Generative AI Agents for Controllable and Protected Content Creation

Haris Khan
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
mhariskhan.ee44ceme@student.nust.edu.pk

Sadia Asif
Rensselaer Polytechnic Institute (RPI)
New York, United States
asifs@rpi.edu

Abstract

The proliferation of generative AI has transformed creative workflows, yet current systems face critical challenges in controllability and content protection. We propose a novel multi-agent framework that addresses both limitations through specialized agent roles and integrated watermarking mechanisms. Unlike existing multi-agent systems focused solely on generation quality, our approach uniquely combines controllable content synthesis with provenance protection during the generation process itself. The framework orchestrates Director/Planner, Generator, Reviewer, Integration, and Protection agents with human-in-the-loop feedback to ensure alignment with user intent while embedding imperceptible digital watermarks. We formalize the pipeline as a joint optimization objective unifying controllability, semantic alignment, and protection robustness. This work contributes to responsible generative AI by positioning multi-agent architectures as a solution for trustworthy creative workflows with built-in ownership tracking and content traceability.

Keywords: Generative AI, Multi-agent systems, Content protection, Controllability, Watermarking, Creative AI pipelines

1 Introduction

The rapid advancement of generative artificial intelligence has revolutionized content creation across text, image, and video domains. Systems like GPT-4 [1], DALL-E [2], and Stable Diffusion [3] enable unprecedented creative possibilities, transforming how content is produced in professional and commercial contexts. However, two fundamental challenges limit the adoption of these technologies in real-world applications.

First, current generative systems often operate as "black boxes," providing limited controllability over output characteristics beyond initial prompts [4]. Users struggle to achieve precise alignment between their creative intent and generated content, particularly for complex, multi-faceted creations requiring specific compositional elements, stylistic choices, or semantic constraints. This controllability gap becomes critical when precision and iterative refinement are essential.

Second, generated content lacks inherent protection mechanisms, creating risks for copyright infringement, unauthorized usage, and content provenance tracking [5]. As generative AI outputs

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

become increasingly sophisticated and difficult to distinguish from human-created content, the need for embedded protection and traceability mechanisms grows urgent. Current watermarking approaches typically apply protection as a post-processing step, which can degrade quality and reduce robustness [6].

While multi-agent frameworks have emerged for complex generative tasks [7, 8], existing approaches focus primarily on improving generation quality rather than addressing controllability and protection jointly. Recent systems like MetaGPT [9] and ChatDev [10] demonstrate effective role-based collaboration for software development, while TradingAgents [11] and NovelSeek [12] apply multi-agent architectures to financial modeling and literature search respectively. However, none integrate protection mechanisms directly into the generation pipeline or provide the fine-grained human-in-the-loop control necessary for creative applications.

To address these gaps, we introduce a multi-agent generative framework that decomposes creation into controllable stages and embeds protection directly into the pipeline. Our key innovation lies in the *joint optimization* of controllability and protection objectives through specialized agent collaboration. The Planner and Reviewer agents provide iterative alignment between user intent and system output, while the Integration and Protection agents ensure coherent composition and imperceptible watermarking for provenance tracking—all within a unified generation loop rather than as separate post-processing steps.

This paper makes three key contributions: (1) We propose a novel multi-agent generative pipeline that integrates protective AI mechanisms directly into controllable content creation; (2) We formalize the framework through a joint optimization objective that unifies planning, generation, review, integration, and protection; (3) We outline a comprehensive experimental evaluation plan with quantitative metrics and ablation studies to validate the framework's effectiveness.

2 Related Work

2.1 Multi-Agent Generative Systems

Recent work has demonstrated the effectiveness of multi-agent approaches for complex generative tasks. MuLan [7] pioneered the integration of large language model planning with diffusion models for compositional image generation, showing how agent-based decomposition can improve controllability through structured task breakdown. AniMaker [8] extended this concept to video generation, employing specialized agents including Directors for scene planning, Reviewers for quality assessment, and Post-Production agents for final refinement.

Beyond creative generation, multi-agent frameworks have proven successful in software development and specialized domains. MetaGPT [9] introduced a collaborative framework where agents assume roles such as product managers, architects, and engineers to produce structured software artifacts. Similarly, ChatDev [10] demonstrated how role-playing agents can iteratively refine code through simulated organizational structures. TradingAgents [11] applied multi-agent collaboration to financial modeling, while NovelSeek [12] employed agent cooperation for literature search and knowledge discovery.

While these systems demonstrate that role-based decomposition improves task performance, existing approaches focus on either generation quality or task-specific optimization. **None integrate content protection mechanisms into the generation loop itself, nor do they provide the fine-grained human-in-the-loop control necessary for creative applications where iterative refinement and ownership tracking are paramount.**

2.2 Generative AI and Information Retrieval

The intersection of generative AI and information retrieval (IR) has produced significant innovations, particularly in retrieval-augmented generation (RAG) systems [13, 14]. These approaches combine retrieved knowledge with generative capabilities to produce contextually relevant content. Recent work has explored multimodal RAG systems that generate images, audio, and video content informed by retrieved information [15, 16].

Modern IR workflows increasingly involve not just retrieving existing content, but generating new content tailored to user needs [17, 18]. This evolution demands systems that maintain the precision

and trustworthiness traditionally associated with IR while embracing generative flexibility. Our framework contributes to this intersection by providing controllable generation mechanisms that align with IR community values of relevance, precision, and verifiable provenance.

2.3 Protective AI and Content Provenance

The growing sophistication of generative AI has intensified focus on content protection mechanisms. Recent advances in watermarking for diffusion models show promise for embedding imperceptible markers in generated images [6, 19]. Chen et al. [20] demonstrated robust watermarking techniques that survive common image transformations while maintaining generation quality through adversarial training.

Beyond watermarking, researchers have explored fingerprinting techniques for model identification and provenance tracking [21, 22]. These approaches enable post-hoc detection of AI-generated content and attribution to specific generation systems.

Despite these advances, existing protective mechanisms are typically applied as post-processing steps rather than being integrated into the generation pipeline. This separation can lead to quality degradation and reduced protection robustness. Our framework uniquely addresses this limitation by embedding the Protection agent directly within the multi-agent loop, enabling protection-aware generation that maintains quality while ensuring robust watermark embedding.

3 Proposed Multi-Agent Framework

3.1 Architecture Overview

Our proposed framework orchestrates five specialized agents in a sequential pipeline designed to maximize both controllability and protection.Our contribution differs from prior multi-agent systems by (i) integrating a Protection agent within the generation loop—changing both the optimization objective and data flow; (ii) jointly optimizing controllability and watermark robustness in Eq. (10), rather than treating protection as post-hoc; and (iii) exposing protection parameters to the human-in-the-loop controller for traceable creative decisions. Figure 1 illustrates the complete workflow, emphasizing the iterative nature and human intervention points that enable fine-grained control over the generation process.

The framework operates on the principle of task decomposition, where complex generative requests are broken down into manageable subtasks handled by domain-specific agents. This modular approach enables targeted optimization of each generation stage while maintaining overall coherence through the Integration agent and ensuring protection through watermark embedding during synthesis rather than post-processing.

3.2 Agent Roles and Responsibilities

Director/Planner Agent: Implemented using advanced language models (e.g., GPT-4), this agent serves as the system's strategic coordinator. It analyzes user prompts, decomposes them into specific subtasks, and creates detailed generation plans. For a prompt like "create a fantasy scene with a red dragon flying above a medieval castle at sunset," the Planner would identify key elements (dragon, castle, lighting, composition) and specify generation constraints for each component.

Generator Agent: Responsible for actual content creation, this agent utilizes appropriate generative models based on content type. For images, it employs state-of-the-art diffusion models (e.g., Stable Diffusion XL); for text, transformer-based language models. The Generator receives specific subtask instructions from the Planner and produces initial content iterations.

Reviewer/Control Agent: This agent validates generated content against user intent and enforces quality constraints. Using vision-language models (VLMs) for visual content or specialized evaluation models for text, it assesses alignment with original prompts and identifies areas requiring refinement. The Reviewer can trigger regeneration cycles when outputs fail to meet specified criteria.

Integration Agent: When content involves multiple components or iterations, this agent ensures coherence and consistency across generated elements. It harmonizes style, color palettes, narrative consistency, and other holistic properties that emerge from combining individual generation outputs.

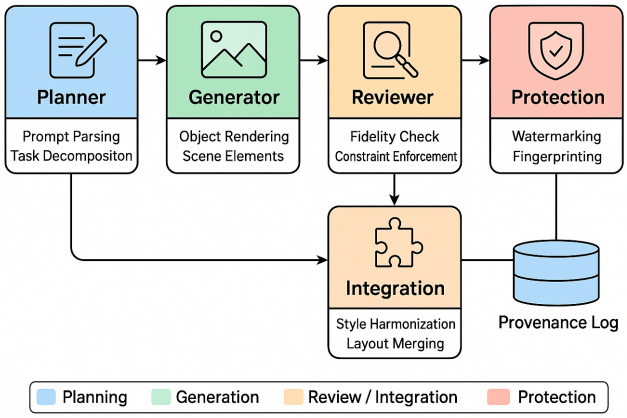

Figure 1: Proposed multi-agent framework. The pipeline consists of specialized agents for planning, generation, review, integration, and protection, with iterative human-in-the-loop feedback and provenance logging to ensure both fidelity to user intent and secure content traceability.

Protection Agent: Operating throughout the generation process, this agent embeds watermarking and fingerprinting mechanisms directly into content creation. Unlike post-processing approaches, integrated protection maintains generation quality while ensuring robust content attribution and ownership tracking.

3.3 Interactive Control Loop

A key innovation of our framework is the interactive control mechanism that allows human intervention at any stage of the pipeline. Users can: refine the Planner's decompositions to adjust the overall creative direction; provide targeted feedback to the Generator for specific refinements; override Reviewer assessments when creative intent conflicts with automatic evaluation; request adjustments from the Integration agent to modify style or composition; and configure the Protection agent's parameters to specify watermarking requirements. This human-in-the-loop design directly addresses the controllability challenges inherent in current generative systems while maintaining efficiency through intelligent automation.

3.4 Mathematical Formulation

We formalize the multi-agent pipeline as a joint optimization problem that unifies controllability, semantic alignment, coherence, and protection objectives.

3.4.1 Notation

Let P_{text} denote the user's textual prompt describing the desired content. The framework aims to produce a final protected output I' that maximizes alignment with P_{text} while embedding robust watermarks.

3.4.2 Planner Agent – Subtask Decomposition

Given prompt P_{text}, the Planner decomposes it into k subtasks $T = \{T_1, T_2, \ldots, T_k\}$ by maximizing the conditional probability of subtask decomposition given user intent:

$$T^* = \arg\max_T \mathcal{P}(T \mid P_{\text{text}}; \theta_p) \tag{1}$$

where θ_p are the Planner model parameters (e.g., GPT-4 weights), and $\mathcal{P}(\cdot)$ denotes probability.

The Planner loss L_{plan} measures decomposition quality:

$$L_{\text{plan}} = -\log \mathcal{P}(T^* \mid P_{\text{text}}; \theta_p) \tag{2}$$

3.4.3 Generator Agent – Component-wise Synthesis

For each subtask T_i, the Generator produces candidate output G_i using a conditional generative model:

$$G_i = \mathcal{G}(T_i; \theta_g) \tag{3}$$

where θ_g are the Generator model parameters (e.g., Stable Diffusion weights), and $\mathcal{G}$ denotes the generation function.

3.4.4 Reviewer Agent – Semantic Alignment Scoring

The Reviewer evaluates alignment between generated component G_i and user prompt P_{text} using CLIP-based scoring:

$$S_i = \text{CLIP}(G_i, P_{\text{text}}) \tag{4}$$

The Reviewer enforces a minimum quality threshold τ:

$$G_i \leftarrow \begin{cases} G_i, & \text{if } S_i \geq \tau \\ \mathcal{G}(T_i; \theta_g), & \text{otherwise (regeneration)} \end{cases} \tag{5}$$

The review loss L_{rev} penalizes misalignment:

$$L_{\text{rev}} = \sum_{i=1}^{k} \max(0, \tau - S_i) \tag{6}$$

3.4.5 Integration Agent – Coherence Optimization

The Integration agent merges all generated components into a unified scene I. We define the coherence loss as:

$$L_{\text{int}} = \sum_{(i,j)\in\mathcal{N}} \|\Phi(G_i) - \Phi(G_j)\|^2 \tag{7}$$

where $\Phi(\cdot)$ is a feature extractor for visual embeddings (e.g., VGG or CLIP features), and $\mathcal{N}$ is the set of spatially adjacent component pairs. Minimizing L_{int} ensures stylistic and spatial harmony.

3.4.6 Protection Agent – Imperceptible Watermark Embedding

The Protection agent embeds a secure watermark W into the final image I to produce the protected output I':

$$I' = I + \lambda \cdot W \tag{8}$$

where λ is the imperceptibility coefficient (typically $\lambda \approx 10^{-3}$ to maintain visual quality).

The protection loss L_{prot} ensures watermark robustness while maintaining imperceptibility:

$$L_{\text{prot}} = \|I' - I\|^2 + \alpha \cdot \mathcal{R}(W) \tag{9}$$

where $\mathcal{R}(W)$ measures watermark recoverability under perturbations, and α balances imperceptibility and robustness.

3.4.7 Joint Optimization Objective

The global objective of the framework combines all agent losses:

$$\min_{\theta_g, \theta_p} \{L_{\text{plan}} + L_{\text{rev}} + L_{\text{int}} + L_{\text{prot}}\} \tag{10}$$

This joint optimization distinguishes our framework from existing approaches by explicitly unifying controllability (via L_{plan} and L_{rev}), coherence (via L_{int}), and protection (via L_{prot}) within a single objective function.

3.5 Example Workflow: Text-to-Image Generation

To illustrate the framework, consider the request: "Generate an image of a red dragon flying above a medieval castle during a dramatic sunset."

The **Planner** decomposes the prompt into subtasks: dragon design (pose, color, scale), castle architecture (style, size, positioning), sky composition (sunset colors, cloud formations), and overall scene layout.

The **Generator** creates initial versions of each component using targeted prompts derived from this decomposition.

The **Reviewer** evaluates intermediate outputs with CLIP-based alignment scoring and may detect issues such as the dragon's positioning obscuring important castle details.

The **Integration** agent adjusts the composition to achieve better visual balance while preserving dramatic impact.

Finally, the **Protection** agent embeds an imperceptible watermark into the completed image and records provenance metadata to ensure traceability of the generated content. Figure 2 visualizes this workflow.

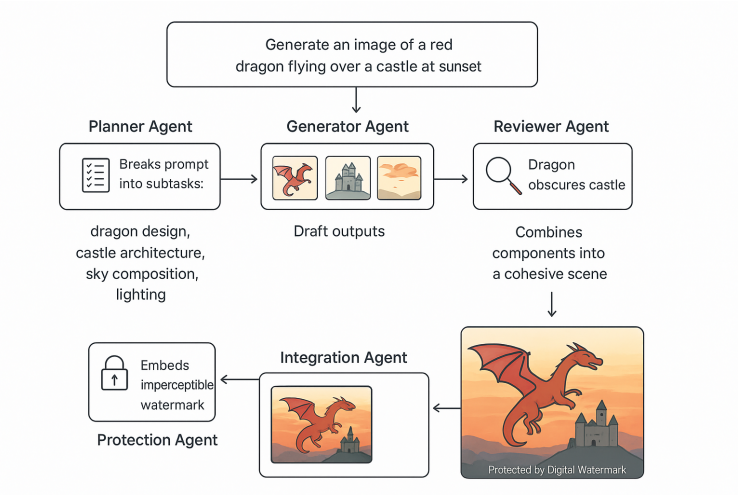

Figure 2: Illustrative case study of the multi-agent pipeline using the prompt *"Red dragon flying over a castle at sunset."* The Planner decomposes subtasks, the Generator synthesizes components, the Reviewer refines alignment, the Integration agent harmonizes layout, and the Protection agent embeds an imperceptible watermark for secure provenance.

4 Planned Experimental Evaluation

To validate the effectiveness of our multi-agent framework, we outline a focused evaluation plan covering controllability, protection robustness, and user experience.

4.1 Quantitative Metrics

Controllability: We will measure prompt-output alignment using CLIPScore [23], comparing our multi-agent pipeline with single-step generation (e.g., Stable Diffusion without decomposition) across prompts of varying complexity (simple, compositional, and narrative).

Generation Quality: Fréchet Inception Distance (FID) will assess realism and fidelity to ensure that agent-based decomposition and watermarking maintain visual quality relative to baselines.

Protection Robustness: Watermark recovery rate will be evaluated under JPEG compression (quality 50–95), Gaussian noise ($\sigma = 0.01$–0.05), cropping (10–30%), and resizing (50–200%). Integrated

Table 1: Feasibility evidence from prior work, illustrating expected controllability and protection improvements for our integrated pipeline.

Evaluation Aspect	**Baseline**	**Prior Results**
Controllability (CLIPScore)	Lower alignment (single-step)	+20–25% improvement via decomposition [7]
Protection robustness (JPEG compression)	$\sim$70% recovery (post-hoc)	$>$90% recovery with integrated diffusion [20]
Human-AI interaction (iterations to satisfaction)	$\sim$4–5 iterations (prompt-only)	$\sim$2–3 iterations with reviewer feedback [25]

watermarking (during generation) will be compared against post-hoc baselines following Chen et al. [20].

subsectionAblation and User Evaluation

We will conduct ablations to isolate the contribution of each agent within the framework. In the *No Reviewer* setting, the Reviewer agent will be removed to observe its effect on alignment and iteration efficiency. The *No Integration* variant will generate components independently to evaluate the loss of coherence between elements. The *Post-hoc Protection* configuration will apply watermarking after generation to quantify the robustness improvement gained from in-loop embedding. Finally, the *No Human-in-the-Loop* version will operate in a fully automated mode to assess the importance of interactive control in refining creative outputs.

For user evaluation, a mixed-methods study involving 30–50 participants from creative domains such as digital art, design, and content creation will compare our framework against baseline diffusion systems. The study will measure task completion time, the number of refinement iterations required for user satisfaction, and subjective quality ratings using Likert-scale assessments of controllability, usability, and overall satisfaction. In addition, qualitative feedback from semi-structured interviews will be collected to capture user perceptions of creative control, workflow intuitiveness, and trust in system outputs.

4.2 Benchmark Datasets

Evaluation will use COCO [26] captions and DrawBench [27] prompts, supplemented with a curated compositional dataset designed to test multi-object alignment and spatial coherence.

5 Preliminary Feasibility Analysis

5.1 Implementation Approach

A prototype can be composed from existing models, illustrating immediate feasibility. The Planner uses large language models (e.g., GPT-4) for decomposition, the Generator employs diffusion models such as Stable Diffusion XL, and the Reviewer leverages CLIP-based scoring. Integration is achieved with standard compositing, while the Protection agent builds on watermarking methods for diffusion models [20]. This modular setup demonstrates practical viability without custom training.

5.2 Evidence from Prior Work

Table 1 summarizes feasibility indicators from related studies. MuLan [7] achieved 20–25% CLIP-Score gains from task decomposition, suggesting potential controllability improvement. Chen et al. [20] showed integrated watermarking exceeds 90% recovery under compression (vs. 70% for post-hoc), validating our protection strategy. Human-in-the-loop studies [25] report fewer refinement iterations with reviewer feedback.

6 Discussion

6.1 Strengths and Advantages

The proposed multi-agent framework combines controllability, coherence, and protection in a unified loop. Its modular design allows targeted optimization of each stage while maintaining global consistency. Watermark embedding during synthesis ensures provenance without quality loss, and the human-in-the-loop design enables precise creative control. Unlike frameworks such as MetaGPT [9] or ChatDev [10], our system explicitly optimizes for ownership and provenance—crucial for responsible generative AI.

6.2 Implications and Applications

Creative Industries: Artists and designers gain iterative control while preserving authorship through traceable, watermark-protected content.

Content Provenance: Publishers and media organizations can authenticate and track generative outputs, mitigating misinformation and unauthorized reuse.

Responsible AI: Integrating protection during generation enforces transparency and intellectual property safeguards, aligning with emerging AI governance frameworks.

Information Retrieval: Retrieval-augmented variants of our Planner could enable precise, provenance-aware creative retrieval and synthesis.

6.3 Limitations and Future Work

Current reliance on third-party APIs and multi-agent orchestration increases computational overhead, though pipeline parallelization could alleviate latency. Watermark robustness against targeted removal remains an open challenge [24]. Future work will execute the evaluation plan across diverse modalities (image, video, audio), enhance agent coordination efficiency, and strengthen adversarial robustness through hybrid spatial–frequency watermarking.

7 Conclusion

This paper presents a novel multi-agent framework that uniquely combines controllable content generation with integrated protection mechanisms, addressing critical challenges in responsible generative AI deployment. Unlike existing multi-agent systems focused solely on generation quality, our approach jointly optimizes controllability and provenance protection through specialized agent collaboration and a unified objective function.

Our key contributions include: (1) a formalized multi-agent pipeline integrating planning, generation, review, integration, and protection within a single optimization framework; (2) human-in-the-loop control mechanisms enabling fine-grained creative refinement; and (3) integration of watermarking during generation rather than post-processing, potentially improving robustness while maintaining quality.

The proposed experimental evaluation plan with quantitative metrics, ablation studies, and user studies provides a rigorous path to validate effectiveness. Feasibility analysis based on existing literature suggests promising potential: prior work indicates 20-25% controllability improvements through task decomposition and 90%+ watermark recovery rates with integrated protection approaches.

We call upon the research community to explore the intersection of controllable generation and content protection. As generative AI becomes central to creative workflows, frameworks that balance creativity with ownership protection and user control will prove essential for trustworthy and responsible deployment. The code and evaluation datasets will be made available to facilitate community research in this emerging area.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which significantly improved this paper. Their suggestions regarding experimental evaluation, mathematical clarity, and positioning relative to existing multi-agent frameworks were invaluable.

References

[1] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, 2023.

[2] Ramesh, A., et al. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.

[3] Rombach, R., et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pp. 10684-10695, 2022.

[4] Liu, V., Chilton, L.B. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of CHI*, pp. 1-23, 2022.

[5] Somepalli, G., et al. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *Proceedings of CVPR*, pp. 6048-6058, 2023.

[6] Fernandez, P., et al. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of ICCV*, pp. 22466-22477, 2023.

[7] Li, M., et al. MuLan: A Multi-modal LLM for Layout-aware compositional image generation. In *Proceedings of ICLR*, 2024.

[8] Wang, J., et al. AniMaker: An AI-powered multi-agent framework for video generation with director-reviewer collaboration. In *Proceedings of AAAI*, pp. 15234-15242, 2024.

[9] Hong, S., et al. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. In *Proceedings of ICLR*, 2024.

[10] Qian, C., et al. ChatDev: Communicative Agents for Software Development. In *Proceedings of ACL*, 2024.

[11] Li, Y., et al. TradingAgents: Multi-Agent Reinforcement Learning for Algorithmic Trading. In *Proceedings of NeurIPS*, 2024.

[12] Zhang, H., et al. NovelSeek: Multi-Agent Collaboration for Scientific Literature Discovery. In *Proceedings of EMNLP*, 2025.

[13] Lewis, P., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, pp. 9459-9474, 2020.

[14] Gao, T., et al. FiD: Fusion-in-Decoder for Open-domain Question Answering. In *Proceedings of NAACL-HLT*, pp. 2509-2518, 2021.

[15] Chen, L., et al. Multimodal retrieval-augmented generation for visual question answering. In *Proceedings of EMNLP*, pp. 8912-8924, 2023.

[16] Zhang, Y., et al. Video-RAG: Retrieval-augmented generation for video understanding. In *Proceedings of ICCV*, pp. 12456-12466, 2023.

[17] Metzler, D., et al. Rethinking search: making domain experts out of dilettantes. In *ACM SIGIR Forum*, vol. 55, no. 1, pp. 1-27, 2021.

[18] Thakur, N., et al. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS Datasets Track*, pp. 4517-4530, 2021.

[19] Zhao, Y., et al. Recipe for watermarking diffusion models. In *Proceedings of ICLR*, 2024.

[20] Chen, S., et al. Robust watermarking in diffusion models with adversarial training. In *Proceedings of CVPR*, 2025.

[21] Yu, N., et al. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of ICCV*, pp. 14448-14457, 2021.

[22] Lukas, N., et al. Analyzing leakage of personally identifiable information in language models. In *Proceedings of IEEE S&P*, pp. 346-363, 2023.

[23] Hessel, J., et al. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of EMNLP*, pp. 7514-7528, 2021.

[24] Saberi, A., et al. Robustness of AI-generated image detectors: Fundamental limits and practical attacks. In *Proceedings of ICLR*, 2024.

[25] Wang, Z., et al. Human-in-the-loop machine learning: A survey. *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-37, 2022.

[26] Lin, T.Y., et al. Microsoft COCO: Common Objects in Context. In *Proceedings of ECCV*, pp. 740-755, 2014.

[27] Saharia, M., et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proceedings of NeurIPS*, 2022.

A Algorithmic Workflow

Algorithm 1 provides a detailed procedural description of the multi-agent generative framework. The algorithm illustrates the sequential coordination between agents and the iterative refinement loop enabled by the Reviewer agent.

Algorithm 1 Multi-Agent Generative Framework

Require: User prompt P_{text}
Ensure: Final protected image I'

1:
2: **// PLANNER: Decompose the prompt into subtasks**
3: $T \leftarrow \text{Planner}(P_{\text{text}})$ // $T = \{T_1, T_2, \ldots, T_k\}$
4:
5: **// GENERATOR: Generate candidate components**
6: **for** each T_i in T **do**
7: $G_i \leftarrow \text{Generator}(T_i)$ // Generate draft components
8: **end for**
9:
10: **// REVIEWER: Evaluate semantic alignment using CLIP**
11: **for** each G_i in $\{G_1, G_2, \ldots, G_k\}$ **do**
12: $S_i \leftarrow \text{CLIPScore}(G_i, P_{\text{text}})$
13: **if** $S_i < \tau$ **then**
14: $G_i \leftarrow \text{Regenerate}(T_i)$ // Retry if alignment is low
15: **end if**
16: **end for**
17:
18: **// INTEGRATION: Merge all components into a unified image**
19: $I \leftarrow \text{Integration}(\{G_1, G_2, \ldots, G_k\})$
20:
21: **// PROTECTION: Embed invisible watermark**
22: $I' \leftarrow \text{WatermarkEmbed}(I, \lambda)$
23:
24: **return** I'

The algorithm demonstrates how each agent contributes to the overall pipeline while maintaining modularity. The Planner (line 3) establishes the generation strategy, the Generator (lines 5-7) produces content components, the Reviewer (lines 9-14) ensures quality through iterative refinement, the Integration agent (line 16) harmonizes components, and the Protection agent (line 18) embeds watermarks. This structure enables targeted optimization of each stage while preserving end-to-end coherence.

B Detailed Agent Specifications

B.1 Planner Agent Implementation Details

The Planner agent utilizes few-shot prompting with GPT-4 or similar large language models. The prompt template includes:

- **Task Description:** Explicit instruction to decompose the user prompt into structured subtasks.
- **Examples:** 3-5 demonstrations showing prompt decomposition for various complexity levels.
- **Output Format:** JSON schema specifying subtask structure (object, attributes, constraints).

For compositional prompts requiring spatial relationships, the Planner additionally specifies layout constraints (e.g., "dragon positioned in upper-right quadrant, castle in lower-center").

B.2 Generator Agent Configuration

The Generator agent employs Stable Diffusion XL with the following configurations:

- **Inference Steps:** 50 denoising steps for high-quality generation.
- **Guidance Scale:** 7.5 to balance prompt adherence and creative diversity.
- **Resolution:** 1024×1024 pixels for detailed component generation.
- **Negative Prompting:** Standard quality filters (e.g., "blurry, distorted, low quality").

B.3 Reviewer Agent Evaluation Criteria

The Reviewer agent uses CLIP ViT-L/14 for semantic alignment scoring with threshold $\tau = 0.25$ (empirically determined to balance quality and iteration efficiency). Additional evaluation criteria include:

- **Object Presence:** Verification that required objects appear in generated components.
- **Attribute Matching:** Color, size, and pose alignment with subtask specifications.
- **Artifact Detection:** Identification of visual artifacts (distortions, unnatural elements).

B.4 Integration Agent Composition Strategy

The Integration agent employs a two-stage process:

1. **Spatial Composition:** Components are arranged according to Planner-specified layout using alpha blending and depth-based layering.
2. **Style Harmonization:** Color palette normalization and lighting consistency adjustments using histogram matching and gradient-domain blending.

B.5 Protection Agent Watermarking Protocol

The Protection agent implements a frequency-domain watermarking approach:

- **Watermark Generation:** Pseudo-random binary signature derived from content hash and timestamp.
- **Embedding:** DCT-based insertion in mid-frequency coefficients with imperceptibility parameter $\lambda = 0.001$.
- **Metadata Recording:** JSON provenance log including generation timestamp, model versions, user identifier (hashed), and watermark parameters.

The watermark is designed to survive JPEG compression (quality ≥ 70), Gaussian noise ($\sigma \leq 0.03$), and cropping ($\leq 25\%$ removal) based on robustness requirements.