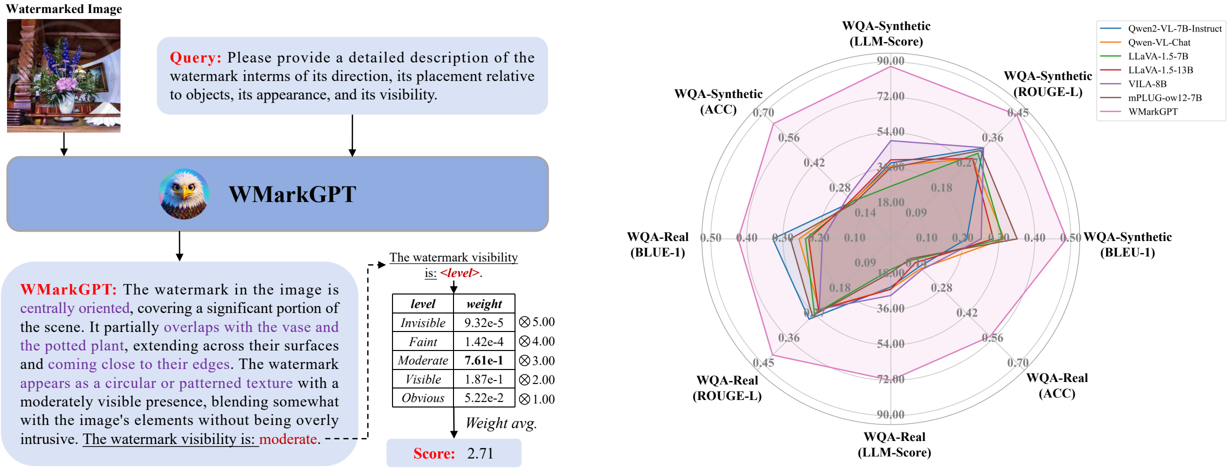
WMarkGPT: Watermarked Image Understanding via Multimodal Large Language Models

Songbai Tan [1 2] **Xuerui Qiu** [3] **Yao Shu** [2] **Gang Xu** [2] **Linrui Xu** [4] **Xiangyu Xu** [5]
Huiping Zhuang [6] **Ming Li** [2] **Fei Yu** [2]



Query: Please provide a detailed description of the watermark interms of its direction, its placement relative to objects, its appearance, and its visibility.

WMarkGPT: The watermark in the image is centrally oriented, covering a significant portion of the scene. It partially overlaps with the vase and the potted plant, extending across their surfaces and coming close to their edges. The watermark appears as a circular or patterned texture with a moderately visible presence, blending somewhat with the image's elements without being overly intrusive. The watermark visibility is: moderate.

The watermark visibility is: *<level>*.

level	*weight*	
Invisible	9.32e-5	⊗5.00
Faint	1.42e-4	⊗4.00
Moderate	**7.61e-1**	⊗3.00
Visible	1.87e-1	⊗2.00
Obvious	5.22e-2	⊗1.00

Score: 2.71

(a) **(b)**

Figure 1. (a) Our WMarkGPT is the first multimodal large language model specifically designed for watermarked image content understanding *without requiring access to the original images*, which are often unavailable in text-driven generative watermarking scenarios. It excels in accurately localizing watermarks, generating detailed semantic descriptions of their characteristics and their impact on image content, and evaluating visibility levels—capabilities that surpass conventional statistical metrics, which depend on original images and are limited to measuring low-level differences. (b) Compared to existing state-of-the-art MLLMs, WMarkGPT demonstrates significantly improved performance across various evaluation metrics.

Abstract

Invisible watermarking is widely used to protect digital images from unauthorized use. Accurate assessment of watermarking efficacy is crucial for advancing algorithmic development. However, existing statistical metrics, such as PSNR, rely on access to original images, which are often unavailable in text-driven generative watermarking and fail to capture critical aspects of watermarking, particularly visibility. More importantly, these metrics fail to account for potential corruption of image content. To address these limitations, we propose WMarkGPT, the first multimodal large language model (MLLM) specifically designed for comprehensive watermarked image understanding, *without accessing original images*. WMarkGPT not only predicts watermark visibility but also generates detailed textual descriptions of its location, content, and impact on image semantics, enabling a more nuanced interpretation of watermarked images. Tackling the challenge of precise location description and understanding images with vastly different content, we construct three visual question-answering (VQA) datasets: an object *location-aware* dataset, a *synthetic* watermarking dataset, and a *real* watermarking dataset. We introduce a meticulously designed three-stage learning pipeline to progressively equip WMarkGPT with the necessary abilities. Extensive experiments

[1] School of management, Shenzhen University [2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) [3] Institute of automation, Chinese Academy of Sciences [4] School of Geosciences and Info-Physics, Central South University [5] Xi'an Jiaotong University, China [6] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology. Correspondence to: Ming Li <liming@gml.ac.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

on synthetic and real watermarking QA datasets demonstrate that WMarkGPT outperforms existing MLLMs, achieving significant improvements in visibility prediction and content description. The datasets and code are released at https://github.com/TanSongBai/WMarkGPT.

1. Introduction

Invisible watermarking has been extensively utilized in digital images to prevent unauthorized use by embedding discernible information, such as 2D logos, which can later be retrieved by specific extractors to provide verifiable proof for various applications (Hosny et al., 2024; Rezaei et al., 2025; Sharma et al., 2024). This is particularly significant in text-to-image generation domains, which hold immense potential for commercial applications (Ma et al., 2024; Fernandez et al., 2023). The primary challenge in watermarking lies in achieving a delicate balance between minimizing its impact on image quality and maintaining robust detectability under various conditions (Guo et al., 2024; Fernandez et al., 2022). As a result, there has been growing interest in the research community toward evaluating the visual and structural effects of watermarking on digital images (Sharma et al., 2024).

Conventional assessments primarily rely on pixel-wise statistical metrics, such as peak signal-to-noise ratio (PSNR) (Korhonen & You, 2012), structural similarity index (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018a), to quantify low-level differences between original and watermarked images (Rachmawanto et al., 2024). However, accessing original images as references is often impractical for text-driven generative watermarking (Ma et al., 2024; Fernandez et al., 2023). Moreover, they often fail to accurately reflect watermarking efficacy, particularly in terms of watermark visibility. As shown in Fig. 2, these metrics often fail to align with human perception. Even for clearly visible watermarks, the metric values are unexpectedly favorable, highlighting their limitations in accurately assessing watermark visibility. To quantify the degree of misalignment, we calculate the Spearman's rank correlation coefficient (SRCC) (Sedgwick, 2014), Pearson linear correlation coefficient (PLCC) (Sedgwick, 2012), and Kendall's rank correlation coefficient (KRCC) (Abdi, 2007) between these metrics and human annotations for watermarking efficacy. As shown in Fig. 3, the comparisons reveal that traditional metrics significantly deviate from human perception. More importantly, these metrics fail to offer a comprehensive evaluation of the broader impact of watermarking on image content.

Recently, multimodal large language models (MLLMs) have achieved significant breakthroughs in image understanding and question answering tasks, effectively addressing conventional vision and language challenges with unprecedented capabilities (Liu et al., 2024b; Zhu et al., 2023; Xue et al., 2024; Liu et al., 2024a). Notably, pioneering efforts such as the large language and vision assistant (LLaVA) integrate a vision encoder with a LLM and leverage GPT-4-generated multimodal data to build robust systems excelling in visual understanding, instruction-following, and complex question answering tasks (Liu et al., 2024b). However, none of these MLLMs can be directly applied to watermarked image understanding, even with nuanced prompting, as they are trained on natural images whose data distribution and semantic content remain unaltered by the integration of distinctly different logo images.

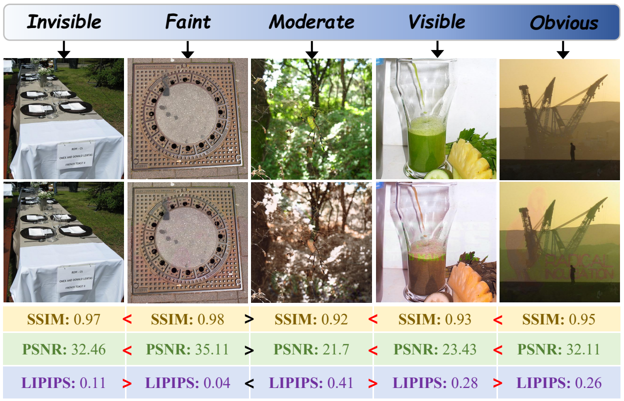

Invisible
Faint
Moderate
Visible
Obvious
SSIM: 0.97 < SSIM: 0.98 > SSIM: 0.92 < SSIM: 0.93 < SSIM: 0.95
PSNR: 32.46 < PSNR: 35.11 > PSNR: 21.7 < PSNR: 23.43 < PSNR: 32.11
LIPIPS: 0.11 > LIPIPS: 0.04 < LIPIPS: 0.41 > LIPIPS: 0.28 > LIPIPS: 0.26

Figure 2. Examples of original and watermarked image pairs with varying levels of watermarking efficacy, as annotated by humans. Traditional similarity metrics, including SSIM (↑), PSNR (↑), and LPIPS (↓), are computed for each pair. The greater and less signs between two samples indicate the relative magnitude of the metric values. Red signs indicate inconsistencies with human perception, while black signs represent consistent relations. The results demonstrate that traditional metrics often fail to accurately reflect watermarking efficacy, particularly in alignment with human evaluations.

To overcome these limitations, we introduce a new MLLM for watermarked image understanding, termed WMarkGPT. This model generates textual descriptions of watermark content, locations, detailed interactions with the main image content, and visibility prediction *with only a watermarked image*, making it particularly suitable for text-to-image generative watermarking, as illustrated in Fig. 1. To achieve it, the model must possess the ability to precisely describe object locations, identify a watermark from desirable foreground contents, and assess its visibility in a watermarked image. To support this learning, we construct three custom visual question-answering (VQA) datasets. The first is an object location-aware dataset, built upon the COCO dataset, which includes image captions and object bounding boxes. GPT-4 is used to generate 100k QA pairs that capture both absolute and relative object positions within a natural image. The second is a watermarking QA dataset, created with 50k synthetic watermarked images using a semi-automatic annotation pipeline. The final dataset is a real watermarking QA dataset, consisting of 2.5k watermarked images generated by several state-of-the-art watermarking algorithms, along with human annotations including watermark visibility and overall scores. The scores are used exclusively for evaluation purposes.

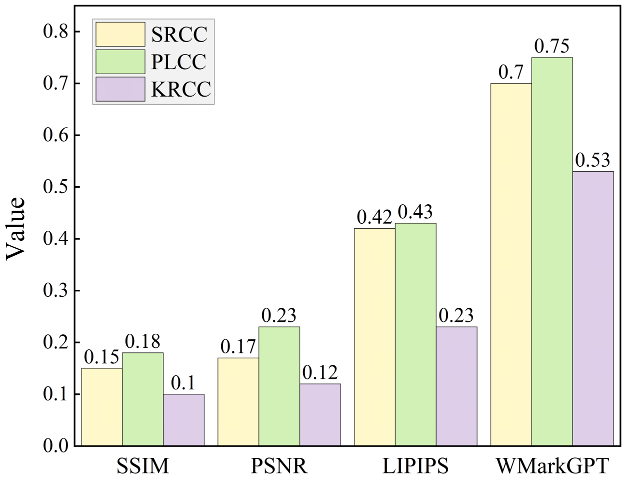


Figure 3. Comparison of various methods for assessing watermarking efficacy, measured by their consistency with human annotations. Higher SRCC, PLCC, and KRCC values indicate better alignment with human perception. Traditional statistical metrics, such as SSIM, PSNR, and LPIPS, show significantly lower consistency and fail to accurately reflect watermarking efficacy. In contrast, our proposed WMarkGPT achieves substantially higher consistency with human annotated scores, demonstrating its superior performance. *Notably*, WMarkGPT is trained only with coarse visibility level descriptions rather than precise scores.

To train our WMarkGPT, we propose a meticulously devised three-stage learning pipeline to progressively endow the model with robust capabilities for understanding watermarked images. Our model architecture comprises a visual encoder, learnable queries, a visual abstractor, and an LLM. In the first stage, we primarily train the visual encoder, learnable queries, and visual abstractor on the object location-aware natural image dataset, enabling the model to precisely recognize object positions. In the second stage, we optimize the entire model on the synthetic watermarking dataset, adapting it to handle images with corrupted data distributions and vastly different semantic contents. Finally, in the third stage, we fine-tune the visual encoder and visual abstractor on the real watermarking dataset, refining the model's performance and familiarizing it with target data distribution. Experimental results on the real watermarking QA dataset demonstrate that our WMarkGPT outperforms existing mainstream MLLMs by approximately 29%, 45%, 150%, and 217% across four metrics, respectively (see Sec. 4.1), highlighting its powerful capabilities in understanding both watermarks and their interactions with relevant semantics.

To summarize, our contributions in this work are three-fold:

- We propose WMarkGPT, the first multimodal large language model tailored for watermarked image understanding. It demonstrates superior performance in generating comprehensive textual descriptions and predicting watermark visibility *without requiring access to the original images*, effectively addressing the significant limitations of existing evaluation methods.
- We construct three visual question-answering benchmark datasets, especially the synthetic and real watermarking datasets, to facilitate research on object location-aware MLLMs and watermarking assessments. These datasets will be publicly released to advance future research.
- We propose a systematic learning paradigm to progressively endow a model with the capability to understand object positioning relationships and corrupted image semantics. To the best of our knowledge, this represents the first attempt to train MLLMs on unnatural images, each composed of a fusion of two significantly different images.

2. Watermarking Question-Answering Datasets

Image watermarking techniques have been extensively explored within the field of computer vision. However, large-scale watermark QA datasets tailored for the fine-tuning of MLLMs (e.g., our WMarkGPT) remain scarce, which hence significantly limits the development of MLLMs capable of comprehensive watermarked image understanding. Prior research has shown the promise of leveraging large-scale generative models like GPT, to synthesize data for multimodal applications, driving substantial progress in model capabilities (Jiao et al., 2024; Liu et al., 2024c; Wu et al., 2023b). Inspired by this, we introduce WQA-Synthetic, the first synthetic watermark QA dataset constructed via a semi-automated pipeline augmented with human oversight on large-scale generative models for comprehensive watermarked image understanding. This dataset comprises artificially generated watermarked images, corresponding questions, detailed watermark descriptions, and associated visibility levels. Furthermore, to capture real-world watermarking scenarios, we conduct subjective experiments and create WQA-Real, the first dataset derived from authentic watermarking cases for improved watermarked image understanding. The distributions of visibility levels in both datasets are illustrated in Fig. 4.

2.1. WQA-Synthetic Dataset

To enhance the capability of MLLMs to understand the details of watermarked images comprehensively, we con-

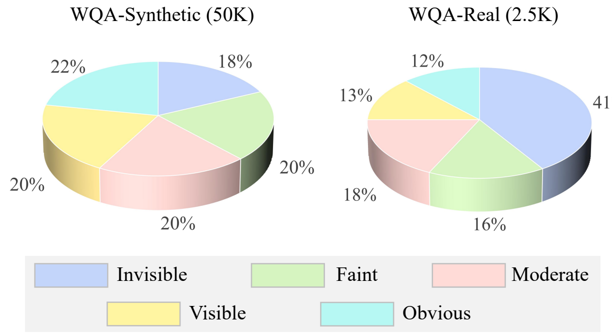


Figure 4. Data distribution of the collected WQA-Synthetic and WQA-Real datasets.

struct a synthetic watermark QA dataset, WQA-Synthetic, through a cost-effective and efficient semi-automatic annotation pipeline introduced below. The dataset consists of 50K data pairs in the format of $\{\mathbf{I}, \mathbf{Q}, \mathbf{T}, \mathbf{V}\}$, where $\mathbf{I}$ represents the watermarked image, $\mathbf{Q}$ denotes the related question, $\mathbf{T}$ is the watermark description, and $\mathbf{V}$ indicates the degree of watermark visibility. To ensure a comprehensive coverage of watermark characteristics, we utilize GPT-4 to generate 30 diverse question templates focusing on different aspects of watermark analysis. For each watermarked image, one question template is randomly selected as $\mathbf{Q}$. The watermark descriptions in the dataset include the position of watermark, its relationship with the main object in the image, watermark-specific features, and the visibility level of watermark, as illustrated in the part 4 of Fig. 5. This structured approach aligns with human perception, where watermark evaluation is not limited to overall visibility but also considers spatial position and feature attributes. Such detailed annotations provide a more comprehensive understanding of watermark content, enabling MLLMs to bridge the gap between perceptual and analytical insights.

The semi-automatic annotation pipeline of WQA-Synthetic is shown in Fig. 5. We randomly select 50K images from the COCO dataset (Lin et al., 2014) and 50K watermark logos from the LOGO-2K dataset (Wang et al., 2020), and then synthesize watermarked images while generating corresponding watermark descriptions using a four-step process: **(1) Step-1:** watermark segmentation; **(2) Step-2:** main object bounding box detection; **(3) Step-3:** watermarked image synthesis; and **(4) Step-4:** question-answering generation. Each stage is outlined in detail below.

Step 1: Watermark Segmentation (Part 1 in Fig. 5). During this stage, we aim to minimize background noise by carefully isolating the area where the watermark is located. Specifically, we use the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate a binary mask that precisely delineates the watermark region. To further refine the segmentation, a background filtering step is applied to eliminate extraneous pixels. The filtered mask is then mapped back onto the original watermarked image, yielding the segmented watermark and its boundary frame coordinates $\{\mathbf{b}_{x_1}^w, \mathbf{b}_{x_2}^w, \mathbf{b}_{y_1}^w, \mathbf{b}_{y_2}^w\}$.

Step-2: Main Object Bounding Box Detection (Part 2 in Fig. 5). To determine the watermark's relative position within the image, we extract bounding box information for the primary objects. With the object bounding box data from the COCO dataset, we employ GPT-4 to identify the main objects by analyzing their dimensions and spatial coordinates. The selected bounding boxes for these objects, denoted as $\{\mathbf{b}_{x_1}^i, \mathbf{b}_{x_2}^i, \mathbf{b}_{y_1}^i, \mathbf{b}_{y_2}^i\}$, are then collected for further processing.

Step-3: Watermarked Image Synthesis (Part 3 in Fig. 5). In this stage, we focus on integrating the selected watermark into the original image while maintaining the visual integrity and coherence. Firstly, the watermark is resized to fit within a predefined bounding box. Subsequently, the watermark is embedded at a random position within the image using a weighted average fusion method. The blending process is governed by a transparency factor α, which controls the visibility and prominence of the watermark. The integration procedure for each pixel within the target region can be mathematically expressed as:

$$\mathbf{I}_e(x, y) = (1 - \alpha) \cdot \mathbf{I}_c(x, y) + \alpha \cdot \mathbf{I}_w(x, y) \cdot \mathbf{M}(x, y), \quad (1)$$

where $\mathbf{I}_e$, $\mathbf{I}_c$, $\mathbf{I}_w$, and $\mathbf{M}$ represent the watermarked image, the original image, the watermark, and the filtered mask, respectively. To maintain visual consistency, a brightness adjustment is applied to the watermarked image $\mathbf{I}_e$ through a linear transformation:

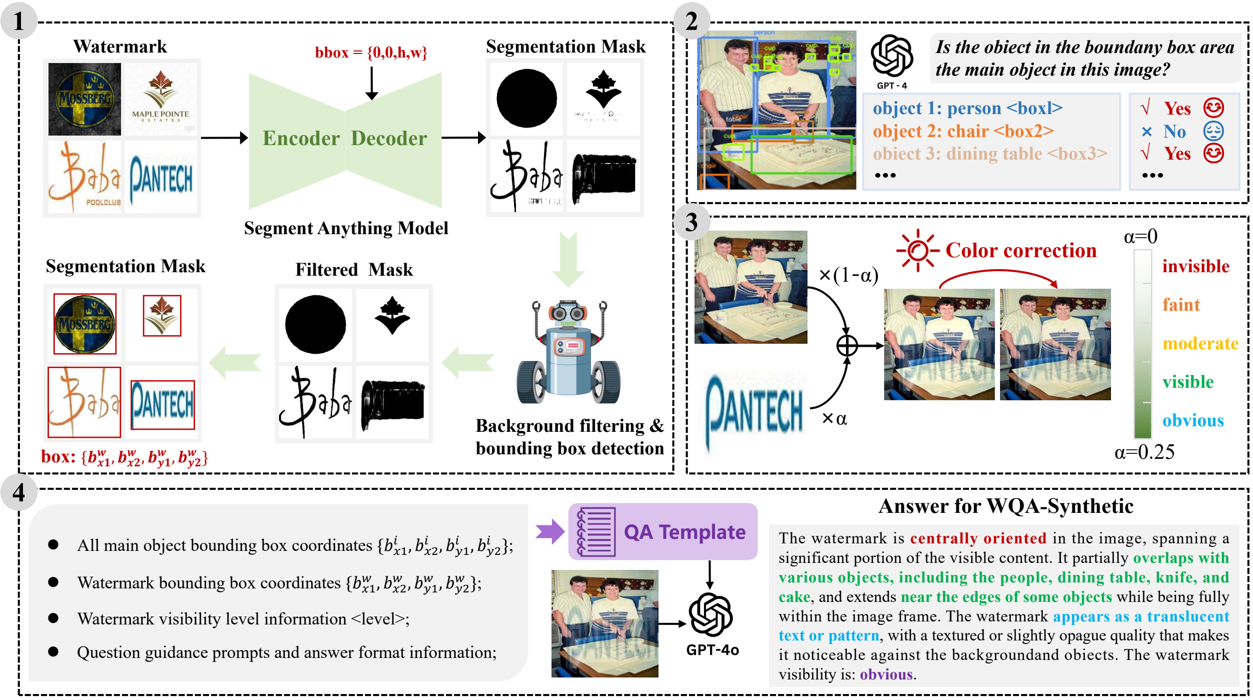
$$\mathbf{I}_e^{'}(x, y) = \frac{\mathbf{I}_e(x, y)}{\frac{1}{n}\sum_n^i \mathbf{I}_e} \cdot \frac{1}{n}\sum_n^i \mathbf{I}_c, \quad (2)$$

The value of α ranges from 0 to 0.25, with each increment of 0.05 corresponding to a one-level increase in watermark visibility. The visibility is categorized into five distinct levels:“invisible”,“faint”,“moderate”, “visible”, and “obvious”, correspondingly. The visibility level $\mathbf{V}$ of the watermark is determined by the chosen value of α, allowing for precise control over its perceptual appearance.

Step-4: Question-Answering Generation (Part 4 in Fig. 5). After the aforementioned stages, we obtain the watermarked image, the bounding box coordinates for both the watermark and the main object, as well as the watermark's visibility level. These information is then integrated into a QA template, which is fed into GPT-4o to generate a watermark description $\mathbf{T}$. This watermark description includes information about the watermark's location and other specific features. The data collection for WQA-Synthetic, consisting of $\{\mathbf{I}_e^{'}, \mathbf{Q}, \mathbf{T}, \mathbf{V}\}$, is then completed.

2.2. WQA-Real Dataset

In addition to the synthetic watermark QA dataset, we also build a real-world dataset, WQA-Real, to further improve the capability of MLLMs to understand watermarked images comprehensively. Different from WQA-Synthetic dataset,



Is the obiect in the boundany box area the main object in this image?

- All main object bounding box coordinates $\{b^i_{x1}, b^i_{x2}, b^i_{y1}, b^i_{y2}\}$;
- Watermark bounding box coordinates $\{b^w_{x1}, b^w_{x2}, b^w_{y1}, b^w_{y2}\}$;
- Watermark visibility level information <level>;
- Question guidance prompts and answer format information;

The watermark is **centrally oriented** in the image, spanning a significant portion of the visible content. It partially **overlaps with various objects, including the people, dining table, knife, and cake**, and extends **near the edges of some objects** while being fully within the image frame. The watermark **appears as a translucent text or pattern**, with a textured or slightly opague quality that makes it noticeable against the backgroundand obiects. The watermark visibility is: **obvious**.

Figure 5. Illustration of WQA-Synthetic semi-automatic annotation pipeline. The watermark description and quality score of watermark information are constructed from a given watermarked image through a four-step process.

the data consists of 2.5K data pairs of $\{\mathbf{I}, \mathbf{Q}, \mathbf{T}, \mathbf{V}\}$. Here, the generation of the related question $\mathbf{Q}$ is the same as the one detailed in Sec. 2.1.

Collection of Real Watermarked Images. To generate watermarked images that closely mirror real-world scenarios, we select a range of representative image watermarking models, including Hidden (Zhu, 2018), BalujaNet (Baluja, 2019), WengNet (Weng et al., 2019), HiNet (Jing et al., 2021), and Safe-SD (Ma et al., 2024). The first four models are post-processing techniques, while the latter is an in-generation approach. Using all five models, we synthesize 10K watermarked images. A real-world watermarked image dataset is then created by randomly selecting 2.5K of these images, ensuring they reflected the distribution of watermark visibility levels. This dataset provides our source of watermarked images $\mathbf{I}$.

Question-Answering and Watermark Visibility Annotating. To obtain watermark descriptions and assess the visibility of watermarked images in real-world scenarios, we conduct a subjective experiment with human participants who are trained to master relevant background knowledge. The participants are trained to describe the watermark by considering factors such as its position, distribution, content, relationship with the main object, features (e.g., transparency, texture), and visibility. These descriptions then form the source for the watermark description $\mathbf{T}$. After completing the description, participants assess the visibility of each watermarked image based on predefined criteria, which forms the watermark visibility $\mathbf{V}$.

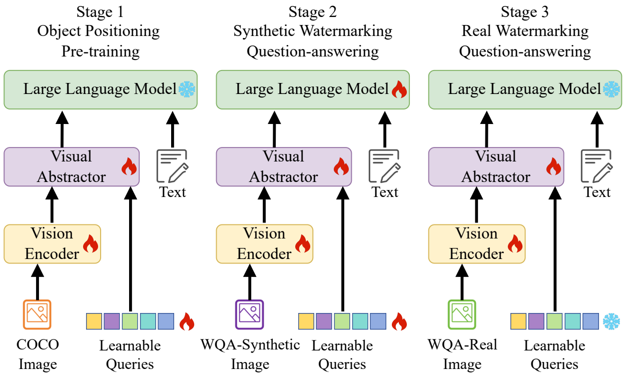


Figure 6. Model architecture and the progressive learning paradigm of WMarkGPT. We employ a meticulously designed three-stage training pipeline to sequentially optimize different model components, progressively enhancing performance and achieving optimal final results.

3. WMarkGPT

3.1. Model Architecture

WMarkGPT is a specialized multimodal large language model explicitly designed for watermarked image comprehension. By seamlessly integrating visual representations with linguistic embeddings, WMarkGPT enables the recognition and detailed articulation of watermarked image content without requiring original reference images. It addresses key limitations of traditional evaluation methods, including the reliance on reference images, the absence of watermark information and localization, lack of descriptions about semantic corruption, and insufficient visibility assessment. Notably, our visibility prediction logits yield scores highly consistent with human ratings, despite these ratings not being used to supervise the model training.

To facilitate these challenging functions, inspired by mPLUG-owl-2 (Ye et al., 2024), our WMarkGPT incorporates an LLM as its core component, complemented by a vision encoder, a visual abstractor, and learnable queries as depicted in Fig. 6. Specifically, WMarkGPT processes a 448×448 watermarked image as input, which is first handled by the vision encoder to produce 64 image tokens, each with a dimension of 1024. These tokens are combined with a set of learnable queries of the same shape and passed through the visual abstractor, where information exchange occurs via multi-head self-attention mechanisms, distilling task-specific semantic representations into the learnable queries. Finally, the refined learnable queries are concatenated with the encoded textual embeddings of a question and jointly processed by the LLM to generate the textual response.

3.2. Training Paradigm

Watermarked images are generated by embedding watermark patterns into original images, resulting in significant domain differences compared to naturally captured images. Addressing the task of describing watermarked images using MLLMs requires overcoming two major challenges. First, the model must accurately describe the spatial relationships between watermark distributions and primary objects within the image. Second, watermarked images differ fundamentally from natural images, as they are the fusion or combination of natural image content and watermark logos, containing two highly distinct visual components. To tackle these challenges, we design a three-stage vision-language training strategy, as illustrated in Fig. 6, to systematically endow the model with these capabilities. In the first stage, the model is trained on an object location-aware QA dataset based on natural images. In the second stage, it is further optimized on a synthetic watermarking QA dataset, and in the final stage, fine-tuned on a real watermarking QA dataset. Throughout all stages, cross-entropy loss is employed to measure the discrepancy between the predicted outputs and ground-truth labels.

Stage-1: Object Positioning Pre-training. This stage aims to enhance WMarkGPT's ability to perceive and understand object positioning within images. To achieve this, we construct the object location-aware QA dataset using the COCO dataset and GPT-4. The COCO dataset provides bounding box annotations for primary objects alongside image captions, forming the foundation for generating two types of question-answer pairs with GPT-4. For absolute positioning, GPT-4 generates questions about objects located in specific regions of the image, such as the top, bottom, left, or right. For relative positioning, GPT-4 creates questions regarding the spatial relationships between objects, such as identifying which objects are above, below, to the left, or to the right of a given object. Using this dataset, we train the vision encoder, visual abstractor, and learnable queries of WMarkGPT, enabling the model to effectively extract and process object positional features.

Stage-2: Synthetic Watermarking Question-answering. In this stage, all trainable parameters are unfrozen for supervised fine-tuning. The model is trained on our high-quality synthetic dataset, WQA-Synthetic, which consists of a large number of watermarked images. This dataset is designed to familiarize the model with the unnatural data distribution and question-answering tasks specific to watermarked images. Through this stage, the model's ability to generate detailed textual descriptions is significantly enhanced, including precise watermark positioning, relevant features, and visibility assessment.

Stage-3: Real Watermarking Question-answering. To further enhance the model's ability to process real watermarked images, we conduct additional fine-tuning using a small but high-quality dataset, WQA-Real, featuring meticulously annotated watermarked images. By this stage, the model has already developed a strong understanding of watermarked images through the earlier training phases. Therefore, we freeze the LLM and learnable queries, focusing fine-tuning exclusively on the vision encoder and visual abstractor components. This targeted fine-tuning ensures alignment with the real data distribution, enabling the model to generate accurate and detailed descriptions, thereby improving its performance.

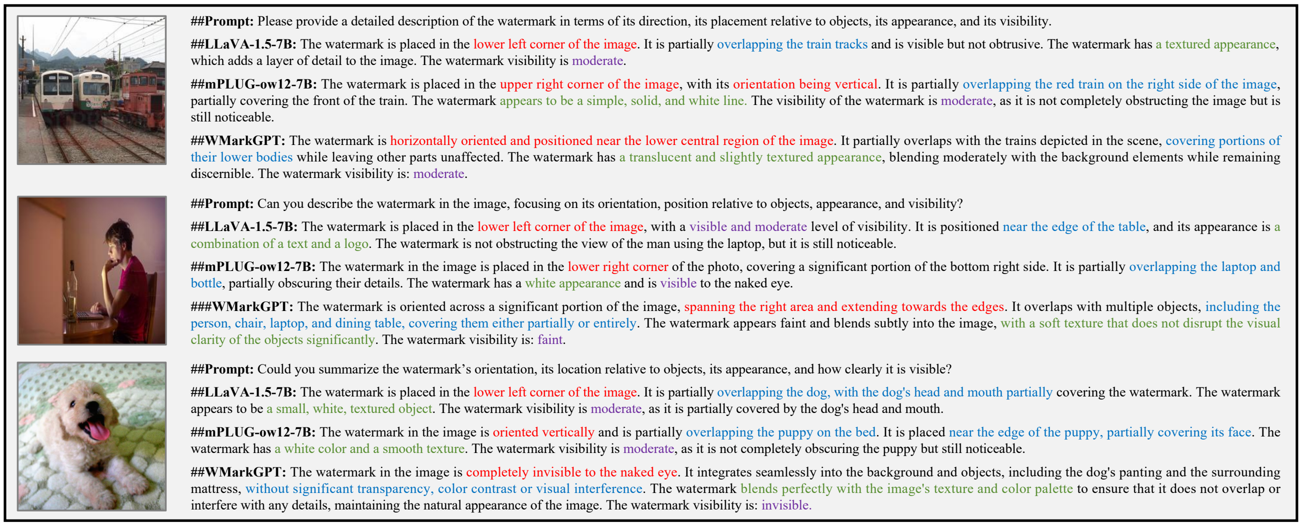
4. Experiments

4.1. Main Results

Quantitative Comparison. To demonstrate the effectiveness of WMarkGPT, we conduct a comprehensive quantitative comparison against several state-of-the-art multimodal large language models (MLLMs) on both the WQA-Synthetic and WQA-Real datasets, including Qwen2-VL-7B-Instruct (Wang et al., 2024), Qwen-VL-Chat (Bai et al., 2023), LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-1.5-13B (Liu et al., 2024a), VILA-8B (Lin et al., 2024), and mPLUG-owl-7B (Ye et al., 2024). The details of the evaluation metrics are available in Appx. G. To ensure a fair comparison, we standardize the input format for each model. The quan-

Table 1. Performance comparison between WMarkGPT and six state-of-the-art MLLMs on the WQA-Synthetic and WQA-Real datasets across four evaluation metrics. It shows that WMarkGPT consistently demonstrates superior performance on these two datasets.

Models	**Backbone**	**WQA-Synthetic**				**WQA-Real**			
		BLEU-1	ROUGE-L	LLM-Score	ACC	BLEU-1	ROUGE-L	LLM-Score	ACC
Qwen2-VL-7B-Instruct	Qwen-7B	0.211	0.326	38.645	0.209	0.328	0.289	24.450	0.156
Qwen-VL-Chat	Qwen-7B	0.315	0.289	37.078	0.211	0.255	0.253	25.450	0.166
LLaVA-1.5-7B	Vicuna-1.5-7B	0.311	0.308	26.778	0.210	0.237	0.270	15.600	0.118
LLaVA-1.5-13B	Vicuna-1.5-13B	0.285	0.289	40.220	0.217	0.228	0.256	25.650	0.134
VILA-8B	LLaMA-2-7B	0.251	0.328	49.961	0.236	0.190	0.253	28.800	0.172
mPLUG-ow12-7B	LLaMA-2-7B	0.351	0.318	36.229	0.201	0.279	0.277	17.150	0.110
WMarkGPT	LLaMA-2-7B	**0.488** (+39.0%)	**0.446** (+36.0%)	**87.751** (+75.6%)	**0.645** (+173.3%)	**0.424** (+29.3%)	**0.418** (+44.6%)	**71.950** (+149.8%)	**0.546** (+217.4%)



##Prompt: Please provide a detailed description of the watermark in terms of its direction, its placement relative to objects, its appearance, and its visibility.

##LLaVA-1.5-7B: The watermark is placed in the lower left corner of the image. It is partially overlapping the train tracks and is visible but not obtrusive. The watermark has a textured appearance, which adds a layer of detail to the image. The watermark visibility is moderate.

##mPLUG-ow12-7B: The watermark is placed in the upper right corner of the image, with its orientation being vertical. It is partially overlapping the red train on the right side of the image, partially covering the front of the train. The watermark appears to be a simple, solid, and white line. The visibility of the watermark is moderate, as it is not completely obstructing the image but is still noticeable.

##WMarkGPT: The watermark is horizontally oriented and positioned near the lower central region of the image. It partially overlaps with the trains depicted in the scene, covering portions of their lower bodies while leaving other parts unaffected. The watermark has a translucent and slightly textured appearance, blending moderately with the background elements while remaining discernible. The watermark visibility is: moderate.

##Prompt: Can you describe the watermark in the image, focusing on its orientation, position relative to objects, appearance, and visibility?

##LLaVA-1.5-7B: The watermark is placed in the lower left corner of the image, with a visible and moderate level of visibility. It is positioned near the edge of the table, and its appearance is a combination of a text and a logo. The watermark is not obstructing the view of the man using the laptop, but it is still noticeable.

##mPLUG-ow12-7B: The watermark in the image is placed in the lower right corner of the photo, covering a significant portion of the bottom right side. It is partially overlapping the laptop and bottle, partially obscuring their details. The watermark has a white appearance and is visible to the naked eye.

###WMarkGPT: The watermark is oriented across a significant portion of the image, spanning the right area and extending towards the edges. It overlaps with multiple objects, including the person, chair, laptop, and dining table, covering them either partially or entirely. The watermark appears faint and blends subtly into the image, with a soft texture that does not disrupt the visual clarity of the objects significantly. The watermark visibility is: faint.

##Prompt: Could you summarize the watermark's orientation, its location relative to objects, its appearance, and how clearly it is visible?

##LLaVA-1.5-7B: The watermark is placed in the lower left corner of the image. It is partially overlapping the dog, with the dog's head and mouth partially covering the watermark. The watermark appears to be a small, white, textured object. The watermark visibility is moderate, as it is partially covered by the dog's head and mouth.

##mPLUG-ow12-7B: The watermark in the image is oriented vertically and is partially overlapping the puppy on the bed. It is placed near the edge of the puppy, partially covering its face. The watermark has a white color and a smooth texture. The watermark visibility is moderate, as it is not completely obscuring the puppy but still noticeable.

##WMarkGPT: The watermark in the image is completely invisible to the naked eye. It integrates seamlessly into the background and objects, including the dog's panting and the surrounding mattress, without significant transparency, color contrast or visual interference. The watermark blends perfectly with the image's texture and color palette to ensure that it does not overlap or interfere with any details, maintaining the natural appearance of the image. The watermark visibility is: invisible.

Figure 7. Qualitative comparison of watermark descriptions between WMarkGPT and other open-sourced MLLMs on watermarked images with varying visibility, which demonstrates the ability of WMarkGPT to accurately identify watermark location and features, even when the watermark is not visible to the human eyes.

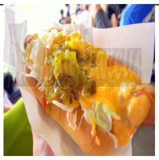
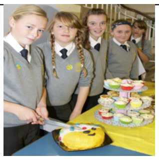
titative results, detailed in Tab. 1, show that WMarkGPT significantly outperforms the other models in terms of watermark description relevance, achieving higher BLEU-1, ROUGE-L, and LLM-Score values. Moreover, WMarkGPT demonstrates superior visibility prediction accuracy, achieving a 217.4% higher ACC than the second-best model, VILA-8B. To further investigate watermark description quality across different visibility levels, we evaluate these models on the WQA-Real dataset across five visibility categories. As shown in Tab. 2, WMarkGPT consistently achieves the highest BLEU-1 scores for watermarked image descriptions across all visibility levels, highlighting its robustness.

Table 2. The BLEU-1 results for watermark descriptions of the five different visibility levels in WQA-Real. Here, L_1, L_2, L_3, L_4, and L_5 represent the visibility levels of "invisible", "faint", "moderate", "visible", and "obvious", respectively.

Models	**WQA-Real**				
	L_1	L_2	L_3	L_4	L_5
Qwen2-VL-7B-Instruct	0.31	0.33	0.33	0.37	0.36
Qwen-VL-Chat	0.24	0.27	0.27	0.26	0.27
LLaVA-1.5-7B	0.24	0.23	0.23	0.26	0.23
LLaVA-1.5-13B	0.23	0.21	0.23	0.25	0.23
VILA-8B	0.19	0.16	0.19	0.21	0.21
mPLUG-ow12-7B	0.27	0.28	0.29	0.29	0.27
WMarkGPT	**0.44**	**0.38**	**0.41**	**0.44**	**0.43**

Qualitative Evaluation. In addition to quantitative comparisons, we perform a qualitative evaluation using watermarked images from diverse scenes in the test set, particularly those with subtle or absent watermarks (Fig. 7). This analysis reveals that WMarkGPT accurately identifies watermark position and features, even in challenging cases. In contrast, models like LLaVA-1.5-7B and mPLUG-owl-7B often provided less accurate descriptions, and, notably, misidentified background patterns as watermarks when the watermark is visually imperceptible. WMarkGPT, however, correctly recognizes the absence of a visible watermark,

Table 3. The visibility prediction probability distributions and quantification method for watermarked images with varying visibility levels. In the table, the watermark in (A) is more visible compared to (B), as indicated by its smaller weighted score. Here, $\mathbf{T}_i$ denotes the watermark description.

Watermarked Image	(A) 				
$\mathbf{T}_i$	The watermark is oriented horizontally across the central portion of the image. ... The watermark visibility is: **visible**.				
Level	L_1	L_2	L_3	L_4	L_5
Probability	9.39	6.24	12.84	**18.92**	17.27
Softmax	0.000	0.000	0.002	**0.837**	0.161
Weight avg.	1.84 (Range:[1:5])				
Watermarked Image	(B) 				
$\mathbf{T}_i$	The watermark is oriented horizontally across the lower portion of the image. ... The watermark visibility is: **faint**.				
Level	L_1	L_2	L_3	L_4	L_5
Probability	13.18	**17.63**	14.3	11.65	9.04
Softmax	0.011	**0.952**	0.034	0.002	0.000
Weight avg.	3.97 (Range:[1:5])				

demonstrating its robust performance. To further analyze watermark visibility quantitatively, we visualize the prediction results for two different visibility levels in Tab. 3. The descriptions from WMarkGPT conclude with the phrase "The watermark visibility is <level>", and we observed a strong correlation between the probability distribution of the <level> output and the actual watermark visibility. Following the approach in (Wu et al., 2023a), we assign numerical values (1-5) to the visibility levels and calculated a weighted score based on the probability of each level. This score objectively reflects watermark visibility, with lower values indicating more visible watermarks. As illustrated in Fig. 3, this quantitative method proves to be a more accurate measure of watermark visibility compared to traditional metrics like PSNR.

4.2. Ablation Study

Effects of Varying Size of WQA-Synthetic. To investigate the impact of the WQA-Synthetic dataset size on model performance, we train WMarkGPT using three different sizes of the training data: 0K, 10K, 30K, and 50K images, respectively. The experimental results, detailed in Tab. 4, demonstrate that model performance improves as the amount of synthetic training data increases. Specifically, using 10K images results in a 61.3% and 46.2% improvement in BLEU-1 and ROUGE-L scores, respectively, compared to the 0K results. Furthermore, increasing the training data from 0K to 50K images led to a 319.5% and 396.4% enhancement in LLM-Score and ACC metrics, respectively.

Table 4. The impact of using varying size of WQA-Synthetic data for fine-tuning. As the training proportion increases, the model's ability to understand watermarks improves, resulting in more accurate watermark descriptions.

Size	**WQA-Synthetic**			
	BLEU-1	ROUGE-L	LLM-Score	ACC
0K	0.351	0.318	36.229	0.201
10K	0.450	0.405	77.706	0.493
30K	0.448	0.407	79.267	0.555
50K	**0.488** (+8.9%)	**0.446** (+9.6%)	**87.751** (+10.7%)	**0.645** (+16.2%)

Table 5. The performance of the model on WQA-Real after three different training stages. S_1, S_2, and S_3 represent the **Stage-1**, **Stage-2**, and **Stage-3**, respectively.

Stages	**WQA-Real**			
	BLEU-1	ROUGE-L	LLM-Score	ACC
Baseline	0.279	0.277	17.150	0.110
S_1	0.281	0.280	30.324	0.235
$S_1 + S_2$	0.290	0.283	40.400	0.320
$S_1 + S_2 + S_3$	**0.424** (+46.2%)	**0.418** (+47.7%)	**71.950** (+78.1%)	**0.546** (+70.6%)

Effects of Progressive Training Pipeline. To assess the impact of our three-stage training pipeline, we evaluated the performance of WMarkGPT after each stage on the WQA-Real dataset. The results, detailed in Tab. 5, show a clear progression of improvement. After the **Stage-1**, the model achieved a 76.81% increase in LLM-Score and a 113.64% increase in visibility prediction accuracy (ACC) compared to the baseline. Following the **Stage-2**, all four evaluation metrics showed further improvement, with the LLM-Score increasing by 33.23%. Finally, the complete three-stage training resulted in significant enhancements over the second stage, with BLEU-1, ROUGE-L, LLM-Score, and ACC increasing by 46.20%, 47.70%, 78.09%, and 70.63%, respectively. These results demonstrate that our stepwise three-stage training effectively enhances the model ability to perceive watermark content.

5. Conclusion

This paper presents WMarkGPT, the first MLLM specifically designed for watermark content understanding. To train the model for watermark content perception, we developed three visual question-answering datasets: an object location-aware dataset, a synthetic watermarking dataset, and a real watermarking dataset. Furthermore, the model employs a three-stage training pipeline that progressively bridges the gap between natural images and watermarked images, achieving superior results in watermark understanding. Future work will explore the application of watermark understanding in other modalities, such as watermarked videos and 3D-generated watermarked content, to further advance watermarking technology.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (62306117 and 62302385), the Guangzhou Basic and Applied Basic Research Foundation (2024A04J3681), and GJYC program of Guangzhou (2024D03J0005).

Impact Statement

This study introduces WMarkGPT, the first multimodal large language model (MLLM) designed for watermarked image analysis without requiring access to original images, addressing a major limitation of existing evaluation metrics. WMarkGPT predicts watermark visibility and describes its location, content, and semantic impact, improving assessment reliability. This research has critical implications for watermark protection and digital content safety, enabling more robust evaluation of watermarking techniques against unauthorized removal and content manipulation.

References

Abdi, H. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510, 2007.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Baluja, S. Hiding images within images. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1685–1697, 2019.

Chen, Y., Sikka, K., Cogswell, M., Ji, H., and Divakaran, A. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14239–14250, 2024.

Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., and Douze, M. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3054–3058. IEEE, 2022.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models, 2023.

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1220–1230, 2022.

Guo, J., Li, Y., Wang, L., Xia, S.-T., Huang, H., Liu, C., and Li, B. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *Advances in Neural Information Processing Systems*, 36, 2024.

Hosny, K. M., Magdi, A., ElKomy, O., and Hamza, H. M. Digital image watermarking using deep learning: A survey. *Computer Science Review*, 53:100662, 2024.

Jiao, Q., Chen, D., Huang, Y., Li, Y., and Shen, Y. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*, 2024.

Jing, J., Deng, X., Xu, M., Wang, J., and Guan, Z. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4733–4742, 2021.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Korhonen, J. and You, J. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pp. 37–38. IEEE, 2012.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F. R., Ji, X., and Li, M. Eventgpt: Event stream understanding with multimodal large language models. *arXiv preprint arXiv:2412.00832*, 2024c.

Lu, Y., Yang, X., Li, X., Wang, X. E., and Wang, W. Y. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

Ma, Z., Jia, G., Qi, B., and Zhou, B. Safe-sd: Safe and traceable stable diffusion with text prompt trigger for invisible generative watermarking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7113–7122, 2024.

Rachmawanto, E. H., Hidajat, M. S., Hermanto, D., Wibowo, D. A., Pratama, Z., Setiarso, I., Astuti, E. Z., Handoko, L. B., and Yaacob, N. M. Quality measurement for imperceptibility watermarking based on psnr and ssim using walsh hadamard transform. In *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 353–358. IEEE, 2024.

Rezaei, A., Akbari, M., Alvar, S. R., Fatemi, A., and Zhang, Y. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pp. 118–136. Springer, 2025.

Sedgwick, P. Pearson's correlation coefficient. *Bmj*, 345, 2012.

Sedgwick, P. Spearman's rank correlation coefficient. *Bmj*, 349, 2014.

Sharma, S., Zou, J. J., Fang, G., Shukla, P., and Cai, W. A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, 83(11): 31829–31891, 2024.

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., Wang, H., and Jiang, S. Logo-2k+: A large-scale logo dataset for scalable logo classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6194–6201, 2020.

Wang, J., Chan, K. C., and Loy, C. C. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Weng, X., Li, Y., Chi, L., and Mu, Y. High-capacity convolutional video steganography with temporal residual modeling. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pp. 87–95, 2019.

Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023a.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023b.

Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M. S., et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., and Huang, F. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018b.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, J. Hidden: hiding data with deep networks. *arXiv preprint arXiv:1807.09937*, 2018.

Appendix

In this appendix, we provide additional technical details, dataset construction methods, and supplementary experiments. Sec. A provides a detailed summary of related works. Sec. B details how our model computes the visibility score for watermarked images. Sec. C presents the text templates used in the production of the object location-aware dataset. Sec. D presents the text templates used in the semi-automatic annotation process of WQA-Synthetic and showcases intermediate results. Sec. E explains the details and standards of the subjective experiments conducted for WQA-Real. Sec. F describes the implementation details. Sec. G elaborates on the evaluation metrics, particularly the LLM-Score and ACC metrics. Sec. H compares our model's visibility score predictions with other score-based methods, highlighting the advantages of our approach.

A. Related Works

Multimodal Large Language Models. Early works like BLIP (Li et al., 2022) and Flamingo (Alayrac et al., 2022) have demonstrated strong cross-modal understanding capabilities through large-scale image-text alignment pre-training. Further advancements, such as MiniGPT-4 (Zhu et al., 2023) and Next-GPT (Wu et al., 2023b), enable transformations between arbitrary modes, while models like BLIP-2 (Li et al., 2023) and DRESS (Chen et al., 2024) incorporate human feedback to improve the alignment with human intentions. Improved datasets, as seen in LLaVA-1.5 (Liu et al., 2024a), and optimizations in attention mechanisms, such as in mPLUG (Ye et al., 2024), have further boosted performance and efficiency. However, most existing MLLMs focus on natural image understanding for tasks like text-image matching or dialogue generation, lacking optimization for watermarked image comprehension. Our work addresses this gap by training MLLMs to analyze unnatural images formed through the fusion of distinct contents, enabling the understanding of object positioning, watermark influence, and visibility assessment.

Watermarking Efficacy Evaluation. Traditional methods for evaluating watermarking primarily rely on pixel-based metrics such as peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and mean squared error (MSE), which quantify pixel-level differences between original and watermarked images. Recent learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018a) is a perceptual similarity metric that measures differences between images based on deep features extracted from a pre-trained neural network, capturing perceptual differences beyond pixel-level distortions. While these metrics effectively capture low-level distortions or feature-level shifts, they fail to accurately assess watermark visibility and influence on semantics. Moreover, existing evaluation methods rely on access to both original and watermarked images. In contrast, our work trains an MLLM to evaluate watermarked images without requiring original references. It provides detailed textual descriptions, including the watermark's location, content, its impact on the image's semantics, and visibility assessment, addressing the limitations of conventional approaches.

B. Visibility Score Prediction

After training, the model demonstrates strong capabilities in summarizing the content of watermarked images and accurately judging the visibility of the watermark. Its output is a plain text description, with the final sentence following a fixed format: "The watermark visibility is: $< level >$." To convert the rating levels back into scores, we define a reverse mapping G from text-defined rating levels to numeric scores, as follows:

$$G : L_i \rightarrow i, \tag{1}$$

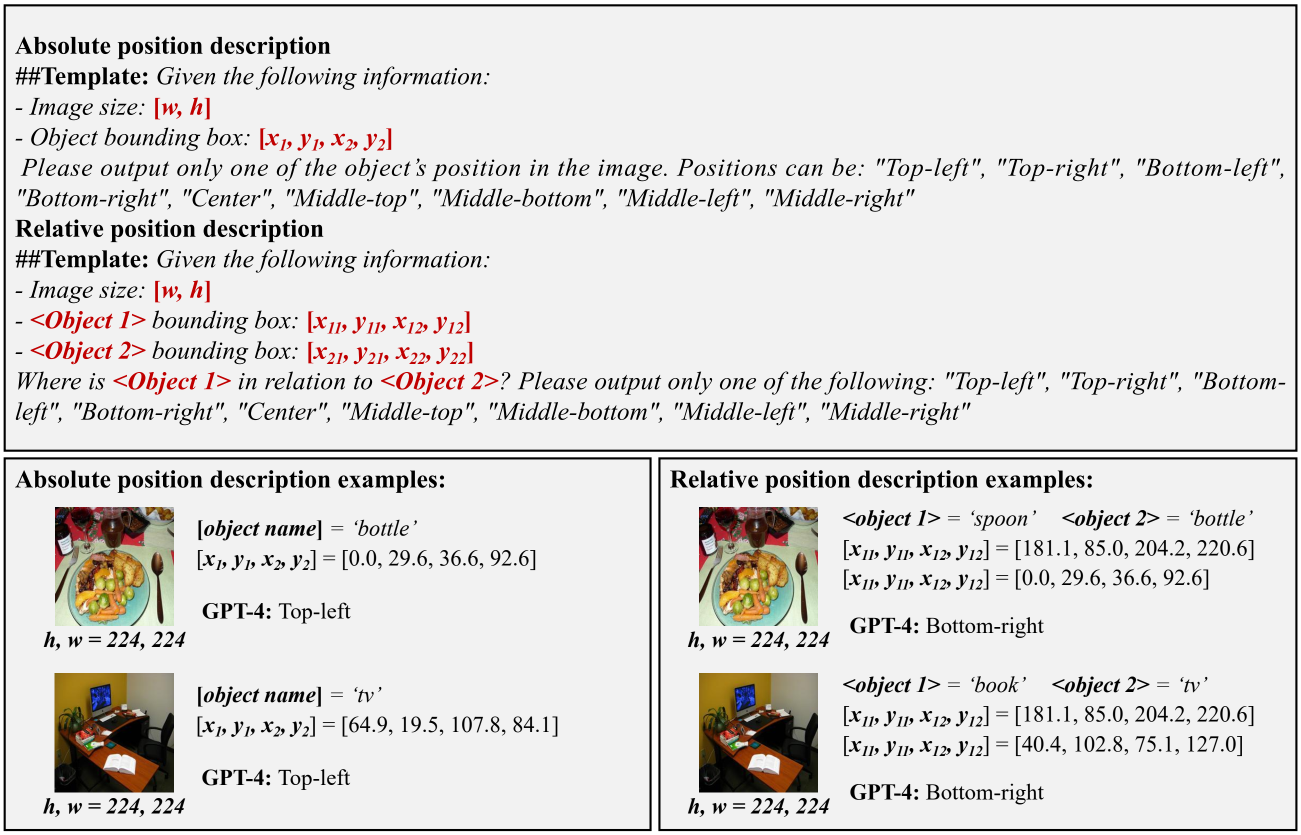
Where $\{L_i \mid i = 1, \ldots, 5\}$represents the five levels from obvious to invisible. Considering that the predicted $< level >$ token by the LMM represents a probability distribution over all possible tokens in the vocabulary, we perform a restricted softmax operation to obtain the probability P_{L_i} of each level L_i, ensuring the sum of P_{L_i} equals 1. The final predicted score of the LMM is computed as follows:

$$Q = \sum_{i=1}^{5} P_{L_i} G(L_i) = \sum_{i=1}^{5} i \cdot \frac{e^{X_i}}{\sum_{j=1}^{5} e^{X_j}}, \tag{2}$$

Here, X_i represents the logits corresponding to each level L_i, ensuring a normalized and interpretable visibility score prediction.

C. Templates of Object Location-aware Dataset

This dataset is primarily designed to enhance the model's ability to perceive the spatial positions of objects. Specifically, we approached the data design from two perspectives: absolute position descriptions and relative position descriptions. For the former, we aim to obtain the absolute positional description of an object in the image. For the latter, we focus on describing the position of one object relative to another. To construct this dataset, we utilized the original bounding box annotations from the COCO dataset. Object boundary information was used to create question templates, and with the help of GPT-4, we generated descriptive information regarding the objects' positions. The details of the question templates and relevant examples are provided in Fig. 1.



Absolute position description
##Template: *Given the following information:*
- Image size: **[w, h]**
- Object bounding box: **[x_1, y_1, x_2, y_2]**
Please output only one of the object's position in the image. Positions can be: "Top-left", "Top-right", "Bottom-left", "Bottom-right", "Center", "Middle-top", "Middle-bottom", "Middle-left", "Middle-right"
Relative position description
##Template: *Given the following information:*
- Image size: **[w, h]**
- ***<Object 1>*** *bounding box:* **[x_{11}, y_{11}, x_{12}, y_{12}]**
- ***<Object 2>*** *bounding box:* **[x_{21}, y_{21}, x_{22}, y_{22}]**
Where is ***<Object 1>*** *in relation to* ***<Object 2>****? Please output only one of the following: "Top-left", "Top-right", "Bottom-left", "Bottom-right", "Center", "Middle-top", "Middle-bottom", "Middle-left", "Middle-right"*

Absolute position description examples:

[***object name***] = *'bottle'*
[x_1, y_1, x_2, y_2] = [0.0, 29.6, 36.6, 92.6]
GPT-4: Top-left
h, w = 224, 224

[***object name***] = *'tv'*
[x_1, y_1, x_2, y_2] = [64.9, 19.5, 107.8, 84.1]
GPT-4: Top-left
h, w = 224, 224

Relative position description examples:

<object 1> = *'spoon'* ***<object 2>*** = *'bottle'*
[x_{11}, y_{11}, x_{12}, y_{12}] = [181.1, 85.0, 204.2, 220.6]
[x_{11}, y_{11}, x_{12}, y_{12}] = [0.0, 29.6, 36.6, 92.6]
GPT-4: Bottom-right
h, w = 224, 224

<object 1> = *'book'* ***<object 2>*** = *'tv'*
[x_{11}, y_{11}, x_{12}, y_{12}] = [181.1, 85.0, 204.2, 220.6]
[x_{11}, y_{11}, x_{12}, y_{12}] = [40.4, 102.8, 75.1, 127.0]
GPT-4: Bottom-right
h, w = 224, 224

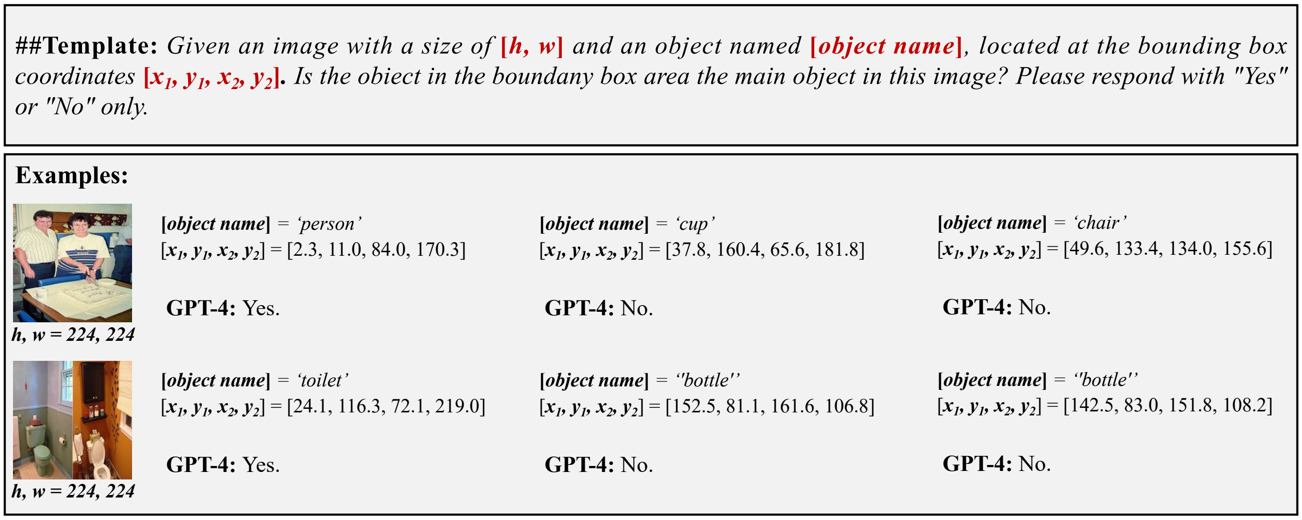
Figure 1. Text template and examples for object location-aware dataset.

D. Templates for WQA-Synthetic

We introduce a cost-effective and efficient semi-automatic annotation pipeline to construct the WQA-Synthetic. This pipeline is divided into four steps, with the second and fourth stages utilizing external LLMs, such as GPT-4 and GPT-4o, to assist in data annotation. In the following, we provide a detailed explanation of the question templates used in these two steps, along with the corresponding intermediate results.

Main Object Box Detection. In this step, we use GPT-4 to judge the main object of the original image. We design a question template by combining the object category related to the original image and the bounding box coordinate information in the image, as shown in Fig. 2.

Question-Answering Generation. After obtaining the main object bounding box coordinates in the original image, watermark bounding box coordinates, and the watermark visibility, we utilize GPT-4o to generate detailed descriptions of the watermarked images. To ensure accurate descriptions, we designed a question template that comprehensively considers both the watermark and original image bounding boxes. This template is used to input the watermark image into GPT-4o for summarization. The specific question template and relevant response examples are shown in Fig. 3. In addition, we designed 30 relevant questions, with each generated text description being randomly matched to one of these questions. An



##Template: *Given an image with a size of* **[*h, w*]** *and an object named* **[*object name*]**, *located at the bounding box coordinates* **[x_1, y_1, x_2, y_2]**. *Is the obiect in the boundany box area the main object in this image? Please respond with "Yes" or "No" only.*

Examples:

h, w = 224, 224

[*object name*] = *'person'* [x_1, y_1, x_2, y_2] = [2.3, 11.0, 84.0, 170.3] **GPT-4:** Yes.	**[*object name*]** = *'cup'* [x_1, y_1, x_2, y_2] = [37.8, 160.4, 65.6, 181.8] **GPT-4:** No.	**[*object name*]** = *'chair'* [x_1, y_1, x_2, y_2] = [49.6, 133.4, 134.0, 155.6] **GPT-4:** No.

h, w = 224, 224

[*object name*] = *'toilet'* [x_1, y_1, x_2, y_2] = [24.1, 116.3, 72.1, 219.0] **GPT-4:** Yes.	**[*object name*]** = *"bottle"* [x_1, y_1, x_2, y_2] = [152.5, 81.1, 161.6, 106.8] **GPT-4:** No.	**[*object name*]** = *"bottle"* [x_1, y_1, x_2, y_2] = [142.5, 83.0, 151.8, 108.2] **GPT-4:** No.

Figure 2. Main object box detection question template and related examples.

Table 1. Classification and scoring criteria for WQA-Real

Level	Score Range	Standard
Invisible	[0,1]	The watermark almost completely disappears, cannot be recognized in the image, and cannot be detected no matter from which angle or condition it is viewed.
Faint	[1,2]	The watermark is not obvious, only under specific conditions or careful observation can barely be seen, usually integrated with the background, the recognition is very low.
Moderate	[2,3]	The watermark is clearly visible in most cases, but not overly prominent, at a relatively unobtrusive level in the middle of the image where an observer can see the watermark without special effort.
Visible	[3,4]	The watermark is clearly visible, the contrast with the background is high, and it is easy to detect, which may slightly affect the overall visual effect of the image, but it does not attract too much attention.
Obvious	[4,5]	The watermark is very prominent, obviously affect the visual effect of the image, easy to attract the attention of observers, may block or interfere with the main content of the image.

example of such a question is: "Please describe the watermark's direction, its placement relative to objects, its texture or appearance, and its visibility."

E. Details of The WQA-Real Production

To construct textual descriptions and visibility scores for real-watermarked images, we conducted a series of subjective experiments. Ten experienced annotators were selected and rigorously trained for the task. The annotators were instructed to provide comprehensive descriptions of the watermark content in the images, including the special location, its relationship with the main objects in the image, its texture and shape characteristics, and its visibility. Finally, the annotators classified the watermark images according to a predefined visibility standard and assigned a visibility score. The specific criteria are outlined in Tab. 1.

F. Implementation Details

In our experimental setup, we randomly select 5K and 0.5K watermarked images from the WQA-Synthetic and WQA-Real dataset correspondingly to construct a diverse test set, and use the remaining images as the training set. As mentioned above, the training process is structured into three stages: In **Stage-1**, we generate 100K position question-answer pairs to train the vision encoder and visual abstractor, with a batch size of 32, a learning rate of 1×10^{-4}, and for a duration of 3 epochs. This larger learning rate is chosen because only the visual components are trained in this stage. In **Stage-2** and **Stage-3**, following the configuration in the mPLUG-owl2, we apply fine-tuning with a batch size of 16, a learning rate of 2×10^{-5},

## System: *You are a helpful watermark discriminator.*
##Template: **<watermarked image>** *The image I provided has a watermark embedded in it. The image contains the following objects with their bounding box coordinates:*
{*object 1:* $[x_{11},y_{11},x_{12},y_{12}]$, *... object i:* $[x_{i1},y_{i1},x_{i2},y_{i2}]$}
The watermark is embedded in the following areas:
$[x_{w1},y_{w1},x_{w2},y_{w2}]$ *(The visibility of watermark is* ***<level>****)*
In the input:
- [x1, y1, x2, y2] represents the bounding box coordinates for detected objects in the image.
- The watermark visibility is described with the following terms: invisible, faint, moderate, visible and obvious. The more obvious the watermark, the worse the watermark embedding effect.

First, describe the specific orientation of the watermark in the image. Next, explain the relative position of the watermark to the objects in the image, indicating whether the watermark is inside, partially overlapping, or near the edge of the objects. Avoid mentioning any bounding box or coordinate information in your response. Then, briefly describe the watermark's characteristics, such as appearance or texture. Finally, assess the watermark visibility using the format: 'The watermark visibility is: [*invisible/faint/moderate/visible/obvious*].
Present all this information in a cohesive and concise paragraph..

Examples:

##GPT-4o: *The watermark in the image is centrally oriented and extends across a significant portion of the scene. It partially overlaps with all the objects, including the bottle, wine glass, dining table, and chair, but it does not fully obscure them. The watermark has a layered and translucent appearance, with discernible text or patterns that give it a textured effect. The presence of the watermark is prominent and easily noticeable. The watermark visibility is: visible.*

##GPT-4o: *The watermark in the image is situated along one side of the image, spanning vertically. It is positioned near the edge of the bed object and partially overlaps with the teddy bear object. The watermark has a faint, blended appearance, making it integrate subtly with the background. The texture appears smooth and does not significantly disrupt the overall image quality. The watermark visibility is: faint.*

##GPT-4o: *The watermark is horizontally oriented and located near the lower center of the image. It partially overlaps with some objects, such as the individuals in the foreground, but does not interfere with peripheral elements like kites in the sky. The watermark features a text-based design with a translucent appearance, blending into the background but still discernible. The watermark visibility is: moderate.*

Figure 3. Text template and examples for Question-answering generation.

and a duration of 5 epochs. Across all three stages, the Adam optimizer and cosine annealing learning rate schedule are used to dynamically adjust the learning rate. The reported performance of WMarkGPT is evaluated using the final weights obtained after fine-tuning and all experiments were conducted on 8 NVIDIA RTX 6000 Ada GPUs.

G. Evaluation Metrics

To comprehensively evaluate the watermark understanding capabilities of WMarkGPT and other MLLMs. Traditional metrics like PSNR are not included in this experimental section because they fail to provide textual descriptions and visibility, which are the core of evaluation. Instead, we refer the reader to Fig. 3 for a comparison of WMarkGPT with these metrics in other aspects. we compare their performance in two major aspects: (1) the relevance of their watermark descriptions, and (2) the accuracy of their visibility predictions.

(1) Relevance of Watermark Description. To evaluate the quality of generated watermark descriptions, i.e., their relevance to the reference text, we employ established natural language processing metrics. Specifically, we calculate BLEU-1 and ROUGE-L scores to measure word-level similarity between the generated description and the reference text. We also employ the LLM-Score, as described in (Lu et al., 2024), to assess semantic relevance. This score ranges from 0 to 4, with higher scores indicating better relevance. For ease of interpretation, the final scores are normalized to a scale of 0 to

Table 2. Classification and scoring criteria for WQ-Real

Models	WQA-Synthetic			WQA-Real		
	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
DBCNN	0.91	0.91	0.75	0.63	0.70	0.46
HyperIQA	0.91	0.91	0.75	0.62	0.65	0.44
TReS	0.90	0.90	0.75	0.46	0.48	0.33
CLIP-IQA	0.89	0.86	0.71	0.43	0.46	0.31
WMarkGPT	**0.93**	**0.93**	**0.78**	**0.71**	**0.74**	**0.54**

100. Specifically, we designed a text template to evaluate the correlation between the generated content and the reference description with the help of GPT-4. The detailed text template is as follows:

"Evaluate the relevance between the following description and ground truth on a scale from 0 to 4. Higher scores indicate better relevance. Return only the numeric score. - Description: $< candidates >$ *- Ground Truth:* $< references >$*"*

(2) Accuracy of Visibility Prediction. A crucial part of evaluating the ability to understand watermarks from a multimodal large language model is assessing the accuracy of its visibility predictions. In our dataset, each watermarked image has a corresponding visibility label, ranging from "invisible" to "obvious" (i.e., "invisible", "faint", "moderate", "visible", "obvious"). Unlike WMarkGPT, which directly outputs visibility predictions, other MLLM baselines describe visibility in various textual formats, often embedding the information within their generated text. To standardize the visibility classification, we use a question template and leverage GPT-4 to categorize the visibility level in the generated text from other MLLM baselines. Finally, we calculate the accuracy (ACC) of these visibility predictions in all images. The question template designed to summarize the visibility of the watermark is as follows:

"You are an expert in image watermark analysis, specializing in assessing the visibility of watermarks. I will provide a textual description of an image watermark, and your task is to evaluate its visibility based on the description. Choose the most appropriate visibility level from the following options: invisible, faint, moderate, visible, or obvious. Provide only the selected visibility level as your output. -Input: $< candidates >$*"*

H. Comparison with Other Score-based Methods

Our model calculates the watermark visibility score by computing the probability of the final level word, providing a more accurate representation of watermark visibility and the degree of image content degradation compared to traditional pixel-wise metrics. To further highlight the advantages of our model, we selected several score-based deep neural network models, including DBCNN (Zhang et al., 2018b), HyperIQA (Su et al., 2020), TReS (Golestaneh et al., 2022), and CLIP-IQA (Wang et al., 2023), and trained and tested them on score prediction tasks using the WQA-Synthetic and WQA-Real datasets according to the configurations from their respective original papers. We then compared the SRCC, PLCC, and KRCC metrics of our model against these score-based methods. The experimental results, shown in Table 2, demonstrate that our method outperforms these score-based approaches in predicting watermark visibility scores, achieving results that are more consistent with subjective ratings. This superiority may be attributed to the fact that most score-based methods are designed to assess image distortion, which does not capture the watermark's semantic content effectively, leading to poorer performance in watermark visibility score prediction tasks.