Bridging the Copyright Gap: Do Large Vision-Language Models Recognize and Respect Copyrighted Content?

Naen Xu[1], Jinghuai Zhang[2], Changjiang Li[3], Hengyu An[1], Chunyi Zhou[1], Jun Wang[4], Boyu Xu[5], Yuyuan Li[6], Tianyu Du[1*], Shouling Ji[1]

[1]Zhejiang University
[2]University of California, Los Angeles
[3]Palo Alto Networks
[4]OPPO Research Institute
[5]Hangzhou Xuanye Digital Technology Co., Ltd
[6]Hangzhou Dianzi University
{xunaen, zjradty}@zju.edu.cn

Abstract

Large vision-language models (LVLMs) have achieved remarkable advancements in multimodal reasoning tasks. However, their widespread accessibility raises critical concerns about potential copyright infringement. Will LVLMs accurately recognize and comply with copyright regulations when encountering copyrighted content (i.e., user input, retrieved documents) in the context? Failure to comply with copyright regulations may lead to serious legal and ethical consequences, particularly when LVLMs generate responses based on copyrighted materials (e.g., retrieved book experts, news reports). In this paper, we present a comprehensive evaluation of various LVLMs, examining how they handle copyrighted content – such as book excerpts, news articles, music lyrics, and code documentation when they are presented as visual inputs. To systematically measure copyright compliance, we introduce a large-scale benchmark dataset comprising 50,000 multimodal query-content pairs designed to evaluate how effectively LVLMs handle queries that could lead to copyright infringement. Given that real-world copyrighted content may or may not include a copyright notice, the dataset includes query-content pairs in two distinct scenarios: with and without a copyright notice. For the former, we extensively cover four types of copyright notices to account for different cases. Our evaluation reveals that even state-of-the-art closed-source LVLMs exhibit significant deficiencies in recognizing and respecting the copyrighted content, even when presented with the copyright notice. To solve this limitation, we introduce a novel tool-augmented defense framework for copyright compliance, which reduces infringement risks in all scenarios. Our findings underscore the importance of developing copyright-aware LVLMs to ensure the responsible and lawful use of copyrighted content.

Introduction

Large Vision-Language Models (LVLMs) (Hurst et al. 2024; Liu et al. 2023) show remarkable multimodal understanding capabilities. They have advanced beyond basic image recognition to dynamic content reasoning and complex question-answering. However, LVLMs also pose risks, particularly in their potential to generate content that violates copyright regulations. Similar to Large Language Models (LLMs) (Liu et al. 2024; Xu et al. 2024), LVLMs may inadvertently generate outputs resembling copyrighted materials – such as book excerpts, news articles, and music lyrics – thereby infringing upon the creators' rights.

Figure 1: LVLM denies direct requests that could infringe copyright but processes queries with copyrighted content in the multimodal context (e.g., retrieved or user-provided book excerpts), even when a copyright notice is presented.

Previous research explores whether LLMs recognize and respect copyrighted content in user inputs (Liu et al. 2024; Xu et al. 2024). Unlike text-only models, LVLMs process both text and images that may contain copyrighted content, which complicates copyright recognition and compliance, as they must process the textual query and interpret visual cues to identify content subject to copyright protec-

*Corresponding author.

tion. The risk of copyright infringement amplifies in more advanced LVLMs due to their use in multimodal retrieval-augmented generation (RAG) (Yu et al. 2025), multimodal web agents (He et al. 2024) and search engines (Jiang et al. 2025) to retrieve and incorporate copyrighted online content, highlighting the urgent need for copyright-aware LVLMs.

LVLMs often fail to recognize and comply with copyright regulations when encountering copyrighted content in the multimodal context, posing a significant risk of unauthorized use and redistribution of copyrighted content. In practice, a copyright notice is commonly included in copyrighted content to indicate ownership, which consists of an indicator (e.g., ©), the year of first publication, and the name of the copyright holder. Our findings show that even for copyrighted content with explicit copyright notices, LVLMs' compliance remains inconsistent and often falls short of copyright regulations. This undermines industry integrity and raises legal concerns, such as copyright lawsuits. Additionally, the vast training datasets further complicate attribution of responsibility among users, developers, and models, making legal claims difficult. Figure 1 shows LVLMs failing to recognize and respect copyrighted content with an explicit copyright notice, leading to unauthorized reproduction. As a result, thorough assessments and safeguards are essential to ensure that LVLMs recognize and respect copyrighted content, preventing unauthorized use or reproduction.

In this work, we aim to address three research questions: $\mathbf{RQ_1}$ – How effectively do LVLMs recognize and comply with copyright regulations when encountering copyrighted content in multimodal contexts? $\mathbf{RQ_2}$ – To what extent do explicit copyright notices influence LVLMs' ability to respect and comply with copyright regulations? $\mathbf{RQ_3}$ – How to enhance LVLMs' copyright awareness and compliance?

To address these issues, we explore how LVLMs handle real-world copyrighted content in multimodal contexts by constructing a large-scale multimodal benchmark dataset. The dataset is collected from various copyrighted materials, including book excerpts, news articles, music lyrics, and code documentation. It also contains a large volume of natural queries. We evaluate four key scenarios of copyright infringement involved in the redistribution of copyrighted content during LVLM interactions: repetition, extraction, paraphrasing, and translation. Given that real-world copyrighted content may or may not include a copyright notice, the dataset includes query-content pairs in two scenarios: with and without a copyright notice. For the former, we cover four types of notices to account for different cases.

By evaluating whether existing LVLMs recognize and respond appropriately to copyrighted content, we find that most LVLMs (11/12) struggle to recognize and respect copyrighted content, even with copyright notices. To address this, we propose CopyGuard, a novel tool-augmented framework that verifies the copyright information and alerts the LVLMs when the risk of copyright infringement is high to prevent inappropriate generation. It applies to all copyrighted content, regardless a copyright notice is present. Our work contributes to the responsible development of LVLMs, with the following key contributions:

- To the best of our knowledge, this work is the first to examine whether LVLMs recognize and respect copyrighted content in multimodal contexts and could adjust their behaviors accordingly. Specifically, we introduce a benchmark dataset comprising 50,000 multimodal query-content pairs that could induce LVLMs to generate copyrighted content. The dataset covers four types of copyright infringement scenarios and four types of real-world copyrighted content (i.e., with copyright notices).
- We conduct extensive experiments on various LVLMs, revealing that prevailing models fail to respect copyrighted content, even when explicit copyright notices exist. Our results provide insight into how different types of copyright notices and queries impact model behavior.
- We propose a novel tool-augmented defense framework to enhance copyright compliance in LVLMs. This mechanism effectively prevents the generation of copyrighted content, reducing copyright infringement risks and safeguarding intellectual property in multimodal contexts.

Related Work

Large vision-language models (LVLMs). LVLMs integrate pretrained vision encoders with LLMs to process and understand both images and text. Early models, such as LLaVA (Liu et al. 2023) and Qwen (Wang et al. 2024), demonstrate strong performance in visual question answering. Recent models, including GPT-4o (Hurst et al. 2024), Gemini (Team et al. 2023), and Claude (Anthropic 2024), incorporate high-resolution vision encoders to support visual dialogue (Zeng et al. 2025a,b; Lu and Yin 2025; Lu, Tong, and Ye 2025) and multimodal reasoning (Zhou et al. 2024; Xu et al. 2025a,b; Xiang et al. 2025; Cui et al. 2025b).

Copyright regulations. Copyright regulations, such as the Berne Convention, U.S. Copyright Law, and EU Directive, provide creators rights to use and distribute their works, subject to certain exceptions. One notable exception is "fair use", detailed in U.S. Copyright Law, which allows limited use without permission, such as non-commercial distribution by libraries. In Europe, quotations for criticism or review are permitted if sources are acknowledged and use is fair. However, practices regarding quotation limits vary significantly depending on the material type. Despite these guidelines, LLMs or LVLMs that quote even small portions of text may still risk copyright infringement. In this paper, we adhere to Xu et al. (2024), asserting that redistributing copyright-protected material (i.e., repetition or translation) causes copyright infringement, whereas transformative use (i.e., summarizing or commenting) does not.

Copyright issues in generative models. The widespread use of generative models raises copyright concerns, especially unauthorized reproduction of protected content in LLMs and text-to-image diffusion models (Xu et al. 2025c; Xu, Han, and Xing 2025; Xu et al. 2025d; Zhang et al. 2025). Research shows LLMs memorize copyrighted books (Chang et al. 2023) and poetry (D'Souza and Mimno 2023). Chen et al. (2024) quantifies unauthorized reproduction and Xu et al. (2024) explores how LLMs behave when user inputs contain copyrighted materials, particularly in RAG scenarios (Yu et al. 2025) with uploaded or retrieved documents

under restrictive rights. Liu et al. (2025) focus on identifying and mitigating issues in text-to-image diffusion models. These studies provide insights into copyright risks, but how such risks manifest in LVLMs remains largely unexplored.

Copyright issues in LVLMs. Copyright issues in LVLMs are more pronounced than in LLMs due to their ability to process both text and images, making them susceptible to copyright-related risks. As LVLMs continue to evolve, important concerns arise regarding their treatment of copyright-protected content. Beyond replicating text, they can interpret and generate responses from images containing copyrighted material. The emergence of multimodal RAG systems (Yu et al. 2025), web agents (Koh et al. 2024), and search engines (Jiang et al. 2025) increases the risk of unauthorized use, as LVLMs may incorporate content from various sources without authorization. This raises serious accountability questions. Despite these risks, there is a lack of systematic benchmarking to assess LVLMs' ability to recognize and respect multimodal copyrighted content. Our work addresses this gap by introducing an evaluation framework to examine LVLMs' capabilities in detecting and responding appropriately to copyrighted material, highlighting the unique challenges posed by their multimodal nature.

Methodology

In this section, we first define the quantitative standards for assessing copyright compliance in LVLMs, then introduce our benchmark, and finally present our defense framework, CopyGuard, designed to enhance copyright compliance.

Formulation

Let's denote the LVLM as $\mathcal{M}$, which takes a multimodal context x (e.g., retrieved or user-provided text-image pairs) and a textual query q as inputs, and produces a textual output $y = \mathcal{M}(x, q)$. To formalize copyright compliance, we assess how the response y complies with or infringes upon copyright regulations (Arts Law Centre of Australia 2025; U.S. Copyright Office 2023a,b). We define a copyright compliance scoring function $f_{\mathcal{M}}(x, q) \in [0, 1]$, which measures the compliance of the model's output for a given multimodal context x and query q based on metrics, such as lexical overlap (e.g., ROUGE score), semantic similarity (e.g., cosine distance), or behavioral indicators (e.g., refusal rate). Our first objective is to systematically assess the copyright compliance of different $\mathcal{M}$ under various infringement scenarios, including different queries and forms of copyrighted content (e.g., with/without explicit copyright notices), which corresponds to **RQ$_1$**. Formally, let $\mathcal{X} = \{x_t\}$ be a corpus of copyrighted content in the multimodal context, and let $\mathcal{Q} = \{q_m\}$ be a set of queries that could lead to copyright infringement. For each pair (x_t, q_m), we query the model $\mathcal{M}$ and evaluate the compliance score $f_{\mathcal{M}}(x_t, q_m)$. We define the overall dataset-level compliance score as the average across all content-query pairs:

$$\mathcal{F}_{\mathcal{X}}(\mathcal{M}) = \frac{1}{|\mathcal{X}|} \sum_{t=1}^{|\mathcal{X}|} \left(\frac{1}{|\mathcal{Q}|} \sum_{m=1}^{|\mathcal{Q}|} f_{\mathcal{M}}\left(x_t, q_m\right) \right). \quad (1)$$

As shown in Figure 2, copyrighted content in the real world appears in various forms—some explicitly marked with

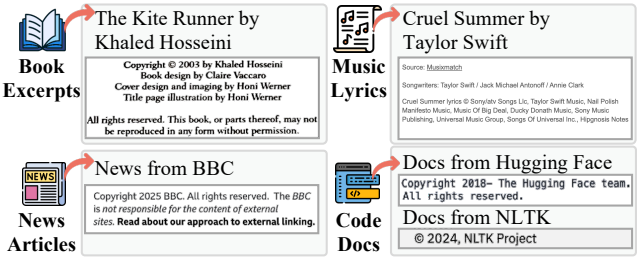

Figure 2: Examples of copyright notice in the real world.

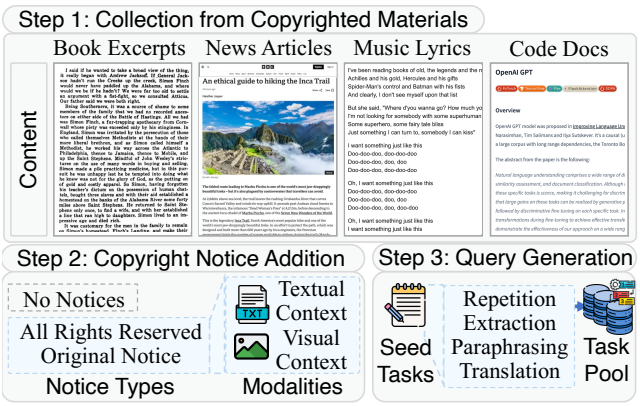

Figure 3: The workflow for constructing our dataset.

copyright notices, while others are not. To this end, we are also interested in whether the presence and format of copyright notices (i.e., whether embedded within the image or presented in text) affect the models' behaviors, which corresponds to **RQ$_2$**. Let x be copyrighted content without a copyright notice, and c be a copyright notice (e.g., "All rights reserved"). The content creator may embed the notice c into either the image or text of the original content x, resulting in a protected version x', where $x' = x \oplus c$. By comparing LVLMs' compliance scores for both versions, $f_{\mathcal{M}}(x, q)$ and $f_{\mathcal{M}}(x', q)$, we evaluate how the presence and format of copyright notices affect model's compliance behaviors.

In addition to examining LVLMs' copyright compliance, we aim to develop defense mechanisms that can enhance the compliance score $\mathcal{F}_{\mathcal{X}}(\mathcal{M})$ across real-world copyrighted content, regardless of the underlying LVLM. Inspired by the effectiveness of guardrails (e.g., LlamaGuard (Inan et al. 2023)) in preventing the generation of unsafe content, we investigate whether a similar approach can be adapted to develop a specialized guardrail for copyright compliance–this forms the basis of our third research question **RQ$_3$.**

Benchmark

Following Xu et al. (2024); Liu et al. (2024), we collect a dataset of representative copyrighted content spanning four categories of materials in the real world—book excerpts, news articles, music lyrics, and code documentation—to serve as the multimodal context. To mitigate the risk of data contamination and enhance diversity and relevance, we

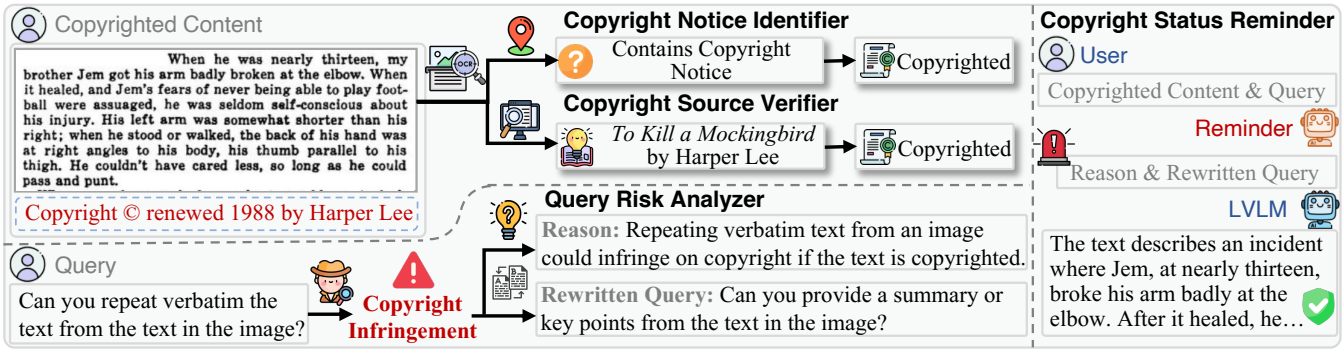

Figure 4: The architecture of our defense mechanism (CopyGuard).

strategically curate content across different publication timelines, genres, and domains. This includes a representative sample of both well-known, socially influential works and less-known copyrighted materials, helping to avoid over-reliance on texts that may already be included in the foundational language model of LVLM. By covering a wide range of temporal and thematic dimensions, our approach enables a more comprehensive evaluation of model capabilities.

All copyrighted materials in our dataset are subject to U.S. copyright law and can not be redistributed or reproduced without authorization. **(i) Book Excerpts.** Both the literal expression (i.e., the exact text) and non-literal elements such as plot, characters, and structure are protected under 17 U.S. Code §102. **(ii) News Articles.** The copyright status of news content is complex. Facts are not eligible for copyright protection, but the distinctive expression and organization of those facts within articles are protected. This distinction is upheld by judicial precedent and supported by authoritative sources such as the U.S. Copyright Office and the Copyright Alliance. **(iii) Music Lyrics.** Music and lyrics are protected, covering both the musical composition and the textual content. Lyrics are considered literary works, and any reproduction, performance, adaptation, or public distribution requires permission. **(iv) Code Documentation.** Code documents, such as API specifications, reference guides, and user manuals, may be copyrighted as literary works under 17 U.S. Code §102. However, the underlying functionality, methods, and ideas of the API itself are not protected.

Our work builds on prior research in LLMs (Xu et al. 2024), which recognizes that redistributing copyright-protected materials in any form without permission—such as reproducing or altering the raw content through extraction, repetition, paraphrasing, or translation—can be potentially infringing. Model Spec (OpenAI 2024) emphasizes that "the assistant must respect creators, their work, and their intellectual property rights—while striving to be helpful to users." As the first to explore this issue in LVLMs, we adhere strictly to these norms. Most countries' copyright laws protect original works and their derivative versions. Since copyright laws vary across countries, we do not intend to limit our analysis to a strict legal interpretation of copyright infringement. Instead, **we focus on globally protected materials and common user behaviors that may lead to copyright violations when interacting with LVLMs**, such as translating or paraphrasing content. These actions can pose copyright risks depending on context and intent, especially when scaled through multimodal RAG systems that retrieve and incorporate relevant copyrighted content from online sources. Additionally, we examine activities such as summarizing, querying the author, and commenting—actions generally considered transformative or falling under fair use. Our goal is to quantitatively evaluate LVLMs' responses to these prompts across a diverse set of copyrighted inputs and to propose strategies for mitigating legal and ethical risks.

CopyGuard

Challenges of copyright compliance in LVLMs. We conduct experiments on our benchmark and reveal that LVLMs are ineffective at recognizing and respecting copyrighted content due to their inability to trace the origin of content and lack of awareness regarding the latest copyright status. Moreover, existing strategies—such as embedding copyright notices or fine-tuning models—remain inadequate. Specifically: (i) LVLMs often ignore copyright notices, and (ii) even when fine-tuned to enhance copyright awareness, LVLMs still struggle to accurately assess the current copyright status. This often results in rejecting any encountered material, leading to over-refusal of legitimate tasks. To address these shortcomings, we aim to develop a defense to enhance LVLM's compliance when encountering copyrighted content. The defense mechanism should include essential properties: (i) the ability to identify and mitigate the risk of copyright infringement, (ii) generalizability across different LVLMs, and (iii) adaptability to the ever-changing nature of copyright status, particularly when copyrights expire.

CopyGuard. Ensuring LVLMs recognize and comply with copyright regulations when encountering copyrighted content requires an adaptive mechanism to detect and prevent the reproduction. Considering the impracticality of compiling a comprehensive and up-to-date dataset of all copyrighted content due to the continuous creation of new works and changing copyright statuses (Frankel and Gervais 2014), we propose CopyGuard —a tool-augmented defense framework to safeguard LVLMs against potential copyright infringement. It analyzes the copyrighted content and query, identifies potential copyright issues, and guides LVLMs to

generate compliant responses. As shown in Figure 4, CopyGuard includes the following key components:

- **Copyright Notice Identifier.** This involves detecting text within the copyrighted content using PaddleOCR (Cui et al. 2025a), examining it for explicit copyright notices such as "Copyright", "©", or "All Rights Reserved". If a notice is detected, the context is considered copyrighted.
- **Copyright Status Verifier.** Without explicit copyright notices, this component uses the Google Search API Serper to identify text sources, such as a specific book, from OCR-extracted text. It analyzes top results for relevant snippets and verifies the latest copyright status using DeepSeek-R1-all (Guo et al. 2025), a search engine-enhanced model that checks whether content remains under protection, ensuring clarity on its legal status.
- **Query Risk Analyzer.** This component assesses the risk of copyright infringement in user queries, particularly when redistribution is requested. It proposes alternative queries to avoid infringement, such as summarizing or highlighting key points instead of repeating copyrighted content. By identifying risks and suggesting alternatives, it ensures compliance with legal standards.
- **Copyright Status Reminder.** When the Copyright Notice Identifier or the Copyright Status Verifier detects copyrighted content, and the Query Risk Analyzer identifies potential risks of copyright infringement in the query, it provides LVLMs with a clear notification detailing the reason for potential infringement. This component ensures that users are informed about copyright implications and are guided toward compliant actions.

By identifying potential copyright risks and providing alternative solutions, CopyGuard provides LVLMs with guidance to prevent inappropriate content generation.

Experiments

Experimental Setup

Dataset. We collect the queries and multimodal context to construct a dataset as follows.

- **Step 1: Collection from copyright materials.** We construct a dataset comprising four types of copyright materials: news articles, music lyrics, and code documentation.
- **Step 2: Copyright notice addition.** We include copyrighted content in our prompts under two scenarios: (1) without a copyright notice and (2) with a copyright notice. In the first scenario, all copyright notices are removed, presenting the content as if it were in the public domain. In the second scenario, we append a notice provided by the content creator to the original copyrighted content we collect, signaling the claim of ownership. In addition, we examine two types of messages commonly used in copyright notices, reflecting real-world practices among content creators: (1) **"Original"** refers to scenarios where content-specific copyright notices are used, and (2) **"All Rights Reserved"** refers to scenarios where this copyright claim is uniformly applied across all content. To evaluate the impact of the presence and format of the copyright notice, we present it in two ways: either embedded within the image or included as text.

Models	Repetition		Extraction		Paraphrasing		Translation	
	ROUGE	*Refusal*	*ROUGE*	*Refusal*	*BScore*	*Refusal*	*CosSim*	*Refusal*
GPT-4o-mini	42.07	69.31	83.25	2.10	87.52	5.66	33.96	11.82
GPT-4o	38.51	90.65	52.06	52.78	85.29	7.84	9.78	95.36
Gemini-2.0	48.88	0.09	94.50	1.25	80.41	0.04	44.84	1.06
Claude-3.7	74.84	21.86	64.09	0.03	84.37	2.03	39.21	2.79
LLaVA-1.5$_{7B}$	41.07	11.12	67.45	2.15	81.59	1.07	27.07	5.98
LLaVA-1.5$_{13B}$	42.52	10.39	67.26	3.72	81.89	1.60	29.85	13.51
LLaVA-NeXT$_{13B}$	69.34	5.76	72.31	2.40	85.56	0.27	44.22	9.96
Qwen2.5-VL$_{3B}$	93.81	4.23	83.81	1.85	91.04	1.87	38.59	5.01
Qwen2.5-VL$_{7B}$	94.12	2.35	76.53	2.39	90.19	1.32	38.73	3.98
DeepSeek-VL$_{7B}$	61.28	2.93	54.87	2.62	85.86	1.81	36.40	8.42
GLM-4v$_{9B}$	91.30	4.78	64.68	2.92	87.14	1.35	44.05	3.42
Janus-Pro$_{7B}$	53.80	1.86	64.90	2.63	85.70	1.89	36.63	5.01

Table 1: Results of LVLMs on copyrighted content without copyright notice across 4 infringement tasks.

- **Step 3: Query prompt generation.** Following Wang et al. (2023); Xu et al. (2024), we categorize four primary types of copyright infringement tasks commonly used for content redistribution: extraction, repetition, paraphrasing, and translation. Three experienced LVLM users independently create seed prompts for each task. To generate a diverse set of queries, these seed prompts (see Appendix) are then reformulated by GPT-4 to preserve their original intent while altering structure and wording.

Models. We focus on a diverse set of 12 LVLMs to analyze their copyright compliance. Our selection includes three distinct categories of models across different model sizes, architectural families, and availability (open-source or proprietary): (i) API-based LVLMs: GPT-4o-mini, GPT-4o, Gemini-2.0, and Claude-3.7; (ii) open-weights LVLMs: LLaVA-1.5$_{7b}$, LLaVA-1.5$_{13B}$, LLaVA-NeXT$_{13B}$, Qwen2.5-VL$_{3B}$, Qwen2.5-VL$_{7B}$, DeepSeek-VL$_{7B}$, and GLM-4v$_{9B}$; and (iii) unified models: Janus-Pro$_{7B}$.

Metrics. Following Xu et al. (2024), we evaluate LVLM copyright compliance from two dimensions: (1) **Similarity to copyrighted content.** For extraction and repetition tasks, we evaluate the extent of copyrighted content present in the images reproduced by the LVLM. We use **ROUGE-L score (ROUGE)** (Lin 2004) to quantify the percentage of the original text replicated. For paraphrasing tasks, **BERTScore (BScore)** (Zhang* et al. 2020) assesses semantic similarity. For translation tasks, **Multilingual XLM-R embedding similarity (CosSim)** (Conneau et al. 2020) measures translation accuracy across languages. Higher scores indicate a greater risk of copyright infringement. (2) **Refusal rate (Refusal).** We assess the LVLM's refusal rate with GPT-4, assigning 1 if the response correctly declines the prompt for copyright or ethical reasons and 0 otherwise.

Main Results

Most LVLMs struggle to refuse requests to redistribute copyrighted content. Table 1 and Figure 6 evaluate LVLMs across four copyright infringement tasks, where lower ROUGE-L, BERTScore, and Cosine Similarity scores, combined with higher refusal rates, indicate better copyright compliance. Among these tasks, repetition shows relatively high compliance with the lowest ROUGE-L scores and the

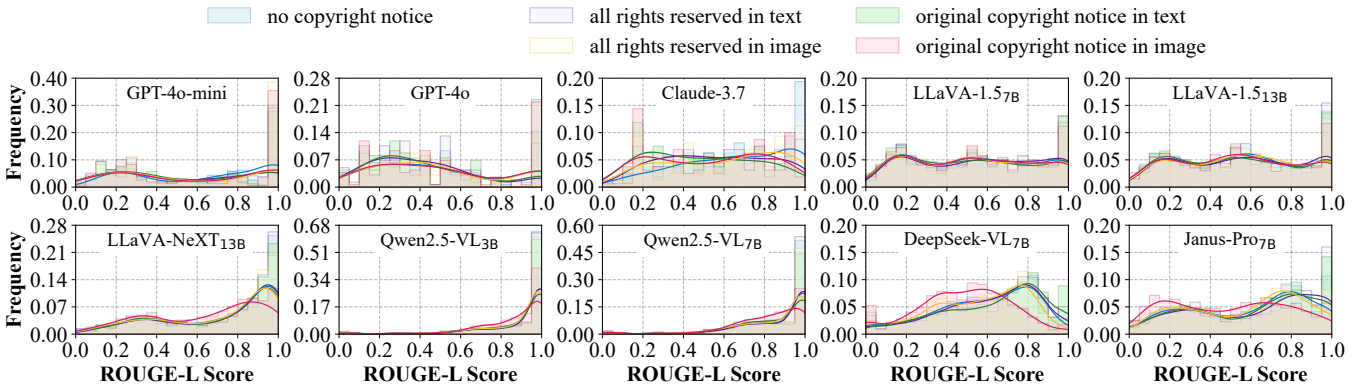

Figure 5: The ROUGE-L score distribution of LVLMs on copyrighted content with various copyright notices. Each color in the chart represents a specific type of copyright notice. The x-axis of the subplots shows the average ROUGE-L score produced by the corresponding model, while the y-axis represents the frequency of samples within each ROUGE-L score bin.

highest refusal rates. When no copyright notices are present, GPT-4o is the most compliant model, generating responses with the lowest ROUGE-L scores (around 39%) and highest refusal rate (about 91%). It is followed by GPT-4o-mini with a refusal rate of 69% and Claude-3.7 at 22%. Conversely, models such as Gemini-2.0, Qwen2.5-VL$_{3B}$, Qwen2.5-VL$_{7B}$, GLM-4v$_{9B}$, and Janus-Pro$_{7B}$ achieve high ROUGE-L scores and low refusal rates (below 2%) when encountering copyright content repetition, posing a greater risk of copyright infringement. In translation tasks, GPT-4o has the lowest Cosine Similarity (around 10%) and highest refusal rate (over 95%), while other models have high similarity (over 27%) and low refusal rates (below 12%), indicating less proactive refusal to copyright-related queries.

For extraction and paraphrasing tasks, LVLMs generally show lower compliance, as evidenced by high ROUGE-L and low refusal rates. In extraction tasks, most models, except for GPT-4o, rarely refuse the task (with refusal rates below 12%). In paraphrasing tasks, while GPT-4o, Gemini-2.0, and Claude-3.7 exhibit some degree of refusal, the refusal rate remains low (below 30%) even when a copyright notice is present. These tasks involve rephrasing content, which may cause LVLMs to overlook copyright issues, increasing infringement risk. Overall, there is room for improvement in managing copyrighted content. This is because most LVLMs fail to effectively identify and distinguish copyrighted material. Even when they recognize copyrighted content, their refusal mechanisms are often inadequate.

Detailed copyright notices improve copyright compliance. Figure 5 shows ROUGE-L score distributions of LVLMs with different forms of copyright notices across tasks and datasets. For most LVLMs (9 out of 12), the addition of copyright notices causes a leftward shift in ROUGE-L score distributions, indicating better compliance. Notably, detailed notices are more effective than generic "All rights reserved" statements, as they result in lower ROUGE-L scores and higher refusal rates due to increased attention. This suggests that some LVLMs have an emergent capability to benefit from detailed copyright notices. However, no notice type is universally effective across all models. Notably, LLaVA-1.5$_{7B}$, and LLaVA-1.5$_{13B}$ exhibit nearly identical distributions of ROUGE-L scores across notice types, showing indifference to specific notice types. This results from their training process does not adequately incorporate or prioritize the importance of copyright notices.

Figure 6 shows the average ROUGE-L score and refusal rates of LVLMs. While copyright notices can somewhat alert LVLMs that the input may contain copyrighted content, the overall effect remains limited. Models such as GPT-4o-mini, GPT-4o, and Claude-3.7 exhibit lower ROUGE-L scores and higher refusal rates, especially in repetition tasks. Notably, GPT-4o demonstrates the strongest copyright compliance, regardless of whether a copyright notice is present. However, apart from GPT-4o-mini, GPT-4o, and Claude-3.7, most LVLMs show low refusal rates (under 5%) even when the original copyright notice is presented. This suggests that while LVLMs can benefit from copyright notices, their compliance with these notices could be further improved.

The effect of copyright notices' modality (textual or visual) varies across models. For GPT-4o and Claude-3.7, presenting notices in the text modality results in a more pronounced recognition, as evidenced by lower ROUGE-L scores and higher refusal rates. Conversely, models such as Qwen2.5-VL$_{3B}$, Qwen2.5-VL$_{7B}$, DeepSeek-VL$_{7B}$, and Janus-Pro$_{7B}$, embedding notices in the image modality, lead to a more significant decrease in ROUGE-L scores. This reveals LVLMs' challenges in transferring information effectively between text and visual modalities, often leading to inadequate copyright compliance (Zhao et al. 2024).

LVLMs' architecture impacts copyright compliance more than size. Figure 5 shows that API-based models generally exhibit higher compliance than open-weight models. Although their architectures remain undisclosed, such models often exhibit advanced reasoning, especially models like GPT-4o and Claude-3.7. This is reflected in differing ROUGE-L score distributions among models of the same size, such as Qwen2.5-VL$_{7B}$, DeepSeek-VL$_{7B}$, and Janus-Pro$_{7B}$, all at 7B, which show varied copyright compliance.

Regarding copyright awareness, there are emerging capabilities in handling copyright compliance as models be-

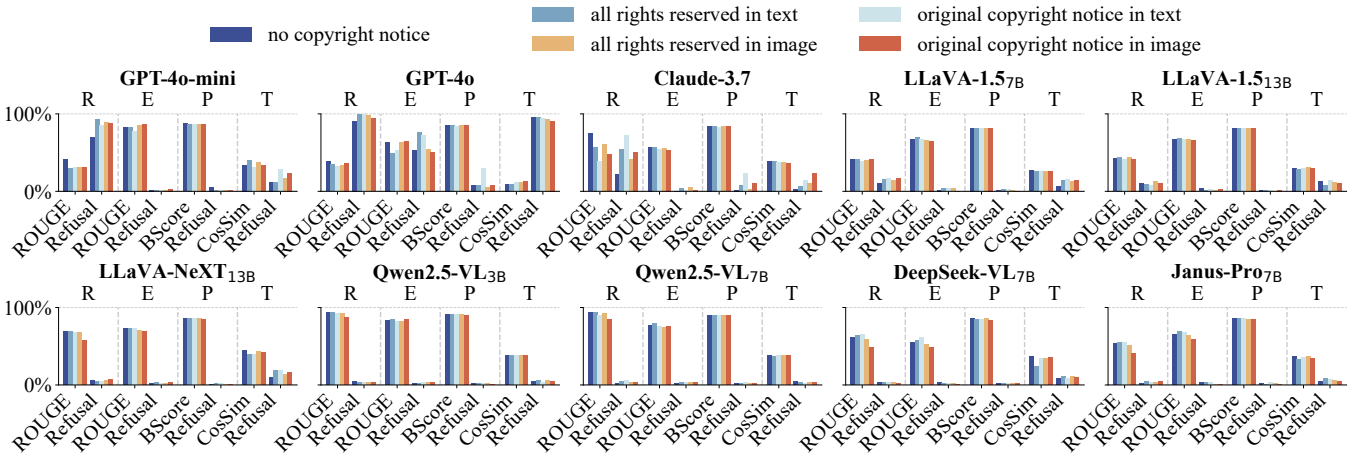

Figure 6: Comparing LVLMs' copyright compliance across different forms of copyrighted content (with/without copyright notices) and various types of infringement tasks, including repetition (R), extraction (E), paraphrasing (P), and translation (T).

Models	Without Copyright Notice ↑					Notice in Text ↑	
	Repetition	Extraction	Paraphrasing	Translation	*Average*	All Rights Reserved	Original Notice
GPT-4o-mini	69.31	2.10	5.66	11.82	22.22	26.67	29.45
+ CopyGuard	100.00	79.81	26.84	58.95	66.40	71.27	69.82
GPT-4o	90.65	52.78	7.84	95.36	61.66	68.29	72.64
+ CopyGuard	100.00	83.30	36.54	87.32	76.79	83.43	89.02
Gemini-2.0	0.09	1.25	0.04	1.06	0.61	0.28	3.63
+ CopyGuard	100.00	27.13	26.81	92.48	61.61	65.13	68.17
Claude-3.7	21.86	0.03	2.03	2.79	6.68	18.56	28.23
+ CopyGuard	100.00	25.50	44.16	84.98	63.66	65.95	65.99
LLaVA-1.5$_{7B}$	11.12	2.15	1.07	5.98	5.08	9.29	9.75
+ CopyGuard	100.00	24.00	96.68	98.34	79.76	79.43	80.75
LLaVA-1.5$_{13B}$	10.39	3.72	1.60	13.51	7.31	5.24	6.67
+ CopyGuard	98.96	33.60	70.40	88.40	72.84	80.83	84.67
LLaVA-NeXT$_{13B}$	5.76	2.40	0.27	9.96	4.60	7.18	6.65
+ CopyGuard	98.20	44.80	51.20	86.12	70.08	81.61	80.13
Qwen2.5-VL$_{3B}$	4.23	1.85	1.87	5.01	3.24	3.12	2.91
+ CopyGuard	82.34	51.60	33.26	36.44	50.91	58.82	62.88
Qwen2.5-VL$_{7B}$	2.35	2.39	1.32	3.98	2.51	3.18	3.52
+ CopyGuard	84.86	60.40	21.28	51.62	54.54	62.47	68.46
DeepSeek-VL$_{7B}$	2.93	2.62	1.81	8.42	3.95	4.33	3.87
+ CopyGuard	86.80	59.64	44.76	54.62	61.46	71.63	73.64
GLM-4v$_{9B}$	4.78	2.92	1.35	3.42	3.12	2.69	2.45
+ CopyGuard	89.26	46.84	28.44	35.30	49.46	57.45	62.22
Janus-Pro$_{7B}$	1.86	2.63	1.89	5.01	2.85	4.27	3.74
+ CopyGuard	92.66	58.80	46.36	48.46	61.57	63.59	66.05

Table 2: The refusal rate of LVLMs after applying CopyGuard. The arrow indicates fewer copyright violations. (%)

come more complex, but these are still limited. For instance, the ROUGE-L score distribution of Qwen2.5-VL$_{7B}$ shifts to smaller values compared to Qwen2.5-VL$_{3B}$, and Qwen2.5-VL$_{7B}$ shows a lower refusal rate compared with Qwen2.5-VL$_{3B}$. Similarly, GPT-4o-mini and GPT-4o, LLaVA-1.5$_{13B}$ and LLaVA-1.5$_{7B}$ display similar trends. The latest model LLaVA-NeXT$_{13B}$ benefits from a copyright notice with a pronounced leftward shift in ROUGE-L score. Nonetheless, its copyright awareness without the presence of a notice is not stronger than in earlier versions. This suggests that while LVLM upgrades improve instruction-following, like recognizing copyright notices, they still lack understanding of copyright laws. To significantly enhance copyright compliance, it is crucial to innovate the models' architectures or implement more effective defensive measures.

CopyGuard effectively enhances the copyright compliance of LVLMs. We evaluate CopyGuard on various types of LVLM models under two scenarios: without and with explicit copyright notices. Specifically, in the former scenario, we assess whether the models can successfully defend against potential infringements. As shown in Table 2, CopyGuard significantly increases the refusal rate in all settings. We find that in the repetition task, CopyGuard achieved the most effective defense, evidenced by a refusal rate over 82%. In the extraction and translation tasks, CopyGuard also demonstrated a high level of defensive effectiveness.

CopyGuard rarely produces false positives. We evaluate previous queries on non-copyrighted content. In 80 interactions per set, we use GPT-4 and manual verification to assess whether LVLMs refuse responses. We find that CopyGuard does not result in any refusals. Additionally, we test non-infringing interactions on benchmarks such as MMMU, MMBench, and MathVista, showing that CopyGuard does not affect LVLMs' general performance for legitimate use.

Conclusion

In this work, we investigate whether LVLMs recognize and respect various copyrighted content in real-world scenarios. We introduce the first benchmark dataset designed to evaluate their copyright compliance across various infringement scenarios, infringement queries and copyrighted content. Through comprehensive experiments, we find that most LVLMs fail to comply with copyright regulations, increasing the risk of copyright infringement. To address this issue, we propose CopyGuard—a tool-augmented framework designed as a defensive copyright guardrail to track up-to-date copyright status and effectively reduce the generation of copyrighted content. Our findings underscore the urgent need to enhance LVLMs' awareness of copyright concerns for their ethical and legal deployment.

Acknowledgments

This work was partly supported by the National Key Research and Development Program of China under No. 2024YFB3908400, NSFC-Yeqisun Science Foundation under No. U244120033, NSFC under No. 62402418, 62402148, the Zhejiang Province's 2025 "Leading Goose + X" Science and Technology Plan under grant No.2025C02034, the Key R&D Program of Ningbo under No. 2024Z115, and the China Postdoctoral Science Foundation under No. 2024M762829.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.anthropic.com/news/claude-3-family.

Arts Law Centre of Australia. 2025. Music and Copyright — Arts Law Centre of Australia. https://www.artslaw.com.au/information-sheet/copyright-in-music-and-lyrics-aitb/.

Chang, K.; Cramer, M.; Soni, S.; and Bamman, D. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7312–7327. Singapore: Association for Computational Linguistics.

Chen, T.; Asai, A.; Mireshghallah, N.; Min, S.; Grimmelmann, J.; Choi, Y.; Hajishirzi, H.; Zettlemoyer, L.; and Koh, P. W. 2024. CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15134–15158. Miami, Florida, USA: Association for Computational Linguistics.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Cui, C.; Sun, T.; Lin, M.; Gao, T.; Zhang, Y.; Liu, J.; Wang, X.; Zhang, Z.; Zhou, C.; Liu, H.; Zhang, Y.; Lv, W.; Huang, K.; Zhang, Y.; Zhang, J.; Zhang, J.; Liu, Y.; Yu, D.; and Ma, Y. 2025a. PaddleOCR 3.0 Technical Report. arXiv:2507.05595.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025b. Multi-Modal Multi-Behavior Sequential Recommendation with Conditional Diffusion-Based Feature Denoising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1593–1602.

D'Souza, L.; and Mimno, D. 2023. The chatbot and the canon: Poetry memorization in LLMs. *Proceedings http://ceur-ws. org ISSN*, 1613: 0073.

Frankel, S.; and Gervais, D. 2014. *The evolution and equilibrium of copyright in the digital age*, volume 26. Cambridge University Press.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, H.; Yao, W.; Ma, K.; Yu, W.; Dai, Y.; Zhang, H.; Lan, Z.; and Yu, D. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6864–6890. Bangkok, Thailand: Association for Computational Linguistics.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jiang, D.; Zhang, R.; Guo, Z.; Wu, Y.; jiayi lei; Qiu, P.; Lu, P.; Chen, Z.; Song, G.; Gao, P.; Liu, Y.; Li, C.; and Li, H. 2025. MMSearch: Unveiling the Potential of Large Models as Multi-modal Search Engines. In *The Thirteenth International Conference on Learning Representations*.

Koh, J. Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.-Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; and Fried, D. 2024. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 881–905. Bangkok, Thailand: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, S.; Shi, Z.; Lyu, L.; Jin, Y.; and Faltings, B. 2025. CopyJudge: Automated Copyright Infringement Identification and Mitigation in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2502.15278*.

Liu, X.; Sun, T.; Xu, T.; Wu, F.; Wang, C.; Wang, X.; and Gao, J. 2024. SHIELD: Evaluation and Defense Strategies for Copyright Compliance in LLM Text Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1640–1670. Miami, Florida, USA: Association for Computational Linguistics.

Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

OpenAI. 2024. Model Specification.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

U.S. Copyright Office. 2023a. Title 17, Chapter 1, §106 of the U.S. Code. https://www.copyright.gov/title17/92chap1.html\#106.

U.S. Copyright Office. 2023b. U.S. Copyright Office Fair Use Index. https://www.copyright.gov/fair-use/.

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

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Xiang, D.; Xu, W.; Chu, K.; Ding, T.; Shen, Z.; Zeng, Y.; Su, J.; and Zhang, W. 2025. Promptsculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 774–786.

Xu, J.; Li, S.; Xu, Z.; and Zhang, D. 2024. Do LLMs Know to Respect Copyright Notice? In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20604–20619. Miami, Florida, USA: Association for Computational Linguistics.

Xu, W.; Xiang, D.; Ding, T.; and Lu, W. 2025a. MMM-Fact: A Multimodal, Multi-Domain Fact-Checking Dataset with Multi-Level Retrieval Difficulty. *arXiv preprint arXiv:2510.25120*.

Xu, W.; Xiang, D.; Liu, Y.; Wang, X.; Ma, Y.; Zhang, L.; Hu, S.; Xu, C.; and Zhang, J. 2025b. Finmultitime: A four-modal bilingual dataset for financial time-series analysis. *arXiv preprint arXiv:2506.05019*.

Xu, Z.; Han, M.; and Xing, W. 2025. EverTracer: Hunting Stolen Large Language Models via Stealthy and Robust Probabilistic Fingerprint. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 7019–7042. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Xu, Z.; Yue, X.; Wang, Z.; Liu, Q.; Zhao, X.; Zhang, J.; Zeng, W.; Xing, W.; Kong, D.; Lin, C.; et al. 2025c. Copyright Protection for Large Language Models: A Survey of Methods, Challenges, and Trends. *arXiv preprint arXiv:2508.11548*.

Xu, Z.; Zhao, X.; Yue, X.; Tian, S.; Lin, C.; and Han, M. 2025d. CTCC: A Robust and Stealthy Fingerprinting Framework for Large Language Models via Cross-Turn Contextual Correlation Backdoor. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6978–7000. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; and Sun, M. 2025. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In *The Thirteenth International Conference on Learning Representations*.

Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; and Wei, X. 2025a. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*.

Zeng, S.; Qi, D.; Chang, X.; Xiong, F.; Xie, S.; Wu, X.; Liang, S.; Xu, M.; and Wei, X. 2025b. JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation. *arXiv preprint arXiv:2509.22548*.

Zhang, J.; Xu, Z.; Hu, R.; Xing, W.; Zhang, X.; and Han, M. 2025. MEraser: An Effective Fingerprint Erasure Approach for Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30136–30153. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; and Chang, B. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations*.

Zhou, S.; Li, L.; Zhang, X.; Zhang, B.; Bai, S.; Sun, M.; Zhao, Z.; Lu, X.; and Chu, X. 2024. LiDAR-PTQ: Post-Training Quantization for Point Cloud 3D Object Detection. In *The Twelfth International Conference on Learning Representations*.