Talk, Imagine, Evolve: A Unified Multimodal Agent for Seamless Visual Generation and Editing

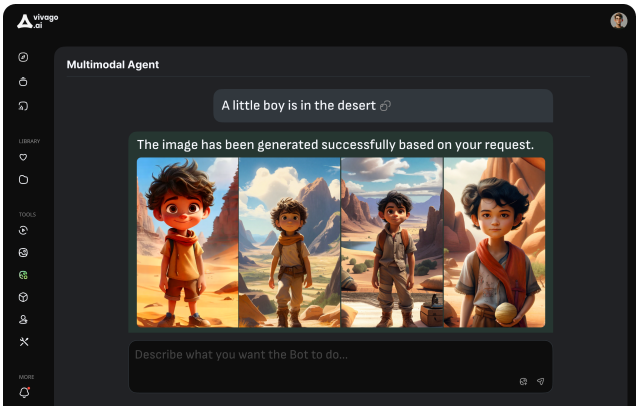
Zhaofan Qiu
HiDream.ai Inc.
Beijing, China
qiuzhaofan@hidream.ai

Zijian Gong
HiDream.ai Inc.
Beijing, China
zijian@hidream.ai

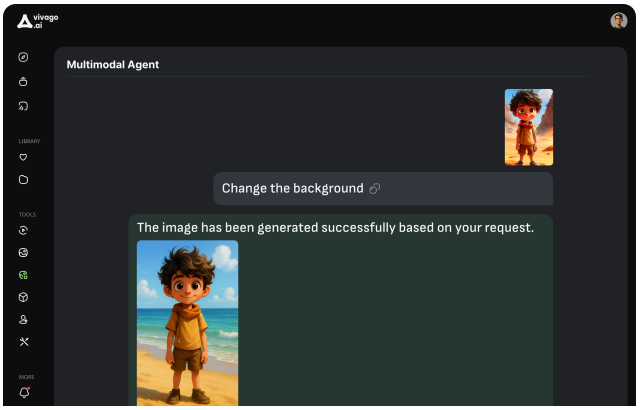
Yingwei Pan
HiDream.ai Inc.
Beijing, China
pandy@hidream.ai

Ting Yao
HiDream.ai Inc.
Beijing, China
tiyao@hidream.ai

Tao Mei
HiDream.ai Inc.
Beijing, China
tmei@hidream.ai



(a) Text-to-image generation.



(b) Instruction-based image editing.

Figure 1: Interface of our multimodal agent (HiDream-Agent).

Abstract

This paper demonstrates a pioneering unified multimodal agent that transforms complex visual content creation into an intuitive, conversational experience, allowing users to talk, imagine, and evolve their ideas. Overcoming the limitations of fragmented multimodal technique tools, our system seamlessly integrates text-to-image generation, instruction-based image editing, text/image-to-video generation, and interactive understanding within a single AI interface. Users of all skill levels can perform sophisticated visual tasks using natural language and visual inputs. The system's architecture features a central Coordinator module processing multimodal inputs and directing tasks to Generation or Chat pathways. For Generation, a Planner utilizes our state-of-the-art specialized models in image/video generation and image editing, while the Chat function facilitates clarification and collaboration. The interactive demonstration will showcase intuitive multimodal input, seamless real-time content creation/editing, dynamic interactive understanding, and a unified workflow. This agent pioneers a new way for accessible, interactive visual storytelling and collaborative content creation in multimodal generative AI.

ACM Reference Format:
Zhaofan Qiu, Zijian Gong, Yingwei Pan, Ting Yao, and Tao Mei. 2025. Talk, Imagine, Evolve: A Unified Multimodal Agent for Seamless Visual Generation and Editing. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3746027.3754467

1 Introduction

The creation and manipulation of multimedia content have become increasingly central to digital communication, entertainment, and professional workflows. However, the techniques/tools available for these multimodal generative tasks often operate in silos, forcing users to master multiple complex models/products for image/video generation and instruction-based image editing. This fragmentation creates a significant barrier, particularly for non-expert users, and hinders a fluid, intuitive creative process where ideas can be seamlessly translated into visual outputs. There is a growing need for systems that can democratize sophisticated visual content creation, making it as natural as describing one's vision.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754467

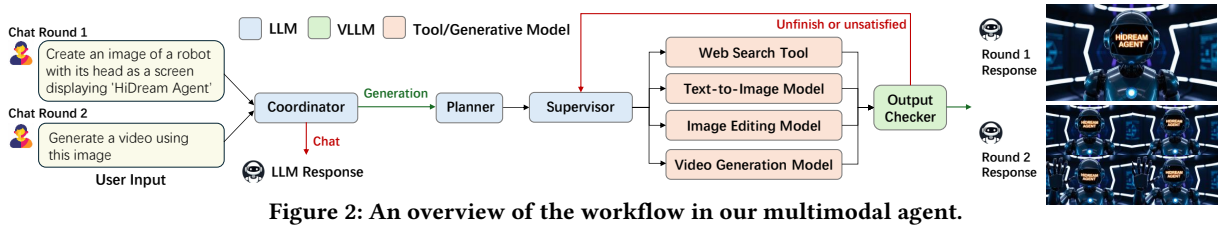


Figure 2: An overview of the workflow in our multimodal agent.

Recent advancements in generative artificial intelligence have yielded powerful diffusion models capable of synthesizing multimodal content with remarkable fidelity [1, 3, 4, 8–12]. Concurrently, the field of AI agents has seen substantial progress, with systems evolving from task-specific automation to more general-purpose assistants capable of understanding complex instructions, planning multi-step actions, and interacting with various tools and environments (e.g., the recent general-purpose AI agent Manus). While these advancements are transformative, a significant opportunity remains in unifying diverse visual content creation modalities—spanning generation (both image and video) and intricate editing—within a single, coherent conversational interface. Current state-of-the-art general-purpose agents, while powerful, may not be explicitly optimized for the iterative and often nuanced dialogue required in creative visual workflows tailored to multimodal field.

In this demonstration, we present a unified multimodal agent (namely HiDream-Agent) designed to bridge this gap. Our system empowers users to perform complex visual content creation and manipulation tasks—including text-to-image generation, instruction-based image editing, and video generation—through natural language dialogue, augmented by optional visual inputs. By integrating a sophisticated Coordinator module to interpret user intent, a Planner module to plan the overall execution process, a Supervisor module to execute specialized generative models, and an Output Checker to verify whether the generated tasks are completed, our agent offers a seamless and intuitive multimedia content creation experience. This demonstration will showcase the agent's ability to interpret multimodal instructions, generate and edit visual content based on dialogue, and engage in context-aware conversations, thereby lowering the barrier to sophisticated visual authoring and fostering a more collaborative human-AI creative process.

2 Technology

Figure 2 shows the workflow of our unified multimodal agent, which is composed of three multimodal generative models (text-to-image generation, instruction-based image editing, and text/image-to-video generation) and several key modules (e.g., Coordinator, Planner, Supervisor, and Output Checker) for scheduling each model.

Specifically, the process begins with User Input, which supports both natural language and visual inputs. This input is then received by the Coordinator, the central hub that manages the workflow. The Coordinator determines whether the task involves Generation (creating/editing images) or Chat (a conversational interaction). If the path is Generation, the Planner comes into play to strategize the necessary steps to fulfill the user's request. The Supervisor oversees the execution of these steps, ensuring they are carried out correctly. Depending on the task, the agent might utilize a state-of-the-art text-to-image Generator, Image Editor, text/image-to-video Generator. For these generation steps, an Image Prompt Refiner is used to optimize the input image prompt for specific generation/editing tasks. Note that the agent may utilize Web Search Tool to gather relevant information from the web. If the path is Chat, the agent engages in a conversational exchange with the user. Finally, the output of the Web Search Tool or Generation Model is sent to an Output Checker module, which uses VLLM to determine whether the current generation task is complete and whether the output quality meets the required standards.

Image and Video Generation. For both image and video generation tasks, we adopt our new image/video generative foundation model (HiDream-I1 [2]) with 17 billion parameters that achieves state-of-the-art performances on several benchmarks (e.g., HPS benchmark [13], GenEval [5], DPG-Bench [6], and VBench [7]). In general, HiDream-I1 is built on a sparse Diffusion Transformer (DiT) structure, which first uses dual-stream DiT architecture to separately process image/frame and text tokens and then employs single-stream DiT architecture to enable multi-modal interaction. Both dual-stream and single-stream DiT architecture are remoulded with dynamic Mixture-of-Experts (MoE) design that aims to dynamically route data through specialized expert modules based on input characteristics.

Instruction-based Image Editing. Here we remould HiDream-I1 with strong "in-context" visual conditioning to support this task, enabling modifications to a source image based on a textual editing instruction. Specifically, we encode both source image and ground-truth target image into VAE latent space, and spatially concatenate their latent maps side-by-side as inputs. Then we fine-tune HiDream-I1 using latent flow matching over 5 million (source image, editing instruction, target image) triplets.

3 System and Evaluation

System and User Interface. As shown in Figure 1, our HiDream-Agent system can be easily accessed through an intuitive conversational AI interface, typically presented within a web browser. In particular, users interact with the agent by providing input through natural language typed into a chat window. The system also supports the uploading of images that the user wishes to edit. Each round of text-to-image generation or image editing typically takes approximately 10 seconds, while the generation of videos (5 seconds) may take around 30 seconds. Conversational responses from the Chat module are delivered in near real-time.

Evaluation. We also conduct user studies to evaluate our HiDream-Agent quantitatively. Specifically, we invite 20 labelers from different educational backgrounds to annotate the generated images/videos. Each labeler is asked to prepare 5 text prompts and 5 image prompts as the inputs of our agent system to generate images/videos or edit images. Then the labelers are asked to evaluate each generated image/video on a three-point ordinal scale (3: Good; 2: Neutral; 1: Bad). Finally, the satisfying rate of all results is 65%.

Acknowledgments. This work was supported in part by the Beijing Municipal Science and Technology Project No. Z241100001324002 and Beijing Nova Program No. 20240484681.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22563–22575.

[2] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705* (2025).

[3] Jingyuan Chen, Fuchen Long, Jie An, Zhaofan Qiu, Ting Yao, Jiebo Luo, and Tao Mei. 2025. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2079–2087.

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

[5] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. GENEVAL: an object-focused framework for evaluating text-to-image alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 52132–52152.

[6] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135* (2024).

[7] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.

[8] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6007–6017.

[9] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*. Springer, 468–485.

[10] Yang Luo, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Zhineng Chen, Yu-Gang Jiang, and Tao Mei. 2024. Freeenhance: Tuning-free image enhancement via content-consistent noising-and-denoising process. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7075–7084.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[13] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341* (2023).