Prompt-Conditioned Scene Reconstruction based on Stable Diffusion for Investigative Hypothesis Testing[*]

Jeong-hwa Ryu, Da-hyung Kim, Su-bin Lee, and Seongmin Kim[†]

Sungshin Women's University, Seoul, Korea
{20231093, 20231065, 20222605, sm.kim}@sungshin.ac.kr

Abstract

The reliability of witness testimony remains a persistent challenge in forensic investigation, as discrepancies between statements and evidence often lead to interpretive ambiguity. To address this issue, this study proposes a forensic framework that integrates Stable Diffusion and CLIP to quantitatively assess the semantic consistency between language and vision. Experimental results show that CLIP similarity systematically decreases with linguistic variations in action, agent, and environment, indicating that the model is more sensitive to action- and identity-centered semantic differences than to simple visual similarity. These findings suggest that diffusion-based generative models can be repurposed for forensic analysis, enabling interpretable, quantitative validation of testimonial coherence and semantic alignment between statements and visual evidence.

Keywords— Stable Diffusion, CLIP, Semantic consistency, Testimony Reliability, Forensic framework, Text-to-Image Generation

1 Introduction

In contemporary forensic practice, the reliability of witness testimony remains a persistent challenge. A recent study in the United Kingdom reported substantial discrepancies between the oral statements of witnesses or victims and their subsequent written statements, with information omissions and distortions occurring in most cases [1]. This problem extends beyond individual memory lapses, suggesting that structural distortions during statement collection and recording can seriously affect the accuracy of legal judgments. Indeed, data from the U.S. National Registry of Exonerations indicate that approximately 63% of wrongful convictions are attributable to forensic and testimonial errors [2, 3].

In other words, while courtroom testimony is assumed to represent factual truth, a range of psychological and procedural factors, including memory distortion, fatigue, stress, and suggestive questioning, continually undermine its reliability [4]. As a result, the boundary between what was seen and what was said becomes blurred, producing a persistent gap between testimony and evidence.

To address this cognitive instability, recent quantitative studies have sought to systematically measure testimonial errors. For instance, several evaluation models have been proposed to assess the reliability of statements by analyzing variables, such as alibi consistency, contradiction rates between statements, and the sequence effect [5, 6]. These developments highlight the need for novel forensic analysis methodologies that move beyond simple manipulation detection and instead quantitatively evaluate the semantic alignment between verbal testimony and visual evidence.

Meanwhile, diffusion models have recently emerged as a powerful framework for high-quality text-to-image generation [7]. Among these models, stable diffusion (SD) has gained particular attention due to its latent-space design, which allows users to express subtle semantic variations through text prompts and to directly manipulate or analyze internal representations during the generation process [8]. In

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. W4, December 16-18, 2025, Sapporo , Japan.

[†] Corresponding author

particular, the use of attention maps enables quantitative tracking of semantic changes resulting from modifications in textual conditions. Owing to these properties, stable diffusion has been applied in various fields, including visual fidelity evaluation of generated images [9], forensic integrity verification of AI-generated content, and anomaly detection [10].

In this paper, we explore the potential of adopting the Stable Diffusion model to address cognitive instability in forensic investigations. Specifically, we propose a forensic simulation framework that integrates stable diffusion with contrastive language-image pre-training (CLIP) to construct an interpretable analytical model capable of visualizing the semantic consistency of testimonies. To demonstrate its feasibility, we conduct a pilot study that systematically varies prompt elements, including action, subject, and scene, which constitute the core components of forensic narrative analysis. Through quantitative evaluation and visual comparison of multiple statements describing the same incident, this study illustrates how diffusion-based generative simulation can serve as a supportive tool for validating and analyzing testimonial coherence in digital forensics.

2 Background and Literature Review

2.1 Stable Diffusion Model

Stable Diffusion provides access to its internal mechanisms, including latent representations, attention maps, and seed fixation. This architectural transparency is particularly valuable for research that requires precise tracking and verification of semantic alignment between text and image, such as semantic consistency analysis. In addition, the model's open design, coupled with the expressive capacity of its latent space, enables detailed examination of the intrinsic semantic relationships between linguistic and visual information. Notably, Stable Diffusion's latent space preserves relatively independent, or disentangled, representations for each semantic unit [8], allowing fine-grained analysis of how text prompts influence generated visual semantics.

Building on these capabilities, recent studies have simultaneously analyzed the semantic consistency and reliability of AI-generated content [9, 11]. This line of research demonstrates that stable diffusion is evolving beyond a simple image generator into an interpretable generative environment capable of supporting semantic structure analysis and reliability evaluation. Accordingly, it serves as an appropriate foundation for assessing the semantic coherence between linguistic statements and visual evidence, as well as for visually exploring that consistency.

2.2 CLIP Model

CLIP is a vision–language model that learns a joint embedding space by aligning paired text and image representations based on contrastive learning [12]. Building on this foundation, Jiang et al. identified the compositional comprehension limitations of CLIP-based models and proposed ComCLIP to address it through causal disentanglement and subimage-level matching [13]. ComCLIP refines visual–linguistic alignment by explicitly tuning the mapping between sentence components (e.g., subject, verb, object) and corresponding visual regions, restoring coherence at the linguistic component level. Also, Tan et al. [14] and Bent et al. [15] demonstrated that CLIP embedding distances significantly correlate with text–image semantic consistency. Our study adopts this structural perspective but shifts the focus from learning-based consistency correction to non-learning semantic detection.

3 Pilot Study Overview

In our pilot study, we utilize CLIP's cosine similarity not as an alignment performance metric but as an internal semantic stability indicator, interpreting its collapse as a signal of semantic drift between images generated by Stable Diffusion. To this end, we extend prior work [14, 15] by applying the same mechanism to a forensic reasoning context, establishing a simulation framework to detect the maintenance or breakdown of semantic coherence among statements describing the same event.

Experimental Setup. Each statement was represented as a text prompt, and an image (512×512 resolution) was generated using the Stable Diffusion v1.5 model (runwayml/stable-diffusion-v1-5) under 50 inference steps and a guidance scale of 7.5, within an Anaconda-based Python virtual environment. The input prompts were derived from the baseline sentence and were categorized into three modification types: (1) Action modification (case A), (2) Subject modification (case B), and (3) Scene modification (case C).

Table 1 summarizes the variation examples derived from the baseline prompt embedding the context, "A man is unloading boxes from a white truck". This design ensures that the core linguistic components relevant to forensic analysis sensitively align with their corresponding visual regions in the generated images. Note that all experiments were performed under identical conditions (e.g., random seed and generation parameters) to ensure consistency across runs. Non-linguistic factors were controlled as much as possible to isolate the effect of prompt semantics on the generated outputs. In total, nine images were produced, corresponding to three variations within each modification type.

Table 1: Semantic variation cases by type (action, subject, scene)

<table>
<tr><th colspan="2">Testimony (baseline)</th></tr>
<tr><td colspan="2">A middle-aged <b>man (B1)</b> wearing a neatly pressed gray work uniform and black safety shoes is <b>unloading (A1)</b> several taped cardboard boxes from the open rear of a clean white delivery truck parked on a paved <b>city street (C1)</b>. His posture is slightly bent forward as he grips one box with both hands, preparing to place it on the curb beside the truck. The vehicle's metallic surface reflects patches of sunlight and nearby buildings, showing subtle reflections of the street environment. Fine dust floats in the bright midday air, and the image captures realistic daylight tones with soft natural shadows and slight 35 mm film grain texture.</td></tr>
<tr><th colspan="2">Variations</th></tr>
<tr><td><b>Case A − action</b></td><td>A2: unloading → loading
A3: unloading → washing (truck)</td></tr>
<tr><td><b>Case B − subject</b></td><td>B2: man → young woman
B3: man → firefighter</td></tr>
<tr><td><b>case C − scene</b></td><td>C2: city street → snow-covered city street
C3: city street → warehouse interior</td></tr>
</table>

4 Result Analysis

For the three types of modified prompts generated by Stable Diffusion, (A) Action, (B) Subject, and (C) Background, we compared the corresponding image triplets within the CLIP embedding space to derive the Semantic Consistency Index (SCI) values. The overall results are illustrated in Figure 1 and Table 2. Subsequently, a detailed quantitative analysis was conducted for each case to examine variations in semantic stability across prompt types.

4.1 Semantic Consistency Index (SCI): Quantitative results

Action Variation. The CLIP similarity map for the action set (A1–A3) indicates that A1–A2 (0.83) are semantically closer than A1–A3 (0.70). However, in the actual images (Figure 2), A1 ("unloading") and A3 ("washing") appear visually more similar, both depict a man beside a truck under nearly identical composition and lighting conditions. The lower similarity score between A1 and A3 therefore reflects CLIP's emphasis on verb-level semantics rather than visual form, interpreting the change in

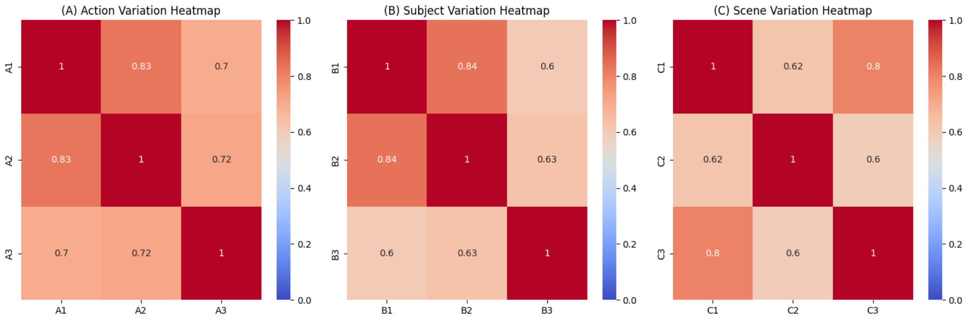

Figure 1: CLIP Cosine Similarity Heatmaps for 3 version

Table 2: Semantic Consistency Index (SCI) Results Across Variation Types

Comparison	Cosine similarity	Interpretation
A1-A2	0.830	Only the action changes within the same context → High consistency
A1-A3	0.700	Change in behavioral purpose → Breakdown of consistency
B1-B2	0.840	Only gender differences exist→ Semantic retention
B1-B3	0.600	Changes in occupational identity → Increased semantic distance
C1-C2	0.620	Weather and lighting changes → Partial drift
C1-C3	0.800	Maintaining core behaviors despite location changes → High consistency

action type (“unloading” → “washing”) as a major conceptual shift even when the visual elements (e.g., person, truck, and background) remain constant. This finding suggests that CLIP’s feature space is more sensitive to linguistic semantics than to structural resemblance, underscoring its language-oriented bias in visual reasoning [16].

Subject Variation. The comparison of B1–B2 (gender change) showed a high similarity of 0.84, whereas B1–B3 (occupational change) dropped sharply to 0.60. In B3, the box (object) was removed, and occupational cues, such as firefighter uniforms and fire trucks, became dominant. Despite visually similar compositions, these images were classified into distinct semantic domains within the CLIP embedding space. Specifically, CLIP encoded “man carrying box” as carrying action + object + contextual background, while “two firefighters” was represented as occupation + uniform + fire truck context [17].

As a result, when the box was removed and the uniform appeared (B3), the model reinterpreted the scene as “firefighter-related” rather than “box-handling”. Consequently, the semantic distance between B1 and B3 expanded, producing a lower CLIP similarity score. Given CLIP’s contrastive learning objective, a decrease in similarity corresponds to an increase in semantic distance, indicating that the model strongly detected semantic differences between the two scenes [18]. Overall, these results suggest that CLIP is more sensitive to social roles and action semantics than to gender attributes.

Scene Variation. The CLIP similarity between C1–C2 was 0.62, while C1–C3 measured 0.80. C2 (snowy scene) exhibited substantial differences in hue, lighting, and texture, producing a large visual

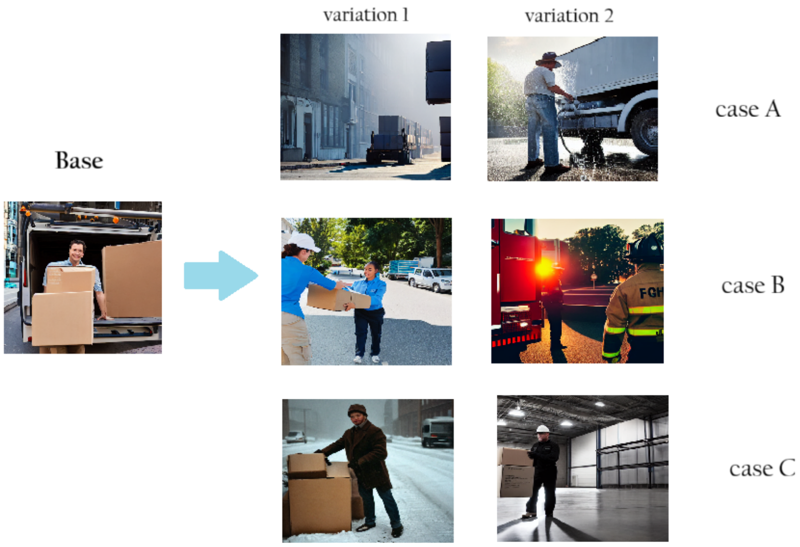

Figure 2: Generated Images Across Action, Subject, and Scene Modifications

deviation from the baseline image. CLIP interpreted this as a change in environmental texture and accordingly assigned a lower similarity score. In contrast, C3 (warehouse interior) preserved the core action of unloading boxes despite the altered background, leading CLIP to assign a higher similarity score (0.80). These results demonstrate that CLIP primarily prioritizes action semantics over scene context when evaluating semantic consistency [19].

4.2 Discussion and Lesson Learned

The overall consistency pattern suggests that CLIP exhibits a hierarchical sensitivity to semantic cues, assigning the greatest weight to action, followed by social role and scene context. This hierarchy implies that semantic coherence between statements is maintained as long as the core action–agent structure remains stable within the embedding space. Once these axes are disrupted through shifts in action or agent identity, CLIP registers a measurable expansion in embedding distance, signaling the onset of semantic drift.

From a forensic perspective, this behavior allows CLIP's cosine similarity to be reinterpreted not as a mere alignment score but as a diagnostic variable that reveals when semantic stability collapses between multiple statements describing the same event [20]. Accordingly, this study positions semantic drift detection as a non-learning, interpretable forensic reasoning procedure, demonstrating that quantitative text–image divergence can serve as an indicator of narrative inconsistency or testimonial deviation.

Limitations. The pilot study demonstrates the feasibility of quantitatively assessing semantic consistency between language and vision, yet it remains a proof-of-concept implementation with several limitations. First, because publicly available pre-trained models were employed in a non-learning configuration, internal noise and inherent model biases could not be fully controlled during the image generation process. Second, the stochastic nature of stable diffusion occasionally produced slight variations across sessions and computing environments, limiting perfect reproducibility. Finally, while CLIP similarity effectively captured internal semantic relations among statements, it could not evaluate the factual accuracy or truthfulness of the described events. In other words, the current framework measures sensitivity to semantic variation rather than determining correspondence to real-world evidence.

5 Conclusion

This study presented a proof-of-concept framework that integrates Stable Diffusion and CLIP to quantitatively analyze the semantic consistency between language and vision. The results showed that CLIP is more sensitive to action- and identity-centered semantic variations than to mere visual similarity, suggesting that the model captures conceptual rather than superficial correspondences. These findings highlight the potential of reinterpreting semantic inconsistencies as forensic indicators, establishing a foundation for future research on AI-assisted statement verification. While the current framework remains a preliminary exploration of semantic sensitivity rather than a comprehensive truth-verification system, future work will expand its scope through external consistency validation using ground-truth images or reference scenes and through human–model alignment analysis. With these advancements, the proposed CLIP-based semantic consistency framework may evolve into a practical methodology for multimodal credibility assessment and digital forensic reasoning.

Acknowledgments

This work is partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00351898 and No. RS-2025-02263915), the MSIT under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

References

[1] Friends of the Forensic Science Club, *"Witnesses and victims' statements correctly collected when written"*, Evidentia University, 2025. https://evidentiauniversity.com/blogs/forensic/witnesses-and-victims-statements-correctly-collected-when-written-forensic-science-club/

[2] Saks, M.J. and Koehler, J.J., *Wrongful convictions and claims of false or misleading forensic evidence*, Journal of Forensic Sciences, 2023, vol. 68, no. 4, pp. 1234-1245, https://onlinelibrary.wiley.com/doi/10.1111/1556-4029.15233

[3] U.S. National Registry of Exonerations, *Causes of Wrongful Convictions*. 2023 report.

[4] Author(s), *The effect of credibility assessment techniques on witness accuracy*. Frontiers in Psychology, 2023.

[5] Qi, et al., *Evaluating the validity of testimony and consistency*. Forensic Psychology Journal, 2024.

[6] Smith, et al., *Systematic analysis of misleading evidence in trials*. Journal of Criminal Justice, 2023.

[7] Research on Image Transformation and Generation using Diffusion Model, Korea University, Seoul, 2025. https://mil.korea.ac.kr/bbs/board.php?bo_table=Research&wr_id=24

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B., *High-Resolution Image Synthesis with Latent Diffusion Models*, arXiv preprint arXiv:2112.10752, 2022. https://arxiv.org/abs/2112.10752

[9] Sharma, R., et al., *Explainable AI-Generated Image Forensics: A Low-Resolution Perspective with Novel Artifact Analysis*, ICCV Workshop on AI for Media Forensics and Deepfake Detection, 2025.

[10] Stamnas, N., et al., *Exposing Deepfakes using Differential Anomaly Detection*, WACV Workshop on AI for Media Forensics and Deepfake Detection, 2024.

[11] Levin, E. and Fried, O., *Differential Diffusion: Giving Each Pixel Its Strength*, arXiv preprint arXiv:2306.00950,June 2023. https://arxiv.org/abs/2306.00950

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever., *Learning Transferable Visual Models From Natural Language Supervision*, arXiv preprint arXiv:2103.00020, 2021. https://arxiv.org/abs/2103.00020

[13] Jiang, Kenan, He, Xuehai, Xu, Ruize, and Wang, Xin Eric. "ComCLIP: Training-Free Compositional Image and Text Matching." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024) – Long Papers*, pp. 6639–6659, 2024.

[14] Tan, Zhaorui and Yang, Xi and Ye, Zihan and Wang, Qiufeng and Yan, Yuyao and Nguyen, Anh and Huang, Kaizhu. "SSD: Towards Better Text-Image Consistency Metric in Text-to-Image Generation." arXiv preprint arXiv:2210.15235, 2022. https://arxiv.org/abs/2210.15235

[15] Bent, B., Pande, A., and Weng, T. "A Semantic Approach to Quantifying the Consistency of Diffusion Model Image Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 312–321, 2024.

[16] Marasovic, A., & Lopatovska, I., *VerbCLIP: Improving Verb Understanding in Vision-Language Models*, arXiv preprint arXiv:2304.10049, 2023. https://arxiv.org/abs/2304.10049

[17] Wolfe, R., & Caliskan, A. *Contrastive Visual Semantic Pretraining Magnifies the Effect of Social Role and Action Semantics*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022.

[18] Yao, T., Pang, Y., Pan, Y., Zha, Z., & Mei, T. *ActionCLIP: A new paradigm for video action recognition* , 2021

[19] Cafagna, M., van Deemter, K., & Gatt, A., *What Vision-Language Models 'See' when they See Scenes*, arXiv preprint arXiv:2109.07301, 2021. https://arxiv.org/abs/2109.07301

[20] Zilliz, *What is Embedding Drift and How Do I Detect It?*, Zilliz.com, 12 Jan. 2025. https://zilliz.com/ai-faq/what-is-embedding-drift-and-how-do-i-detect-it