diag}(\sigma_I^2(\cdot))$ is the diagonal covariance matrix and h_I denotes the latent embeddings of the image-content. The two embeddings are then concatenated and passed through a learnable fusion module F_ψ:

$$h_f = F_\psi(h_w, h_I) = \mathrm{MLP}_\psi\left([h_w \| h_I]\right) \tag{3}$$

where $[\cdot\|\cdot]$ denotes feature concatenation and MLP_ψ maps the concatenated representation into a joint latent embedding.

We integrate the watermark and image embeddings into a fused representation h_f, but this alone does not guarantee sufficient preservation of information from both sources. To explicitly quantify and control such nonlinear statistical dependencies, we employ the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation, which is quantify complex nonlinear statistical dependencies, making it particularly suited for capturing the intricate relationships between watermark and image embeddings. Formally, HGR is defined as:

$$\mathrm{HGR}(X,Y) = \sup_{f,g} \frac{\mathbb{E}[f(X)\, g(Y)]}{\sqrt{\mathbb{E}[(f(X))^2]\,\mathbb{E}[(g(Y))^2]}} \tag{4}$$

Specifically, we compute:

$$HGR_w = \mathrm{HGR}(h_w, h_f), \quad HGR_I = \mathrm{HGR}(h_I, h_f) \tag{5}$$

representing the correlation between fused latent embedding h_f and watermark embedding h_w, as well as between h_f and image-content embedding h_I. To ensure balanced information fusion from both sources, we formulate a balanced HGR-based regularization term:

$$\mathcal{L}_{\mathrm{HGR}} = \frac{1}{2}\left(HGR_w + HGR_I\right) - \lambda_{\mathrm{bal}}\left(HGR_w - HGR_I\right)^2 \tag{6}$$

where first term maximizes overall statistical dependence between the fused representation and both embeddings, while the second term penalizes significant differences in the contributions between HGR_w and HGR_I. Therefore, the fused representation is encouraged to retain watermark-related information for robust recovery and image-content information for visual fidelity in a balanced manner.

Although HGR regularization encourages fused representation h_f to preserve watermark-related information, it does not directly ensure that the embedded message can be recovered from the final generated image after VAE decoding and realistic distortions. Therefore, we further introduce a message similarity loss to provide supervision for watermark recoverability:

$$\mathcal{L}_{\mathrm{sim}} = 1 - \frac{\mathbf{w}\cdot\hat{\mathbf{w}}}{\|\mathbf{w}\|\,\|\hat{\mathbf{w}}\|} \tag{7}$$

where $\hat{\mathbf{w}}$ is the decoded watermark prediction from the watermark decoder D_w.

In addition to recoverability, in our dual-encoder latent fusion structure, the watermark embedding h_w and the image-content embedding h_I are first fused into the joint representation h_f, which is then used to generate the injected latent residual $\Delta\mathbf{z}$. Since h_f is explicitly encouraged to preserve watermark-related information through the HGR regularization, the model may otherwise increase watermark recoverability by injecting an excessively large residual into the clean latent representation, which can cause the watermarked latent representation to deviate significantly from the original clean latent. We define the injected latent residual as: $\Delta\mathbf{z} = \mathbf{z}_{wm} - \mathbf{z}_c$, where $\mathbf{z}_c$ and $\mathbf{z}_{wm}$ denote the clean and watermarked latent representations. We then constrain the relative energy of the injected residual:

$$\mathcal{L}_{\mathrm{res}} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{W}\|\Delta\mathbf{z}_{i,j}\|_2^2}{LW} \tag{8}$$

where i, j index the locations of the latent feature map with resolution $L \times W$. By penalizing the magnitude of $\Delta\mathbf{z}$, this constraint prevents the watermark signal from being embedded through excessively large latent perturbations, thereby encouraging a compact watermark injection while preserving the original generative behavior.

Finally, combining all these terms, we formulate our overall training objective for encoder–decoder pre-training as:

$$\min_{\theta,\omega} \mathcal{L}_{\mathrm{Imp}} = \mathcal{L}_{\mathrm{sim}_{(\theta,\omega)}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}_{(\theta)}} - \lambda_{\mathrm{HGR}}\mathcal{L}_{\mathrm{HGR}_{(\theta,\omega)}} \tag{9}$$

where θ and ω denote the parameters of the watermark encoder and image encoder.

Building upon the implicit embedding of watermark, we further introduce the implicit watermark verification. The objective is to determine whether a suspected T2I service employs protected LoRA under a black-box setting, where only

generated outputs are observable and internal model parameters remain inaccessible. Given a suspected T2I service, we verify by querying the service with a set of n semantically coherent prompts $\{y_i\}_{i=1}^n$ and collecting the corresponding generated images $\{I_i\}_{i=1}^n$. To verify the embedded watermark, we employ the pre-trained watermark decoder D_w, which is designed to recover the watermark message from implicitly embedded patterns in the generated images. Watermark verification is conducted by comparing the computed accuracy, which is empirically calibrated to ensure robustness and minimize false positives.

Explicit watermark embedding. While the implicit watermark embedding discussed above provides a robust yet imperceptible watermark, an explicit watermark serves a complementary and essential role, providing an immediate, visible indication of unauthorized or trial usage. Specifically, our explicit watermark embedding mechanism aims to generate a consistent, clearly identifiable logo or text pattern at a predefined spatial location within generated images. Unlike implicit embedding, the explicit watermarking should ensure perceptual visibility to fulfill its deterrent purpose, yet it should minimally interfere with overall image aesthetics.

To realize explicit watermark embedding, we select a set of diverse images that are background-agnostic (typically $100 \sim 500$) and overlay each image with an identical logo, placed consistently in the same fixed location with uniform size and transparency. Alongside the watermarked images, we generate a binary mask M_{logo}, marking the precise logo region as 1, with all other areas as 0. Formally, given an original clean image I, the explicitly watermarked training sample $\hat{I}$ is obtained as:

$$\hat{I} = (1 - M_{\text{logo}}) \odot I + M_{\text{logo}} \odot I_{\text{logo}} \tag{10}$$

where I_{logo} denotes the predefined logo pattern, and $\odot$ represents element-wise multiplication.

For embedding explicit watermarks in a parametrically efficient manner without modifying the backbone diffusion model, we employ a lightweight LoRA optimization strategy. Specifically, we freeze the backbone diffusion model (e.g., the stable diffusion model) and introduce LoRA modules exclusively in the decoder blocks of U-Net, in particular in the terminal layer closest to the output potential representation. By restricting the trainable parameters in this way, we ensure that explicit watermarks can be reliably generated without significantly affecting the generalization ability of the original model. Given an input latent representation derived from a clear image, we optimize the LoRA parameters to reconstruct the corresponding image with an explicit watermark. Let $\hat{I}$ denote the real image with an explicit watermark, and I^* be the output image generated by the diffusion model augmented with the LoRA module. To guide the explicit watermark embedding, we define a reconstruction loss that emphasizes the consistency of the watermarked region while maintaining the quality of the overall image generation:

$$\mathcal{L}_{\text{Vis}} = \mathcal{L}_{\text{gen}}(I^*, \hat{I}) + \lambda_{\text{logo}} \| M_{\text{logo}} \odot (I^* - \hat{I}) \|_1 \tag{11}$$

where the first term $\mathcal{L}_{\text{gen}}(I^*, \hat{I})$ is the standard image reconstruction objective (e.g., mean squared error or perceptual loss), and the second term explicitly enforces pixel-level accuracy within the logo region, balanced by hyperparameter λ_{logo}.

4.3 Authorization and Watermark Control

Having established the embedding mechanisms for implicit and explicit watermarking, we next explore how to selectively activate and manage and detect these watermarking features in real applications. To this end, we first introduce a multi-head LoRA architecture equipped with a gating mechanism and a permission-based authorization framework, as illustrated in Figure 2. This design enables fine-grained control over watermarking visibility, allowing the system to dynamically switch to explicit or implicit watermarking modes based on the user's authorization status. Specifically, This architecture consists of three independent LoRA branches, each specialized for a distinct functionality: *i) LoRA-main ($\mathcal{H}_{\text{main}}$)*: Encodes primary generative capabilities specific to tasks or styles, trained on the target domain dataset without watermark constraints. *ii) LoRA-vis ($\mathcal{H}_{\text{vis}}$)*: Generates visually explicit watermark logos in generated images. *iii) LoRA-imp ($\mathcal{H}_{\text{imp}}$)*: Embeddings robust and invisible watermark signals within images.

During the training phase, the backbone diffusion model remains frozen, and each LoRA head is trained independently or jointly with separate optimization objectives, as discussed earlier. In inference, each head is selectively activated or deactivated via the corresponding binary gates: $g_{\text{main}}, g_{\text{vis}}, g_{\text{imp}} \in \{0, 1\}$, allowing the runtime configuration to flexibly determine the generated outputs' watermark properties. Formally, the weight matrices of the multi-head LoRA are represented as low-rank decompositions added to the frozen pre-trained weight matrix W. Specifically, each LoRA head is defined as follows:

$$W' = W + g_{\text{main}} B_{\text{main}} A_{\text{main}} + g_{\text{vis}} B_{\text{vis}} A_{\text{vis}} + g_{\text{imp}} B_{\text{imp}} A_{\text{imp}} \tag{12}$$

where W represents the frozen weight matrix of the backbone diffusion model and each LoRA head consists of two low-rank matrices $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$, with $r \ll d$ denoting the low-rank dimension.

For practical control of watermark visibility, we introduce authorization via digitally signed license files issued upon commercial purchase or subscription. A typical license file includes: (i) *User Identifier*, uniquely identifying the authorized user; (ii) *Hardware or Account Binding*, restricting activation to a specific deployment environment; (iii) *Gate Control Parameters*, setting $g_{\text{vis}} = 0$ upon authorization to disable explicit watermarking; and (iv) *Digital Signature*, generated using RSA or ECDSA, to ensure license authenticity and integrity. At service initialization, the following validation processes are performed: (i) read the license file and verify its digital signature to prevent unauthorized modifications; (ii) verify the hardware or account bindings specified in the license file; and (iii) upon successful validation, update the gate value to reflect the authorized usage settings. For a valid license, the gate value is typically set to $(g_{\text{main}}, g_{\text{vis}}, g_{\text{imp}}) = (1, 0, 1)$, which allows the main function to be enabled while

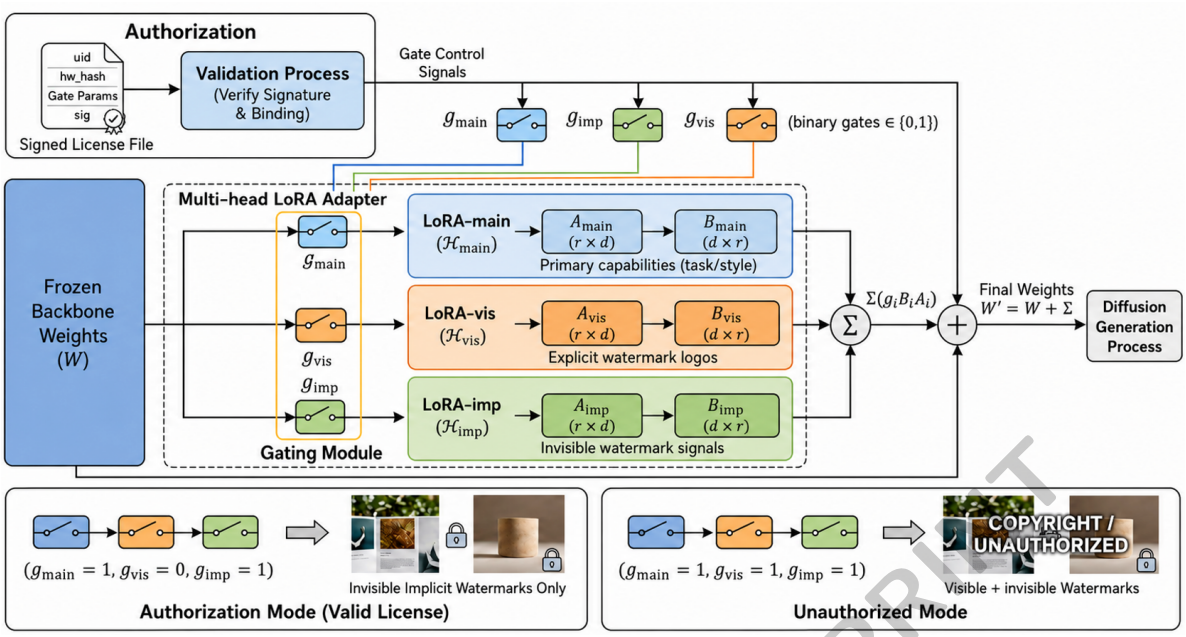

Figure 2: Overview of $LoRA^2D$ framework's multi-head LoRA adapters architecture and authorization management.

implicitly embedding the watermark without displaying the explicit watermark. Well-designed gating mechanisms enable different runtime watermarking behaviors based on the user's authorization status:

- Unauthorized Mode ($g_{\text{vis}} = 1$): The explicit LoRA watermark is active, producing a clearly visible logo pattern in the resulting image that serves as an important reminder of unauthorized use. Meanwhile, the implicit watermark remains embedded at all times to ensure strong copyright validation.
- Authorization Mode ($g_{\text{vis}} = 0$): Explicit watermark generation is disabled to avoid obvious visual distractions to authorized users. However, implicit watermark embedding is kept active ($g_{\text{imp}} = 1$), enabling continuous copyright tracking and black-box validation by securing ownership through invisible watermarking signals.

This comprehensive authorization and watermark control design ensures that explicit watermarks effectively prevent unauthorized redistribution, while implicit watermarks reliably and discreetly track model usage. This approach achieves commercial viability while rigorously safeguarding intellectual property.

5 Experiments

5.1 Experimental Setting

Model and Datasets. We evaluate our method on four task-specific datasets designed for stylized text-to-image generation. These include two open-domain character datasets, *One Piece* [YaYaB, 2022], *Simpsons* [Norod78,], and artist-style datasets *Wikiart* (Van Gogh and Picasso). All experiments are conducted using Stable Diffusion V1.5 as the backbone model. LoRA modules are fine-tuned on each dataset to enable task-specific generation, and the same backbone configuration is used across all methods to ensure fair comparison.

Evaluation Metrics. We evaluate the effectiveness of the watermark through the accuracy of its extraction. Additionally, Fresnel Epistatic Distance (FID) [Heusel *et al.*, 2017] and CLIP score [Radford *et al.*, 2021] are employed to evaluate the visual quality and semantic consistency of generated images.

Baseline. As LoRA specific watermarking remains largely unexplored, we construct comparison baselines by first embedding watermarks into the training images and subsequently fine-tuning LoRA modules on the watermarked datasets. We adopt one traditional watermarking method, DFT-SVD [Varghese *et al.*, 2016], two learning-based approaches: RoSteALS [Bui *et al.*, 2023] and TrustMark [Lu *et al.*, 2024]. For each baseline, watermark embedding is applied at the data level prior to LoRA training, and watermark presence is evaluated on images generated by the trained LoRA.

5.2 Experiment Results

Comparison Study. We evaluated the effectiveness of our implicit watermarking method against several on two distinct task-specific datasets. The primary criterion for evaluating watermarking effectiveness is the accuracy of implicit watermark extraction accuracy, which measures the ability to reliably decode embedded watermark information from generated images. For each watermarking method, we first generate images using a LoRA-adapted diffusion model trained on these datasets, then apply the trained watermark decoder to extract the implicit watermark information. We then computed the extraction accuracy for each dataset, then calculated the overall average in two types of task-specific datasets to de-

Dataset	Metrics	DFT-SVD	RoSteALS	TrustMark	w/o Watermark	Ours
Character datasets One Piece; Simpsons	Accuracy ↑	0.478	0.651	0.805	–	**0.913**
	FID↓	154.99	164.73	166.00	151.73	**151.38**
	CLIP Score↑	29.04	28.90	27.20	29.31	**29.22**
Artist-style datasets Wikiart Van Gogh; Picasso	Accuracy ↑	0.489	0.491	0.755	–	**0.835**
	FID↓	**239.23**	251.08	251.34	236.13	240.29
	CLIP Score↑	26.04	26.66	26.90	27.15	**26.91**

Table 1: Comparison of implicit watermark effectiveness with baseline method. The best results are highlighted in bold.

rive a performance metric. As shown in Table 1, our proposed $LoRA^2D$ consistently achieves superior watermark extraction accuracy in all evaluated datasets, outperforming both traditional watermarking baselines and advanced learning-based methods. Our method achieves a significantly higher average compared to the best-performing baseline, clearly demonstrating the robustness and reliability of our implicit watermark embedding strategy. These results highlight the effectiveness of $LoRA^2D$ in embedding imperceptible yet reliable watermark signals, effectively securing copyright ownership of LoRA modules in real-world deployment scenarios.

T2I Generation Quality Evaluation. We assess the fidelity of our watermarking method by examining whether the watermarked LoRA module retains the generative performance of the clean model. Using FID and CLIP scores as benchmarks for perceptual quality, we observe in Table 1 that our method yields the lowest FID across Character datasets. This indicates that the watermarking process causes minimal degradation to the image distribution. Additionally, the CLIP scores align closely with those of the non-watermarked model, demonstrating that semantic alignment is not compromised. Visual comparisons provided in Figure 3 support these quantitative findings, revealing that the watermarked model preserves stylistic integrity and visual quality, devoid of any visible distortions or artifacts.

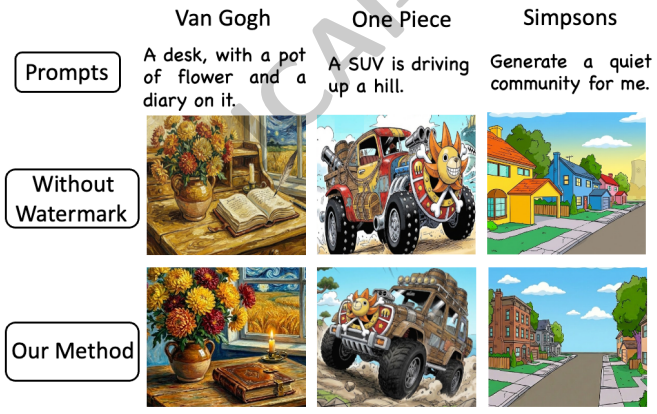

Figure 3: Comparison of images generated by the clean LoRA and the watermarked LoRA, highlighting the preservation of generative quality.

Robustness Study. The experimental results summarized in Figure 4 demonstrate that our proposed approach offers notably improved robustness against LoRA merging attacks compared to TrustMark. To reflect realistic attack conditions, we consider a scenario in which an attacker, lacking access to the original training dataset, attempts to compromise the watermark by merging the protected LoRA module with another unrelated LoRA module. Under such conditions, our method consistently achieves high watermark extraction accuracy, exhibiting only minor degradation across both character-based and artist-style datasets. In contrast, the baseline method suffers significant performance reductions, especially noticeable in character-based datasets. These findings confirm the inherent resilience of our watermark embedding strategy, highlighting its robustness to realistic parameter-merging attacks designed to evade watermark verification.

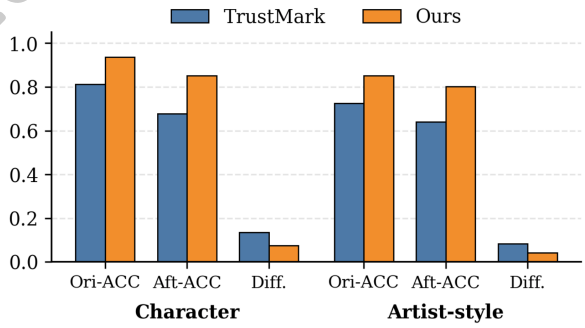

Figure 4: Comparison of our method and TrustMark under attacks.

6 Conclusion

Despite the growing popularity of T2I diffusion models, prior studies have largely neglected copyright protection tailored specifically for LoRA modules. To address this gap, we introduce the $LoRA^2D$ framework, an authorization-controlled dual-watermarking solution explicitly designed for LoRA copyright protection. Our approach integrates explicit watermarking, which visibly marks images generated in unauthorized or trial scenarios, and implicit watermarking, offering persistent, invisible encoding to enable robust ownership verification under adversarial alterations or unauthorized redistributions. Experiments confirm the effectiveness of our method, laying a strong foundation for secure and practical watermarking strategies in modular generative AI.

References

[Aldhaheri *et al.*, 2024] Lameya Aldhaheri, Noor Alshehhi, Irfana Ilyas Jameela Manzil, Ruhul Amin Khalil, Shumaila Javaid, Nasir Saeed, and Mohamed-Slim Alouini. Lora communication for agriculture 4.0: Opportunities, challenges, and future directions. *IEEE Internet of Things Journal*, 2024.

[Andrade and Yoo, 2019] Roberto Omar Andrade and Sang Guun Yoo. A comprehensive study of the use of lora in the development of smart cities. *Applied Sciences*, 9(22):4753, 2019.

[Borji, 2022] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022.

[Bui *et al.*, 2023] Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. Rosteals: Robust steganography using autoencoder latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 933–942, 2023.

[Cao *et al.*, 2024] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

[Chou *et al.*, 2023] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023.

[Feng *et al.*, 2024] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024.

[Heusel *et al.*, 2017] 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. *Advances in neural information processing systems*, 30, 2017.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Huan and Shun, 2025] Muchen Huan and Jianhong Shun. Fine-tuning transformers efficiently: A survey on lora and its impact. 2025.

[Kalajdzievski, 2023] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

[Kwon, 2024] Dong-Hyun Kwon. Analysis of prompt elements and use cases in image-generating ai: Focusing on midjourney, stable diffusion, firefly, dall· e. *Journal of Digital Contents Society*, 25(2):341–354, 2024.

[Liu *et al.*, 2023] Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*, 2023.

[Liu *et al.*, 2024] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):33, 2024.

[Lu *et al.*, 2024] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024.

[Luo *et al.*, 2025] Zihao Luo, Xilie Xu, Feng Liu, Yun Sing Koh, Di Wang, and Jingfeng Zhang. Privacy-preserving low-rank adaptation against membership inference attacks for latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5883–5891, 2025.

[Ma *et al.*, 2025] Zhiyuan Ma, Yuzhu Zhang, Guoli Jia, Liangliang Zhao, Yichao Ma, Mingjie Ma, Gaofeng Liu, Kaiyan Zhang, Ning Ding, Jianjun Li, et al. Efficient diffusion models: A comprehensive survey from principles to practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Norod78,] Norod78. Simpsons-blip-captions. https://huggingface.co/datasets/Norod78/simpsons-blip-captions.

[Parthasarathy *et al.*, 2024] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*, 2024.

[Radford *et al.*, 2021] 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.

[Ray and Roy, 2020] Arkadip Ray and Somaditya Roy. Recent trends in image watermarking techniques for copyright protection: a survey. *International Journal of Multimedia Information Retrieval*, 9(4):249–270, 2020.

[Rombach *et al.*, 2022] 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.

[Shen *et al.*, 2025] Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, et al. Efficient diffusion models: A survey. *arXiv preprint arXiv:2502.06805*, 2025.

[Sundaram *et al.*, 2019] Jothi Prasanna Shanmuga Sundaram, Wan Du, and Zhiwei Zhao. A survey on lora networking: Research problems, current solutions, and open issues. *IEEE Communications Surveys & Tutorials*, 22(1):371–388, 2019.

[Tu *et al.*, 2024] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *Visual Intelligence*, 2(1):34, 2024.

[Varghese *et al.*, 2016] Justin Varghese, Saudia Subash, Omer Bin Hussain, Krishnan Nallaperumal, Mohammed Ramadan Saady, and Mohamed Samiulla Khan. An improved digital image watermarking scheme using the discrete fourier transform and singular value decomposition. *Turkish Journal of Electrical Engineering and Computer Sciences*, 24(5):3432–3447, 2016.

[Wang *et al.*, 2025a] Zichong Wang, Zhipeng Yin, Zhen Liu, Roland HC Yap, Xiaocai Zhang, Shu Hu, and Wenbin Zhang. Redefining fairness: A multi-dimensional perspective and integrated evaluation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 336–353. Springer, 2025.

[Wang *et al.*, 2025b] Zichong Wang, Zhipeng Yin, Liping Yang, Jun Zhuang, Rui Yu, Qingzhao Kong, and Wenbin Zhang. Fairness-aware graph representation learning with limited demographic information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 354–371. Springer, 2025.

[Wang *et al.*, 2026a] Zichong Wang, Zhipeng Yin, Zhong Chen, Jack Yang, Jun Liu, and Wenbin Zhang. Learning counterfactual fairness from authentic generation. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026b] Zichong Wang, Zhipeng Yin, Mo Sha, Xiaofeng Gao, Xiaoli Li, and Wenbin Zhang. Towards fair graph learning without demographic supervision. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026c] Zichong Wang, Zhipeng Yin, and Wenbin Zhang. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. *Advances in Neural Information Processing Systems*, 38:64883–64909, 2026.

[Yang *et al.*, 2024] Menglin Yang, Jialin Chen, Jinkai Tao, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Regina Zhang, Min Zhou, et al. Low-rank adaptation for foundation models: A comprehensive review. *arXiv preprint arXiv:2501.00365*, 2024.

[YaYaB, 2022] YaYaB. One piece blip captions. https://huggingface.co/datasets/YaYaB/onepiece-blip-captions/, 2022.

[Yin *et al.*, 2025] Zhipeng Yin, Zichong Wang, Avash Palikhe, Zhen Liu, Jun Liu, and Wenbin Zhang. Amcr: A framework for assessing and mitigating copyright risks in generative models. *28th European Conference on Artificial Intelligence*, 2025.

[Yin *et al.*, 2026a] Zhipeng Yin, Zichong Wang, Zhong Chen, Jack Yang, Xin Ning, and Wenbin Zhang. Disentangled graph-enhanced large language models for fair learning. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Yin *et al.*, 2026b] Zhipeng Yin, Zichong Wang, Weifeng Xu, Jun Zhuang, Pallab Mozumder, Antoinette Smith, and Wenbin Zhang. *Digital Forensics in the Age of Large Language Models*, pages 55–82. Springer Nature Switzerland, 2026.

[Zhang and Ntoutsi, 2019] Wenbin Zhang and Eirini Ntoutsi. Faht: An adaptive fairness-aware decision tree classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1480–1486. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Zhang and Weiss, 2021] Wenbin Zhang and Jeremy C Weiss. Fair decision-making under uncertainty. In *2021 IEEE international conference on data mining (ICDM)*, pages 886–895. IEEE, 2021.

[Zhang and Weiss, 2022] Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, 2022.

[Zhang *et al.*, 2018] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pages 159–172, 2018.

[Zhang *et al.*, 2021] Wenbin Zhang, Liming Zhang, Dieter Pfoser, and Liang Zhao. Disentangled dynamic graph deep generation. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021.

[Zhang *et al.*, 2023a] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

[Zhang *et al.*, 2023b] Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[Zhang, 2024a] Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3):386–395, 2024.

[Zhang, 2024b] Wenbin Zhang. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

[Zhao *et al.*, 2023] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[Zhou and Nabus, 2023] Kai-Qing Zhou and Hatem Nabus. The ethical implications of dall-e: Opportunities and challenges. *Mesopotamian Journal of Computer Science*, 2023:16–21, 2023.