ISSN 2618-1568 Vol. 7, Issue 5: 48-52, DOI: 10.25236/FAR.2025.070507

Multi-Agent Systems for Collaborative Art Creation

Yi Luo

The Baldwin School, 701 Montgomery Avenue Bryn Mawr, PA, 19010, US
yluo@baldwinschool.org

Abstract: *The field of AI-driven art and content generation has witnessed significant advancements. However, existing models typically operate as monolithic systems, lacking the nuanced colaboration, cross-domain capabilities, and iterative refinement necessary to meet complex user needs. This paper introduces the Multi-Agent System for Collaborative creation (MASC) framework, a novel approach integrating the collaborative, routing, and evaluative strengths of Multi-Agent Systems (MAS) with state-of-the-art generative models, such as Denoising Diffusion Probabilistic Models (DDPMs) and Diffusion Transformers (DiTs) for image synthesis. MASC leverages a team of specialized agents/modules (Task Agent, Domain Analysis Agent, Deepthink Module, Generator Module, Reflection Module) interacting through defined protocols and iterative loops. This enables the system to understand user intent, enrich creative prompts, generate multi-domain content (text, images, video, etc.), and perform evaluation-driven optimization. We elaborate on the MASC architecture, component roles, communication mechanisms, and its integration with core generative processes. Experimental results suggest MASC holds advantages in enhancing the detail richness and user intent alignment of generated content, opening new avenues for exploring more complex, cross-domain, and conceptually driven AI content creation.*

Keywords: *Multi-Agent, Artistic Creation, Diffusion Models, Large Language Models*

1. Introduction

In recent years, the domain of AI-driven art generation has undergone rapid development. Early methods like Generative Adversarial Networks (GANs) [1] and Variational Autoencoders (VAEs) [4] laid the groundwork for the field. Currently, diffusion models [2] have emerged as the state-of-the-art, particularly Denoising Diffusion Probabilistic Models (DDPMs), which exhibit remarkable capabilities in generating high-fidelity and diverse images. Recent architectural innovations, such as Diffusion Transformers (DiTs) [6], have further boosted model scalability and performance, rivaling other generative paradigms.

However, despite the power of these generative models, they often function as single, monolithic systems focused on a specific domain (e.g., image generation). This paradigm lacks the nuanced collaboration, negotiation, and iterative refinement characteristic of human creative processes, especially in team or studio settings [3]. Furthermore, extending them directly to handle complex, coordinated tasks across multiple modalities (text, image, video, etc.) remains challenging. Multi-Agent Systems (MAS) offer a paradigm for simulating such interaction and coordination. In MAS, multiple agents with specialized roles can collaborate towards a common goal.

Consequently, we introduce the MASC framework. It utilizes a team of AI agents and modules (e.g., Task Agent, Domain Analysis Agent, Deepthink Module, Generator Module, Reflection Module). These components interact to route tasks, understand and enrich user instructions, execute generation using appropriate models, and finally, perform iterative optimization through reflective evaluation. While a single generative model primarily generates based on direct input prompts, MASC can decompose complex creative objectives for specialized agents, enrich details via "deep thinking," and iteratively refine through "reflection." This mimics human collaboration and revision processes [3] and can potentially yield outputs that better reflect complex intentions or emergent concepts.

2. Related Work

Diffusion Models (DDPM): The core mechanism of diffusion models involves a forward noising process and a learned reverse denoising process. By gradually adding Gaussian noise to data and then

ISSN 2618-1568 Vol. 7, Issue 5: 48-52, DOI: 10.25236/FAR.2025.070507

learning to remove it in the reverse process, the work by Ho et al. (2020) [2] was seminal in this area.

Diffusion Transformers (DiT): DiT [6] introduces a scalable, Transformer-based backbone for diffusion models. Its architecture processes latent representations by patchifying them and using a series of Transformer blocks on these tokens. Conditional information (like timesteps, class labels) is incorporated using methods like adaptive layer normalization (adaLN). Compared to traditional U-Net structures, DiT demonstrates superior scalability and achieves excellent performance on image generation tasks.

Other Generative Models (GANs, VAEs): Generative Adversarial Networks (GANs) [1] generate realistic images through adversarial training between a generator and a discriminator but face challenges like mode collapse and training instability. Variational Autoencoders (VAEs) [4] are probabilistic generative models that learn latent representations via an encoder and decoder. They facilitate latent space manipulation but sometimes produce blurrier images. These models represent foundational work in generative art.

MAS Concepts: A Multi-Agent System (MAS) consists of multiple interacting autonomous agents [7]. Agents possess characteristics like autonomy, local perspective, and decentralization. Interactions between agents can be cooperative, competitive, or mixed .

MAS for Art and Creativity: Some frameworks explicitly apply MAS to artistic creation, for example, using artist and critic agents that interact. Many recent MAS systems leverage Large Language Models (LLMs) to drive agent reasoning and communication capabilities [5].

3. Method

3.1 System Architecture Overview

The proposed MASC framework coordinates a set of specialized AI agents and modules to handle cross-domain content generation tasks, leveraging iterative optimization to enhance the quality and user satisfaction of the final output. The core idea is to utilize a multi-agent system for task routing, deep instruction understanding, prompt enrichment, and reflective iteration to guide various generative models.

The process begins with user input, typically containing the target domain (e.g., image generation) and specific instructions. This input is first received by the Task Agent, which acts as the system's entry point and dispatcher. The Task Agent identifies the domain required for the task and initiates the corresponding processing pipeline, assigning the task to the specialized agent for that domain (e.g., ArtAgent).

The Specialized Agent is responsible for deeply analyzing the user's instructions, understanding specific requirements and constraints. The analysis result is then passed to the Deepthink Module. This module's core function is to expand and enrich the user's original prompt, adding more detail to make it more vivid, descriptive, and concrete, thereby providing more effective guidance for the generator.

The enriched prompt is sent to the corresponding Generator Module (e.g., a DDPM/DiT-based Image Generator). This module performs the actual content synthesis task, generating an initial version of the output (text, image, or video).

The generated output is then submitted to the Reflection Module for evaluation. This module assesses the output based on the original user intent, the requirements by the Deepthink Module, and potentially general quality standards. If the output meets the requirements, the process concludes, yielding the final result. If not, the Reflection Module initiates an iterative loop, triggering regeneration of the content. This process may repeat until a satisfactory result is generated or a preset iteration limit is reached. This architecture simulates complex creative workflows through modularity and iterative feedback, adapting to diverse user needs and content domains.

3.2 Component Definitions

The MASC framework comprises several distinct components, each with a specialized role in the collaborative creation process. The key components are defined as follows:

Task Agent: Acts as the primary interface and central coordinator.

ISSN 2618-1568 Vol. 7, Issue 5: 48-52, DOI: 10.25236/FAR.2025.070507

Function: Receives the initial, often high-level, user request. It identifies the appropriate core task type (e.g., video generation) and the target domain. It then initiates the workflow by dispatching the task and associated instructions to the relevant Specialized Agent.

Key Feature: Provides a unified entry point for diverse creative tasks and enables dynamic pipeline assignment, crucial for the framework's modularity and extensibility.

Specialized Agents (e.g., Article Agent, Image Agent, Video Agent): Domain-specific agents responsible for interpreting user intent within their area of expertise.

Function: Receives the dispatched task from the Task Agent. It performs an in-depth analysis of the user's instructions, extracting key entities, constraints, stylistic requirements, and semantic meaning relevant to its domain. For instance, the ‘Video Agent’ identifies core concepts like "beautiful", "sea", and the "video generation" task itself from the initial request. This structured understanding is then passed to the Deepthink Module.

Implementation: Often leverages Large Language Models (LLMs) for sophisticated natural language understanding capabilities to parse and interpret complex user requirements accurately.

Deepthink Module: The core engine for creative expansion and prompt enrichment.

Function: Takes the structured or analyzed input from the Specialized Agent (e.g., the keywords from the ‘Video Agent’). Its primary role is to significantly elaborate on the initial concept, transforming concise requirements into a rich, detailed, and evocative narrative or description. As illustrated it can expand simple keywords ("beautiful, sea") into a detailed scene description ("The rugged coastline is dotted with dark, weathered rocks..."),

Providing a much more concrete and inspirational blueprint for the generation process. This step aims to bridge the gap between abstract user intent and the detailed input needed by powerful generative models.

Implementation: Typically relies on advanced generative LLMs capable of creative writing, reasoning, and incorporating diverse details (sensory information, context, mood, etc.) to produce high-quality, enriched prompts.

Generator Modules (e.g., Article Generator, Image Generator, Video Generator): The components responsible for the actual content synthesis.

Function: Receives the highly detailed, enriched prompt from the Deepthink Module. It employs a state-of-the-art generative model specifically suited for the target domain (e.g., a dedicated video generation model like the ‘Video Gen’ a DDPM/DiT for images [2, 6], or an LLM for text) to generate the creative output based on the detailed instructions provided by the enriched prompt.

Implementation: Utilizes domain-specific, pre-trained generative models. For instance, the Image Generator might use Stable Diffusion or DiT models [6], while a Video Generator would employ specialized video synthesis architectures.

Reflection Module (Reflect Agent): The quality control and iterative refinement unit.

Function: Receives the generated output from the Generator Module. It evaluates this output against a set of criteria, which can include alignment with the enriched prompt from the Deepthink module, adherence to the original user intent, domain-specific quality metrics (e.g., visual coherence, aesthetic appeal), and potentially safety checks. Based on the evaluation, it decides if the output is satisfactory. If not, it initiates an iterative refinement loop, potentially providing feedback to the Generator (e.g., adjust parameters) or even back to the Deepthink module (e.g., modify prompt) to guide subsequent generation attempts.

Implementation: Can employ a variety of techniques, including Vision-Language Models (VLMs) for assessing image/video content and aesthetics, LLMs for evaluating text or providing textual critique, or specialized metrics for specific quality dimensions. Its decision logic drives the iterative nature of the MASC framework.

ISSN 2618-1568 Vol. 7, Issue 5: 48-52, DOI: 10.25236/FAR.2025.070507

4. Experiments

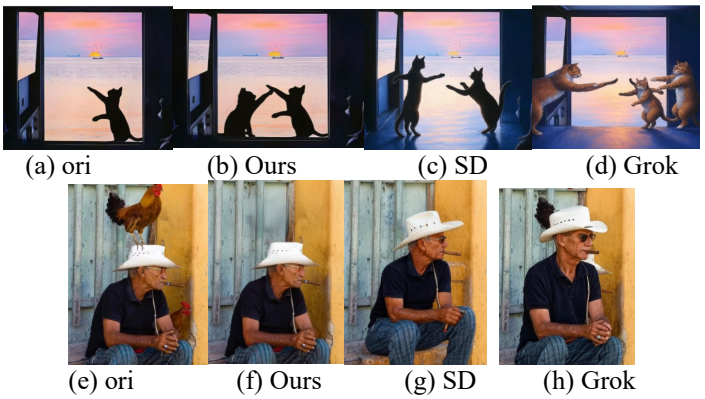

Figure 1 Qualitative comparison for image editing

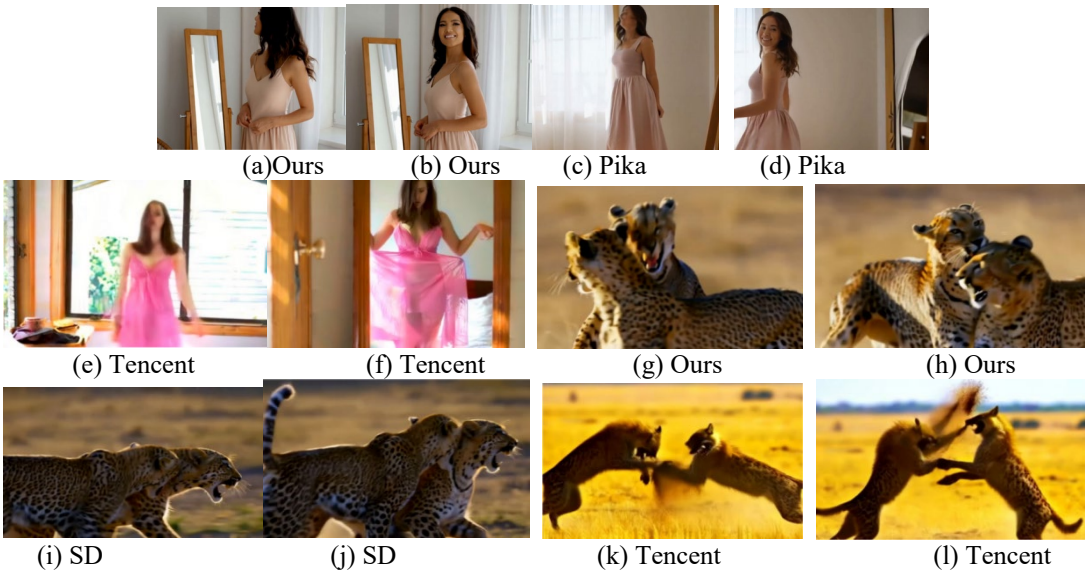

Figure 2 Video Gen

To demonstrate the effectiveness and capabilities of the proposed Multi-Agent System for Collaborative creation (MASC) framework, we conducted a series of experiments focusing on various creative generation tasks. Our evaluation primarily relies on qualitative comparisons between the outputs generated by MASC and those produced by representative baseline or state-of-the-art generative models. These comparisons aim to illustrate MASC's potential advantages in interpreting user intent, enriching creative details through the Deepthink and Reflection mechanisms, and achieving high-quality results across different modalities and tasks, such as image editing, video generation, and style transfer. The following figure 1, 2 and 3 present visual results from these comparative experiments.

ISSN 2618-1568 Vol. 7, Issue 5: 48-52, DOI: 10.25236/FAR.2025.070507

(e) SD (f) SD

Figure 3 Style Transfer

5. Conclusion

We introduced the Multi-Agent System for Collaborative creation (MASC) framework, a novel approach designed to enhance AI-driven content generation by leveraging the principles of multi-agent collaboration and iterative refinement. Addressing the limitations of monolithic generative models, MASC utilizes a synergistic team of specialized agents—including Task, Domain Analysis, Deepthink, Generator, and Reflection modules—to interpret complex user intent, enrich creative concepts, manage multi-modal synthesis, and optimize outputs through evaluation loops.

Our experimental results, showcased through qualitative comparisons across various tasks like image editing, style transfer, and video generation, demonstrate the efficacy of the MASC framework. The system consistently produces high-quality multimedia content that exhibits enhanced detail richness and strong alignment with user objectives, outperforming standard approaches in capturing nuanced requirements.

References

[1] Ian J Goodfellow et al. "Generative adversarial nets". In: Advances in neural information processing systems 27 (2014).
[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: Advances in neural information processing systems 33 (2020), pp. 6840–6851.
[3] Naomi Imasato et al. "Creative Agents: Simulating the Systems Model of Creativity with Generative Agents". In: arXiv preprint arXiv:2411.17065 (2024).
[4] Joon Sung Park et al. "Generative agents: Interactive simulacra of human behavior". In: Proceedings of the 36th annual acm symposium on user interface software and technology. 2023 pp. 1–22.
[5] William Peebles and Saining Xie. "Scalable diffusion models with transformers". In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, pp. 4195–4205.
[6] Michael Wooldridge. An introduction to multiagent systems. John wiley & sons, 2009.