

Zero-shot Copyright Risk Assessment of Character Designs via Vision-Language Reasoning

Seongsoon Kim^{1,*} Jinseop Shin^{1,*} Seongchan Kim^{1,2,†} Minki Kim¹ Wonsu Kim^{1,2}

¹ Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

² UST-KISTI Applied AI, Daejeon, Republic of Korea

{seongkim, js.shin, sckim, mk.kim, wonsukim}@kisti.re.kr

Abstract

Copyright and trademark disputes are increasing due to growing conflicts over visual intellectual property (IP). These cases often involve subjective judgment and specialized legal expertise, making visual similarity assessment both time-consuming and difficult to standardize. This study investigates whether Vision-Language Models (ViLMs) can support structured reasoning about creative expression, a key factor in copyright infringement analysis. We propose a multi-step evaluation protocol that guides ViLMs through abstraction, filtration of generic elements, and comparison of expressive features, reflecting legal principles such as the Abstraction–Filtration–Comparison (AFC) test. Based on this framework, we develop a ViLM-based system that analyzes character and trademark images using staged prompts. We validate the method on a dataset of real-world legal disputes from South Korea and compare its performance with a CLIP-based similarity baseline. Experimental results show that ViLM achieves 87.10% accuracy on a six-point similarity scale and an F1 score of 0.9697 in a binary setting, outperforming CLIP in identifying legally meaningful expressive content. These results suggest that ViLMs can support expert-level reasoning in copyright analysis and may be useful in other knowledge-intensive multimodal applications.

1. Introduction

1.1. Growth of Visual IP Disputes

As the character and visual branding industries continue to expand, legal disputes involving the unauthorized use of visual designs such as characters, logos, and packaging have become increasingly common. These designs carry significant commercial value, and their originality and visual dis-

tinctiveness are now subject to growing legal scrutiny.

This trend is clearly reflected in recent national statistics: in 2023, the number of industrial property dispute mediation cases in South Korea more than doubled compared to the previous year, reaching an all-time high. Notably, 84% of these cases were initiated by individuals or small businesses, underscoring the need for more accessible and scalable mechanisms to resolve visual IP conflicts.

However, determining whether a visual design constitutes copyright or trademark infringement remains a labor-intensive and highly subjective task. The process typically relies on the judgment of domain experts, with no widely adopted, standardized criteria. For example, assessing whether a character’s long ears or stylized flower petals represent a creative expression or a generic motif often requires referencing many prior works and applying nuanced legal criteria. This ambiguity has led to inconsistent judgments and frequent disputes, highlighting the urgent need for a structured, reproducible framework to support such decisions.

1.2. ViLMs for Structured Copyright Reasoning

Recent advances in Vision-Language Models (ViLMs) have substantially improved their ability to understand complex visual inputs and generate reasoned outputs in natural language. Models such as CLIP [10], Flamingo [1], BLIP-2 [7], and MiniGPT-4 [17] have demonstrated strong performance in tasks involving open-ended visual question answering, image-text alignment, and multimodal reasoning.

While most prior applications of ViLMs have focused on generic tasks (e.g., captioning, retrieval), recent work has begun to explore their potential in more knowledge-intensive domains, such as legal decision-making [4, 8, 14]. These studies suggest that ViLMs can support structured reasoning and symbolic interpretation, though challenges in alignment and reliability remain.

In the legal domain, copyright infringement assessments often follow formalized protocols such as the Abstrac-

*Equal contribution

†Corresponding author

tion–Filtration–Comparison (AFC) test. These frameworks decompose the visual evaluation process into structured stages, making them well-suited to prompt-based ViLM execution. Rather than relying solely on pixel-level similarity, legal experts evaluate whether specific expressive elements have been copied—such as layout, gesture, or motif structure. This distinction motivates our investigation into whether ViLMs can emulate such structured reasoning.

We frame our investigation around two central research questions:

RQ1. *Can expert decision-making in copyright assessment be translated into a structured, step-by-step evaluation protocol?*

RQ2. *Can recent ViLMs carry out this reasoning process in a manner that is both reliable and interpretable, surpassing traditional similarity-based approaches?*

To address these questions, we propose a prompt-based framework that guides ViLMs through a multi-step visual reasoning process inspired by legal tests. We evaluate this system on real-world trademark dispute cases annotated by legal experts and compare its performance against an embedding-based similarity classifier as a baseline.

1.3. Contributions

This study presents a multimodal reasoning framework for assessing visual copyright infringement using a vision-language model (ViLM). Our method decomposes the expert decision process into structured steps and prompts that reflect key stages of legal evaluation, such as filtering generic elements and identifying expressive features. The prompt sequence was developed in consultation with a licensed attorney certified in copyright appraisal, ensuring that the reasoning process aligns with practical legal standards.

Our key contributions are as follows:

- We define a new evaluation task for visual copyright infringement that emphasizes expressive content over superficial resemblance, bridging legal standards and AI-based similarity assessment.
- We design a prompt-based reasoning protocol inspired by the Abstraction–Filtration–Comparison (AFC) test, enabling interpretable, stage-by-stage analysis within a single prompt structure.
- We develop and evaluate a ViLM-based system using real-world Korean trademark dispute data, achieving 87.10% top-1 accuracy on a six-point similarity scale and an F1 score of 0.9697 under a binary classification setting.
- We perform a comparative analysis against a CLIP-based similarity baseline, showing that ViLM achieves higher precision (0.9732 vs. 0.9697), higher recall (0.9663 vs.

0.9552), and better handling of cases involving abstract but legally relevant similarities.

To the best of our knowledge, this is the first study to apply ViLMs to structured legal reasoning in visual copyright assessment, grounded in real-world legal cases. The remainder of this paper is organized as follows: Section 2 reviews related work in copyright analysis and ViLM-based reasoning. Section 3 describes our method, including the prompt design and model architecture. Section 4 presents the experimental setup, comparative baselines, and results. Section 5 discusses key findings, failure cases, and limitations. Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1. Visual Similarity and Copy/Trademark Detection

Early approaches to visual similarity in copyright and trademark detection largely relied on deep feature retrieval or metric learning with convolutional networks. For instance, pre-trained CNNs have been adopted for trademark similarity retrieval [2], and robust photo identification pipelines were proposed to detect copyright violations under various manipulations [5]. Domain-specific methods have also explored motion/pose cues for animated or character-style content [6]. While effective in constrained settings, these methods are often tied to low-level resemblance or domain-specific priors, limiting generalization to diverse design spaces and offering limited alignment with legal doctrines.

The recent rise of vision-language models (ViLMs) has motivated a shift from purely perceptual similarity to multimodal reasoning. Foundational contrastive models (e.g., CLIP) [10] and language-bridged frameworks such as BLIP-2 [7], Flamingo [1], and MiniGPT-4 [17] improved open-ended visual understanding; advances in pretraining objectives like SigLIP further stabilized image–text alignment [15]. Building on these, recent work began examining copyright-specific detection using ViLMs. Xu et al. [14] analyze GenAI-related copyright infringement detection with large ViLMs and contrastive baselines, reporting high recall but residual false positives driven by superficial cues. Complementarily, *CopyJudge* [8] proposes prompt-based procedures inspired by legal heuristics to assess diffusion outputs with known provenance. In contrast to these directions, our work targets *real-world* design pairs and formulates a *post-hoc*, interpretable risk assessment pipeline that produces stepwise explanations (description → similarity → creative classification → risk scoring) aligned to legal reasoning stages.

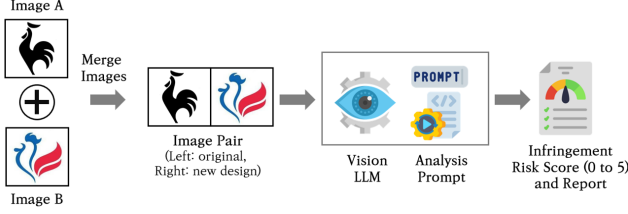


Figure 1. End-to-end workflow of our ViLM-based copyright infringement assessment system. The pipeline takes a pair of visual designs (original and candidate), applies a structured prompt-based reasoning process via a Vision-Language Model (ViLM), and produces a final infringement risk score.

2.2. Legal Reasoning and Vision-Language Alignment

Legal assessment goes beyond perceptual similarity and requires consistency with legal principles (e.g., idea-expression dichotomy, substantial similarity). Schefler et al. [11] propose a quantitative framework to formalize substantial similarity, emphasizing the need for principled metrics rather than superficial similarity measures such as edit distance or other surface-level overlaps. On the evaluation side, *LegalBench* [4] systematically benchmarks LLMs on textual legal reasoning, underscoring both the promise and current limitations of large models in faithfully following legal criteria. At the modeling level, a growing body of work in multimodal cognition examines how vision-language models (ViLMs) perform in higher-level reasoning tasks (e.g. causal inference, physical reasoning, theory of mind). In particular, Schulze Buschoff et al. [3] systematically probe ViLMs’ capabilities in intuitive physics, causal reasoning, and intuitive psychology, revealing both promising strengths and notable gaps in approaching human-level reasoning. Methodologically, we build on structured prompting paradigms, including chain of thought and least to most prompting, to elicit decomposed, inspectable reasoning traces [13, 16]. We also account for recent findings on prompt robustness and failure modes [9], underscoring the need for structured and controllable reasoning traces. Unlike black-box similarity scores, our protocol explicitly encodes legal-style reasoning stages and produces an interpretable rationale at each step. Empirically, we compare against a CLIP baseline [10] and show that a legally structured, prompt-driven pipeline better captures expressive (i.e., protectable) overlap than surface resemblance alone in real-world disputes.

3. Proposed Method

3.1. Task Definition and Evaluation Objectives

We address the task of visual copyright infringement assessment between two design images: one original x_o and

one potentially infringing x_c . Unlike traditional classification tasks, this problem requires legally-informed visual reasoning. To model this process, we propose a prompt-based evaluation protocol that emulates the sequential cognitive stages of human experts. Our goal is to guide ViLMs through structured visual assessments (Figure 1), thereby improving interpretability and aligning model reasoning with real-world legal workflows.

Although our method relies on a single prompt, the internal structure of the prompt encodes four distinct stages of expert judgment (abstraction, motif comparison, expressive filtration), and infringement scoring. Because decoder-only language models like Gemma [12] process text autoregressively, the model follows these stages sequentially during inference. This enables the ViLM to emulate a composed reasoning function:

$$\phi(x_o, x_c) = f_4(f_3(f_2(f_1(x_o, x_c)))) \quad (1)$$

where each f_i corresponds to a specific legal reasoning step embedded in the prompt template. This design allows the model to produce interpretable, legally-aligned outputs while preserving the efficiency of single-pass generation.

3.2. Prompt-Based Reasoning Framework

The protocol consists of four sequential prompts, each corresponding to a distinct cognitive operation in human expert judgment. Rather than relying on an all-in-one prompt, our framework guides the model through stepwise visual analysis, with each stage isolating a specific aspect of the reasoning process. This modular structure improves interpretability and provides diagnostic insight into where model decisions may succeed or fail (Table 1). Our staged prompting is conceptually related to chain-of-thought prompting [13], where decomposing complex reasoning into smaller steps enhances alignment with human judgments.

Importantly, our prompt sequence is explicitly designed to reflect the Abstraction–Filtration–Comparison (AFC) test used in U.S. copyright law. Prompt 1 extracts abstract design features, Prompt 2 filters out generic elements, and Prompts 3 and 4 conduct expressive comparison and scoring. This mapping ensures that the model’s reasoning aligns with legal doctrine. The design was developed in consultation with a licensed Korean attorney certified in copyright appraisal, ensuring consistency with both U.S. and Korean legal standards.

3.2.1. Design Feature Extraction.

The initial prompt asks the model to describe each image solely based on what appears without comparing it to others or jumping to any kind of interpretation at the moment. At this point, the aim of the analysis is just to lay out what the design looks like in terms of things like shape, proportion, outline, style, level of abstraction, and so on. This kind of

Table 1. Overview of the Prompt-Based Evaluation Framework for Visual Copyright Infringement Assessment (summarized; full prompt text omitted)

Prompt	Title	Purpose
1	Design Feature Extraction	Analyze each design separately. Describe shapes, proportions, key elements, stylistic features (e.g., outline, symmetry, color), and the level of abstraction. Avoid interpretation; only state what is visually present.
2	Design Similarity Comparison	Compare both designs side-by-side. Identify any visibly shared structures, motifs, layout similarities, or stylistic parallels (e.g., same curve repetition, alignment, thickness). Focus on surface-level visual similarity.
3	Similarity Classification	For each identified similarity, assess whether it reflects a common design element (e.g., a generic shape) or a distinct creative expression. Indicate whether the copying appears incidental or substantial.
4	Infringement Risk Scoring	Assign a risk score (0–5) based on how much the new design visually overlaps with the distinctive features of the original. Justify the score based on observed features only. Avoid ambiguous scoring.

step mirrors what human reviewers tend to do first: just taking in the visual impression before deciding whether there might be an infringement issue later on.

3.2.2. Design Similarity Comparison.

In the second stage, the model now directly compares the two designs. It identifies structural overlaps, recurring motifs, and compositional similarities. Importantly, this prompt restricts the model to surface-level comparisons focusing on what is visually shared without evaluating whether those similarities constitute infringement. This step is intended to emulate the visual alignment process used in expert analysis.

3.2.3. Similarity Classification.

The third prompt moves from identification to interpretation. Here, the model evaluates whether the observed similarities stem from commonplace design conventions (e.g., basic geometric shapes) or whether they reflect unique, expressive choices that may warrant protection. This is the first stage in which legal reasoning is introduced, as it mirrors the abstraction and filtration process employed by human experts to distinguish between generic and protectable features.

3.2.4. Infringement Risk Scoring.

Based on the previous reasoning outcome, the model finally gives a score from 0 to 5 (or 0 to 4 depends on scale), indicating how likely the new design is to infringe on the original. Note that the score comes with a short explanation why the model thinks like that based on previous observations derived from early part of the prompt, offering a compact summary that can be compared with expert assessments.

By organizing the entire reasoning process into the steps of recognition, comparison, and legal judgment as shown above, we are able to focus on the model’s individual output

at each step, analyze the rationale for the judgment, and interpret the results. It is also possible to see where the model’s conclusions differ from expert opinion, which gives possibility for improving the system in the future.

3.3. Infringement Scoring and Interpretation

In the final step, the model produces a score from 0 to 5, estimating how likely the new design is to infringe on the original. This score turns the model’s visual reasoning into a form that can be interpreted and compared more easily. We interpret the score ranges as follows:

Scores in the range of **0 to 2** indicate a *low risk of infringement*. In these cases, the model recognizes that any similarities between the two designs are either incidental or arise from the use of common, generic design elements such as basic shapes, standard color schemes, or symmetrical forms that are unlikely to be protected as original expression under copyright law.

A score of **3** reflects a *moderate level of risk*. This outcome suggests that the designs share some non-trivial visual elements such as similar layouts, motif arrangements, or repeated structures but that the degree of resemblance does not rise to the level of substantial copying. The similarity may result from stylistic convergence or accidental overlap rather than clear appropriation of expressive features.

Finally, scores in the range of **4 to 5** denote a *high risk of infringement*. In such cases, the model identifies substantial visual overlap in terms of expressive layout, distinctive compositional structure, or unique stylistic features that are specific to the original design. These high scores correspond to cases where the visual correspondence is both detailed and creative in nature, and would plausibly be flagged by human experts as potentially infringing. We treat the assigned score as a proxy for the model’s final infringement judgment and use it as the basis for quantitative evaluation

Table 2. Copyright Infringement Risk Scoring Guidelines

Score Range	Interpretation
0–2	Low infringement risk — Similarities are considered incidental, involving common or generic visual elements with no clear creative overlap.
3	Moderate risk — The design includes a notable combination of similar elements, but the alignment and composition are not clearly distinctive enough to imply direct copying.
4–5	High infringement risk — Substantial visual similarity exists, especially in creative layout, expressive features, or unique composition, indicating potential raise of concern.

against human-labeled ground-truth decisions.

To examine whether a coarser-grained scoring scheme could improve robustness, we additionally designed an alternative five-point scale (0–4), in which adjacent score ranges were collapsed. In this variant, scores of 0–1 were defined as indicating dissimilarity, while 2–4 were interpreted as reflecting varying degrees of similarity. This simplified scale was used alongside the original six-point version for comparative evaluation in later experiments.

3.4. System Information

3.4.1. Model and System Configuration.

Our system is built on Gemma 3 27B [12], an open-weight, decoder-only vision-language model by Google. It accepts interleaved image-text inputs in an autoregressive format and handles long contexts up to 128K tokens. For visual encoding, it employs a SigLIP-based encoder shared across all model sizes, converting images into token sequences aligned with the language stream. While optimized for 896×896 resolution, it supports larger or irregular inputs via adaptive windowing, enabling fine-grained reasoning over object details and embedded text. The experiments were conducted on a server equipped with multiple Intel Xeon Gold 6240 processors (total 72 physical cores), over 770GB of RAM, and a single NVIDIA A100 40GB GPU. With this setting reasoning task over the evaluation dataset took approximately 25 hours to complete.

4. Experiment

4.1. Dataset

To construct a realistic evaluation dataset for visual similarity assessment, we collected trademark rejection notices issued by examiners from a national intellectual property database.¹ Each case includes a pair of images: a new (rejected) trademark design and a prior-registered design cited as the reason for rejection. These image pairs serve as real-world examples of claimed visual similarity in legal decisions.

¹The dataset was collected from an official national trademark registry and will be made available upon request.

However, because examiner opinions reflect the perspective of a limited number of officials and may not align with broader public perception, we conducted a human validation study. Ten independent human raters evaluated each image pair in a binary forced-choice task, indicating whether the pair appeared visually similar or dissimilar. To ensure high inter-rater agreement, we retained only those 3,764 image pairs for which all raters unanimously agreed on the similarity label. This strict filtering minimizes subjectivity and increases the reliability of ground truth annotations.

It is important to note that the dataset is inherently imbalanced. Since trademark examiners tend to flag potentially infringing designs, most image pairs represent borderline or strong visual similarity. In fact, 96% of the unanimously agreed pairs are labeled as *similar*. This distribution reflects real-world IP adjudication scenarios, where false negatives (failing to identify infringing similarity) pose a higher risk than false positives. As such, we emphasize recall as a key metric for evaluating model utility in legal screening pipelines.

4.2. Models and Evaluation Setup

We compared two approaches:

- **ViLM-based system:** A structured reasoning framework using a Vision-Language Model (ViLM) guided by multi-stage prompts reflecting legal tests (e.g., Abstraction-Filtration-Comparison). The model outputs a score (0–5 or 0–4) for each image pair based on expressive similarity.
- **CLIP [10]:** A general-purpose model (openai/clip-vit-base-patch32) that computes cosine similarity between image embeddings. Since CLIP lacks legal awareness, it represents a strong perceptual baseline.

For ViLM, we evaluated performance under both a six-point scale (0–5) and a five-point scale (0–4). Binary classification was obtained by thresholding the model scores: scores ≤ 2 (or ≤ 1 , respectively) were labeled as *dissimilar*. For CLIP, we varied the cosine similarity threshold from 0.50 to 0.85 in increments, and report the corresponding accuracy, precision, recall, and F1 score for each setting.

Table 3. Comparison of confusion matrices: ViLM vs. CLIP (threshold = 0.50). ViLM uses legal-style reasoning; CLIP uses cosine similarity on raw embeddings.

Method	Ground Truth / Prediction	Pred Similar	Pred Dissimilar
ViLM (0–5)	Actual Similar	3,152	466
	Actual Dissimilar	85	61
ViLM (0–4)	Actual Similar	3,496	122
	Actual Dissimilar	96	50
CLIP (0.50)	Actual Similar	3,456	162
	Actual Dissimilar	108	38

Table 4. Precision, recall, and F1 score comparison for each method.

Method	Precision	Recall	F1
ViLM (0–5)	0.9737	0.8713	0.9199
ViLM (0–4)	0.9732	0.9663	0.9697
CLIP (0.50)	0.9697	0.9552	0.9623

4.3. Main Results

Table 3 shows the confusion matrices for ViLM and CLIP. The ViLM-based method exhibits strong alignment with human judgments, particularly when using the 0–4 similarity scale. It correctly identifies 3,496 out of 3,618 human-labeled similar cases, while producing relatively few false positives (96) and false negatives (122).

In contrast, CLIP (with a cosine similarity threshold of 0.50) shows slightly weaker alignment: it correctly classifies 3,456 similar cases but misses 162 (false negatives) and falsely labels 108 dissimilar pairs as similar (false positives). These misclassifications are reflected in the model’s recall and precision performance.

Table 4 summarizes the corresponding precision, recall, and F1 scores for each model. ViLM (0–4) achieves a precision of 0.9732, a recall of 0.9663, and an F1 score of 0.9697 outperforming CLIP, which attains an F1 score of 0.9623. The drop in recall for CLIP highlights its tendency to miss cases judged similar by humans, especially when expressive or abstract features are involved.

Overall, these results demonstrate that ViLM’s structured, reasoning-based approach is better suited for capturing nuanced visual similarity, which is critical in copyright contexts where both false positives and false negatives carry meaningful consequences.

4.4. CLIP Threshold Analysis

To better understand the limitations of embedding-based similarity, we evaluated CLIP across a range of cosine similarity thresholds. As shown in Figure 2, CLIP achieves high recall at lower thresholds (e.g., 0.50–0.60), correctly identifying most similar cases. However, this comes at the cost of reduced precision, as it tends to over-identify similarity and

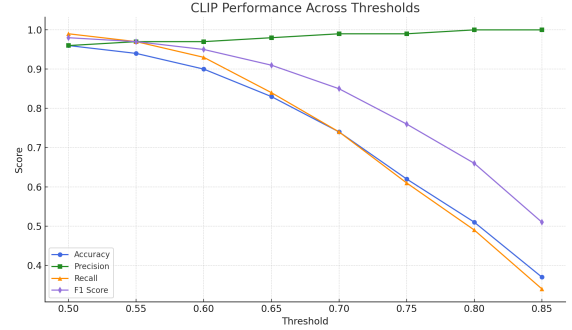


Figure 2. CLIP performance across cosine similarity thresholds. Lower thresholds yield higher recall but reduced precision, while higher thresholds improve precision at the expense of recall.

produces a large number of false positives. As the threshold increases, precision improves, but recall sharply declines, resulting in a steep tradeoff and a rapid drop in F1 score.

This pattern is partly due to the composition of our dataset, in which most pairs are labeled similar by human annotators. At lower thresholds, CLIP benefits from this imbalance by labeling most inputs as similar, thus achieving high recall. However, when the threshold is raised to reduce false positives, CLIP begins to miss many cases that humans still regard as similar. This discrepancy suggests that CLIP’s similarity judgments based solely on cosine distance between global embeddings are not aligned well with human perception, especially for expressive or stylistic similarity.

These findings indicate that CLIP prioritizes coarse visual features which may not adequately capture the semantic or expressive content of a design. In contrast, our ViLM-based method, structured to emulate legal reasoning through abstraction and filtration demonstrates greater robustness across cases, including those involving subtle or symbolic similarities.

4.5. Analysis and Implications

These findings highlight the importance of structured reasoning in high-stakes applications. While CLIP offers strong performance in perceptual similarity detection, it fails to distinguish between superficial resemblance and substantive expression. In contrast, ViLM prompts aligned with legal criteria enable interpretable, step-by-step evaluation consistent with human legal reasoning.

Importantly, all evaluations were conducted in a *zero-shot* setting without fine-tuning on task-specific labels. This underscores the generalization ability of our ViLM approach and its potential utility as a decision support system in copyright-sensitive domains.

5. Discussion

5.1. Case Study Analysis

To assess the interpretability and robustness of ViLM’s reasoning, we conducted qualitative analysis across representative cases including successful detection, misclassification, and ambiguous outcomes. We also compared ViLM’s structured reasoning to CLIP’s embedding-based similarity. Table 5 summarizes five illustrative examples used in this analysis.

The first two cases (1-1 and 1-2) clearly demonstrate ViLM’s capacity to detect expressive similarity that goes beyond superficial shape or texture. Unlike pixel-level matching or low-level embedding comparison, ViLM identifies the spatial layout, functional alignment, and symbolic reuse of visual elements, features central to copyright law. For instance, in Case 1-1, the model not only detected mirrored postures and anatomical alignment (neck, claws, wings), but also emphasized their relational structure and compositional logic. Likewise, in Case 1-2, ViLM recognized that although the new design simplified the original, the expressive arrangement (rider’s pose, horse motion, silhouette flow) was preserved pointing to substantial similarity. These examples show that when properly prompted, ViLMs can emulate structured legal heuristics such as the Abstraction–Filtration–Comparison (AFC) test by isolating creative components and assessing their reuse in context. The model’s scoring was consistent with human expert judgments, suggesting that symbolic reasoning over visual features is both feasible and reliable under our prompting scheme.

In contrast, Case 2 reveals a notable failure. Although both designs shared an elongated, branching form, they originated from distinct typographic roots. The model over-relied on surface similarity and failed to account for design context. Similarly, Case 3 illustrates how ViLM missed human-perceived rhythm and abstraction, underestimating similarity in minimalist settings. These limitations underscore the gap between symbolic reasoning and gestalt perception.

Case 3 shows the model’s judgment diverged from that of human reviewers, likely due to the ambiguous nature of the visual similarity. The original design consists of three circular elements stacked vertically with a dotted texture, while the new one features a smooth, red curve that loosely resembles an “S” shape. Despite their structural and stylistic differences, all human annotators unanimously labeled the pair as *similar*, likely due to the shared visual rhythm, vertical orientation, and gesture-like impression created by the flowing forms. However, the model assigned a relatively low infringement risk score of 1. This implying that the model thinks the two designs differed significantly in expressive features. According to the model’s

step-by-step reasoning output, vertical stacking and curved form were interpreted as common compositional devices rather than indicators of copying. The model emphasized that the original’s granular texture and discrete segmentation contrast with the new design’s clean, continuous stroke suggesting that these factors outweigh surface level resemblance. This case illustrates a challenge for vision-language models. Although they are effective at identifying structural differences, it seems that they overlook higher-level visual similarities or stylistic patterns that human reviewers tend to recognize or that can catch only by human imagery. This limitation becomes more pronounced in abstract or minimalist designs, where legal notions like “overall impression” or “creative expression” are difficult to capture through model-based reasoning.

Finally, Case 4 illustrates a fundamental limitation of embedding-based methods. While CLIP overlooked the similarity between a newly filed logo and the registered Mary Quant mark due to color and edge variations, ViLM’s symbolic assessment correctly flagged potential infringement. This demonstrates ViLM’s advantage in interpretability and alignment with legal heuristics.

5.2. Limitations and Practical Implications






While our framework shows strong potential for structured visual copyright analysis, some limitations remain particularly in relation to its scope and operational boundaries.

First, although our prompting strategy and similarity rating scheme were developed in consultation with legal domain experts, the outputs themselves are not legally enforceable judgments. Copyright infringement assessments in practice are often complex and depend on contexts such as prior art, market positioning, cultural relevance, and judicial precedent that lie beyond the current input space of ViLMs. Therefore, while our model mirrors legal reasoning patterns and benefits from expert-informed criteria, it should be understood as an assistive system that supports rather than replaces human legal expertise.

Second, like many large language models, ViLM occasionally defaults to ambiguous or neutral ratings (e.g., ‘3’) when uncertainty is high. This reflects a known behavioral tendency observed in reasoning tasks [16], and highlights the importance of prompt robustness [9]. We addressed this by enforcing stricter prompt instructions, which improved decisiveness. However, this process revealed high sensitivity to phrasing, raising reproducibility concerns in legal-adjacent applications.

Despite these challenges, the structured reasoning outputs generated by our system provide valuable support in real-world workflows. By decomposing visual analysis into interpretable stages (description, comparison, classification, and scoring) the framework enables traceable, auditable assessments. These can be integrated into preliminary screen-

Table 5. Representative case studies comparing ViLM judgments and human evaluations across different scenarios.

Case Type	Image Pair (Left: Original, Right: New)	ViLM Judgment	Model Reasoning and Assessment (Summarized)
Case 1-1 Stylized Animals		Similar (Score 5)	Identified expressive similarity beyond shape—e.g., posture, neck curve, wing/foot elements. Mirrored structure triggered infringement concern.
Case 1-2 Rider Silhouette		Similar (Score 5)	Detected shared layout and expression despite reduced detail. Focused on core compositional reuse.
Case 2 Abstract Letterforms		Similar (Score 4)	Misclassified due to structural resemblance; ignored typographic context and stylistic divergence.
Case 3 Ambiguous Design		Dissimilar (Score 1)	Underestimated compositional rhythm perceived by humans; overlooked holistic abstraction.
Case 4 CLIP vs ViLM		Similar (ViLM) Dissimilar (CLIP)	ViLM correctly prioritized the iconic shape and compositional abstraction of the flower mark, while CLIP was distracted by the embedded text and failed to capture their visual similarity.

ing processes within copyright offices, legal audits, or content moderation pipelines to flag potentially infringing designs for human review.

Ultimately, our findings suggest that ViLMs, when guided by tailored prompts incorporating expert knowledge and task decomposition, can perform legally relevant visual reasoning with high interpretability and operational efficiency. Future work may explore instruction tuning, domain-specific reward modeling, or multimodal calibration to further enhance reliability and extend applicability across broader categories of intellectual property.

6. Conclusion

In this paper, we proposed a structured, prompt-based framework for assessing visual similarity and potential copyright infringement using recent vision-language models (ViLMs). Our approach mirrors the step-by-step reasoning of human experts by decomposing the legal judgment process into modular prompts, covering abstrac-

tion, filtration, comparison, and risk scoring. This design aligns with real-world legal heuristics such as the Abstraction–Filtration–Comparison (AFC) test, enabling interpretable assessments beyond surface resemblance.

We evaluated our method using a dataset of real trademark application cases from South Korea, including examples with strong human consensus. The ViLM-based system achieved 87.10% accuracy in identifying `similar` cases under a six-point scale. When binarized to a five-point scale (`similar` vs. `dissimilar`), ViLM attained a precision of 0.9732, a recall of 0.9663, and an F1 score of 0.9697, demonstrating close alignment with expert annotations across both positive and negative cases. Qualitative case studies further showed that ViLM could reason about expressive similarity, compositional logic, and motif reuse, key factors in copyright and trademark disputes.

To benchmark our approach, we compared ViLM’s output to CLIP [10], a widely used embedding-based model. CLIP achieved a slightly lower F1 score of 0.9623 at a

cosine threshold of 0.50, with lower recall (0.9552) and slightly lower precision (0.9697). While CLIP maintained high recall at low thresholds, it exhibited a steep precision–recall tradeoff and failed to capture many subtle similarities identified by human reviewers. This contrast highlights the limitations of cosine similarity as a proxy for legal similarity. In contrast, ViLM’s reasoning-based output enabled symbolic interpretation and justified decision-making, which are essential in legal review workflows.

Nonetheless, we identified several challenges. The model sometimes misclassified typographic or abstract forms, indicating a lack of contextual understanding about design conventions. In ambiguous or minimalist cases, it occasionally underestimated holistic similarity as perceived by human reviewers. These limitations point to future directions for improvement, including fine-tuning with domain-specific datasets, integrating retrieval-augmented generation (RAG) for contextual grounding, and extending the framework to other visual IP domains such as character design, UI elements, or packaging.

Ultimately, our goal is not to replace legal expertise, but to augment it. The proposed ViLM framework offers transparency, consistency, and efficiency, particularly valuable in high-volume IP screening scenarios where manual review is costly. As visual IP continues to gain commercial and legal significance, we believe structured multimodal reasoning systems can play a meaningful role in bridging the gap between AI capabilities and expert decision-making in copyright law.

Acknowledgments

This research was supported by the Korea Creative Content Agency (KOCCA) under the Leading Copyright Technology Development (R&D) Program (RS-2024-00397382) and by the Korea Institute of Science and Technology Information (KISTI) under grants N25NT016-25 (NTIS No. 2370000138), K25L1M4C4 (NTIS No. 2710087200).

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, and et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 1, 2
- [2] Hayfa Alshowaish, Yousef Al-Ohali, and Abeer Al-Nafjan. Trademark image similarity detection using convolutional neural network. *Applied Sciences*, 12(3):1752, 2022. 2
- [3] Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7:96–106, 2025. 3
- [4] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, and et al. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 3
- [5] Doyoung Kim, Sungwoong Heo, Jiwoo Kang, Hogab Kang, and Seonghoon Lee. A photo identification framework to prevent copyright infringement with manipulations. *Applied Sciences*, 11(19), 2021. 2
- [6] De Li, Lingyu Wang, and Xun Jin. Cartoon copyright recognition method based on character personality action. *EURASIP Journal on Image and Video Processing*, 2024(1), 2024. 2
- [7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 1, 2
- [8] Shunchang Liu, Zhuan Shi, Lingjuan Lyu, Yaochu Jin, and Boi Faltings. Copyjudge: Automated copyright infringement identification and mitigation in text-to-image diffusion models. *arXiv preprint arXiv:2502.15278*, 2025. 1, 2
- [9] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models, 2022. 3, 7
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1, 2, 3, 5, 8
- [11] Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for copyright law’s substantial similarity. *arXiv preprint arXiv:2206.01230*, 2022. 3
- [12] Gemma Team. Gemma 3 technical report, 2025. 3, 5
- [13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 3
- [14] Qipan Xu, Zhenting Wang, Xiaoxiao He, Ligong Han, and Ruixiang Tang. Can large vision-language models detect images copyright infringement from genai?, 2025. 1, 2
- [15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [16] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. 3, 7
- [17] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1, 2