MedDiT: A Knowledge-Controlled Diffusion Transformer Framework for Dynamic Medical Image Generation in Virtual Simulated Patient

Yanzeng Li[1,2], **Cheng Zeng**[3], **Jinchao Zhang**[2], **Jie Zhou**[2] and **Lei Zou**[1]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Pattern Recognition Center, WeChat AI, Tencent Inc.
[3]School of Computer Science, Wuhan University
liyanzeng@stu.pku.edu.cn, zengc@whu.edu.cn, {dayerzhang, withtomzhou}@tencent.com,
zoulei@pku.edu.cn

Abstract

Medical education relies heavily on Simulated Patients (SPs) to provide a safe environment for students to practice clinical skills, including medical image analysis. However, the high cost of recruiting qualified SPs and the lack of diverse medical imaging datasets have presented significant challenges. To address these issues, this paper introduces MedDiT, a novel knowledge-controlled conversational framework that can dynamically generate plausible medical images aligned with simulated patient symptoms, enabling diverse diagnostic skill training. Specifically, MedDiT integrates various patient Knowledge Graphs (KGs), which describe the attributes and symptoms of patients, to dynamically prompt Large Language Models' (LLMs) behavior and control the patient characteristics, mitigating hallucination during medical conversation. Additionally, a well-tuned Diffusion Transformer (DiT) model is incorporated to generate medical images according to the specified patient attributes in the KG. In this paper, we present the capabilities of MedDiT through a practical demonstration, showcasing its ability to act in diverse simulated patient cases and generate the corresponding medical images. This can provide an abundant and interactive learning experience for students, advancing medical education by offering an immersive simulation platform for future healthcare professionals. The work sheds light on the feasibility of incorporating advanced technologies like LLM, KG, and DiT in education applications, highlighting their potential to address the challenges faced in simulated patient-based medical education.

1 Introduction

Medical education plays a crucial role in preparing future healthcare professionals, relying extensively on Simulated Patients (SPs) to provide a safe and controlled environment for practicing clinical skills [Gaba, 2007; Ziv *et al.*, 2006; Sanko *et al.*, 2013; Mesquita *et al.*, 2010]. However, the traditional use of SPs presents significant challenges, primarily due to the high costs associated with recruiting and training qualified individuals [Hillier *et al.*, 2020; Felix and Simon, 2019]. While the development of Large Language Models (LLMs) offers the potential to build practical virtual SPs (VSPs) [Chen *et al.*, 2023; Benítez *et al.*, 2024; Holderried *et al.*, 2024; Li *et al.*, 2024a], the scarcity of diverse and comprehensive medical imaging datasets complicates the ability of VSPs to provide varied and realistic training scenarios [Glatard *et al.*, 2012; Baraheem *et al.*, 2023; Wang *et al.*, 2021].

To address these limitations, we introduce MedDiT, a novel VSP framework designed to enhance the educational experience. The core of MedDiT's functionality is the integration of patient Knowledge Graphs (KGs) [Fensel *et al.*, 2020; Gyrard *et al.*, 2018]. These KGs meticulously describe patient attributes and symptoms, serving as a foundation to guide the behavior of LLMs. By dynamically prompting LLM behavior, MedDiT ensures that patient characteristics are accurately represented, effectively mitigating issues such as hallucination during medical conversations [Li *et al.*, 2024a]. In addition to its conversational capabilities, MedDiT incorporates a series of well-tuned Diffusion Transformer (DiT) models [Yang *et al.*, 2023a; Pan *et al.*, 2023]. These models can generate medical images that correspond to the specified patient attributes within the KGs [Yang *et al.*, 2023b], providing a realistic and varied set of scenarios for students to engage with.

Essentially, MedDiT is a multi-agent system that integrates KG agent, chat agent, and image generation agent. This integration is centered around KG data, facilitating interaction and enabling the system to dynamically generate medical images that align with simulated patient symptoms. Through a practical demonstration, we showcase MedDiT's ability to simulate diverse patient cases and generate the corresponding medical images. This not only enhances the learning experience by offering an abundant and interactive platform but also advances medical education by providing an immersive simulation environment.

2 Methodology

In this study, we propose a framework for generating knowledge-driven imagery from structured KGs, delineated

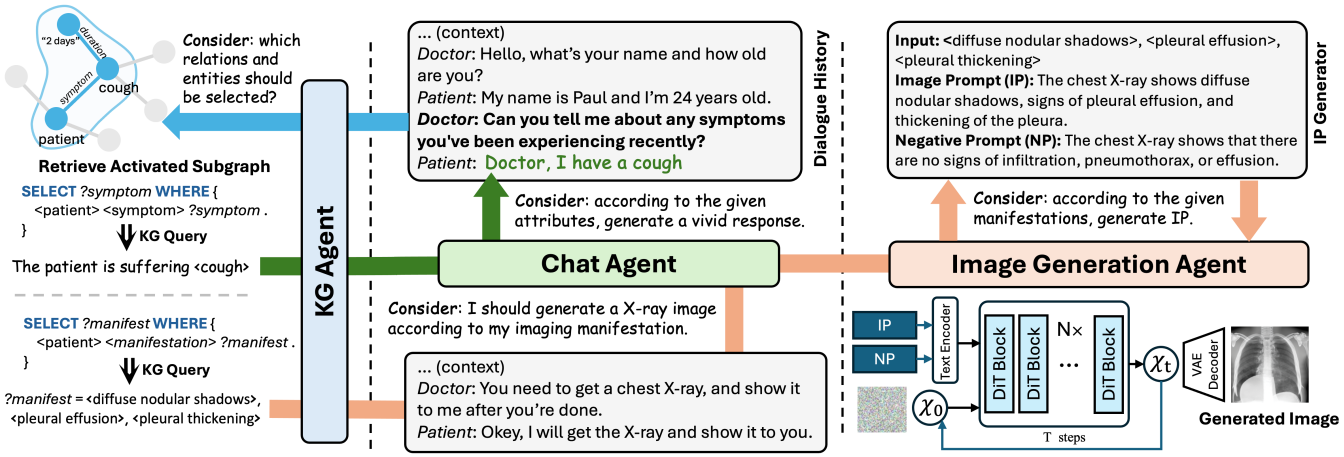

Figure 1: The overview diagram of MedDiT. There are 3 distinct LLM-based agents for controlling information flow across various modalities, including graph, text and image.

through the following processes:

A KG can be denoted as $G = (E, R, T)$, where E represents entities, R is the set of relations, and $T \subseteq E \times R \times E$ is the collection of knowledge triples $\langle s, p, o \rangle$, with $s, o \in E, p \in R$ being subjects, objects, and predicates, respectively. We define a function that represents the knowledge implied by arguments as:

$$\text{Knowledge} : X \to \mathcal{K},$$

where X can be a graph, text, or images, and $\mathcal{K}$ is the set of knowledge representations, implying that knowledge can be represented by $\langle s, p, o \rangle$.

Initially, we perform subgraph retrievel, where a subgraph $G' = (E', R', T')$ is derived from the original patient graph $G = (E, R, T)$, ensuring that $E' \subseteq E$, $R' \subseteq R$, $T' \subseteq T$, and $\text{Knowledge}(G') \subseteq \text{Knowledge}(G)$. This subgraph serves as the foundational data structure for subsequent operations. The retrieved subgraph G' can be represented by:

$$f : G' \to \text{Text}, \text{s.t.} \quad \text{Knowledge}(\text{Text}) \approx \text{Knowledge}(G'),$$

which maps the subgraph to a textual representation while preserving the inherent knowledge, such that the knowledge contained in the text equals that of the subgraph. In practice, we employ a strong LLM as the f for encoding subgraph to text as losslessly as possible. Subsequently, this textual representation $\text{IP} = f(G')$ is utilized as an image prompt within the DiT model, denoted as

$$I = \text{DiT}(\text{IP}),$$

and the total progress can be denoted as:

$$g : \text{Text} \to I, \text{s.t.} \quad I \models \text{Knowledge}(G'),$$

ensuring that the generated image I accurately reflects the knowledge embedded in G'. This comprehensive approach g describes the integration of graph-based data with text and image modalities, fostering enhanced interpretability and control in image synthesis.

3 System Design & Architecture

We developed MedDiT with a focus on modularity and configurability. The KG, chat, and multimodal components are constructed as microservices, while a unified LLM server is utilized to support the entire prompt workflow. In practice, we employ the Qwen2 72B instruction-tuned version [Yang *et al.*, 2024] as our backbone LLM for building all the agents. The overview of MedDiT is illustrated in Figure 1. Figure 2 presents the demonstration of MedDiT.

KG-enhanced LLM-based VSP. To construct a controllable and less-hallucinatory VSP system, we introduce the patient KG as the core information source as [Li *et al.*, 2024a]. A KG agent is applied to retrieve the conversation-related subgraph by discerning the user's intention and generating a SPARQL query to extract related entities, relations, and attributes in KG. Subsequently, the extracted subgraph is linearized into natural language to serve as role-setting for prompting the chat agent's responses. In this manner, MedDiT maintains dialogue consistency effectively, reduces the hallucinations, and saves token costs.

KG-controlled DiT Model. We utilize HunyuanDiT [Li *et al.*, 2024b] as our backbone generative model and train a LoRA adaptor for the transfer model to adapt to our target domain [Hu *et al.*, 2021]. In this demonstration, we selected a subset of Chest X-ray images from the Open-i dataset [Demner-Fushman *et al.*, 2016] for training, comprising 3,314 images along with their corresponding textual descriptions. The hyperparameters used to train the DiT model for generating responses during medical conversations are displayed in Table 1. The unmentioned parameters are aligned with the configuration of [Li *et al.*, 2024b].

Subsequently, we construct an agent to generate image prompts from the structured manifestations, which are stored in the KG and retrievable by KG agent. Once generated, the images are displayed in dialogue flow for students' further analysis.

SP Evaluation. As a VSP system, it is crucial to assess the dialogue history between the SP and students, score the inter-

Name	Description	Setting
$size$	Image size	1024×1024
d_{rank}	LoRA rank	64
t_{lora}	Weights applied with LoRA	$W_q, W_k, W_v, W_{\text{out}}$
d_{text}	Hidden size of the T5 encoder	2048
d_{clip}	Hidden size of the CLIP encoder	1024
l_{text}	Maximum length of the T5 encoder	256
lr	Learning rate	1e-4
$epoch$	Training epochs	100
$optim$	Optimizer for training LoRA adaptor	Adam

Table 1: Hyperparameter settings for our approach.

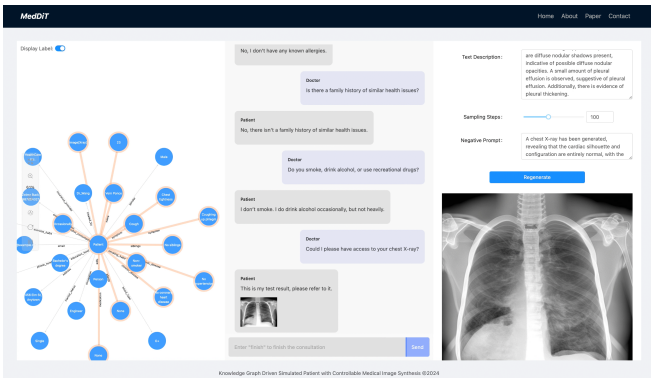

Figure 2: Screenshot of MedDiT. Left: Visualization of patient KG, indicating the activated subgraph for prompting conversaion. Center: The dialogue interface. Right: The DiT model interface.

actions, and provide feedback to help students enhance their medical conversation skills through iterative practice. Following the instructions from [Li *et al.*, 2024a], we utilize various prompts and auxiliary models to analysis the dialogue history across multiple aspects, including the completeness of necessary information, thoroughness of symptom inquiry, and the emotions conveyed to the patient. Finally, a detailed evaluation report, including scoring and advice, is generated by the LLM based on these criteria and the gold standard for the corresponding VSP case. In Listing 1, we present an example of a generated assessment designed to evaluate a student's performance. MedDiT concludes with a comprehensive score and a letter of advice to guide the student in their future practice.

4 Conclusion and Future Work

In conclusion, MedDiT provides a comprehensive and practical solution for building VSP systems that can dynamically generate medical images. By developing a multi-agent system based on LLMs, MedDiT has successfully integrated KGs to control the chat flow and utilized DiT models to produce medical images that align with the context and symptoms of patients. The system's ability to generate diverse and realistic medical images, combined with its interactive conversational capabilities, offers an abundant learning experience that can significantly improve medical education. In future work, we plan to further expand the scope of the KG to encompass a broader range of medical conditions, symptoms, and patient profiles. We will also increase the number of SP cases and explore training more kinds of medical image adaptors. Furthermore, we aim to conduct evaluation experiments on MedDiT to test large vision models and investigate the possibility of using large multi-modal models for comprehensive diagnosis.

Listing 1: An assessment example of medical conversation.

```
The ID for this assessment is {{ID}}, and
    your score is 69/100 points. Here are
    some suggestions for you:

### Summary:
During the process of assessing patients'
    health conditions, you have paid
    attention to many key factors, such as
    marriage, smoking, infectious diseases,
    vomiting, hepatitis, fatigue,
    tuberculosis, and parents' information.
    However, you seem to have not fully
    covered some important details, such as
    genetic history and symptoms.
As a medical professional, [...omit]

### Improvement Guidance:
Genetic History: You can briefly inquire
    about the patient's family medical
    history, [...omit]
Symptom Assessment: In every consultation, it
     is essential to fully understand the
    patient's symptoms. [...omit]
Comprehensiveness and Practicality: Your work
     is constantly committed to expanding
    knowledge boundaries and improving
    clinical skills. [...omit]
On this basis, every effort you make
    demonstrates an increasing attention to
    and understanding of patients' health,
    which is a significant improvement. [...
    omit]
```

Ethical Statement

This work utilizes AI-assisted models for medical education, which may generate inaccurate or risk content. All case materials (including medical images) have been rigorously reviewed by experts to ensure clinical validity. Strict expert oversight and adherence to standard medical protocols are mandatory to mitigate potential risks.

Acknowledgements

This work was partially supported by the National Key Research and Development Program of China (No. 2024YFF0907603), and National Key Laboratory of Data Space Technology and System. Lei Zou is the corresponding author of this paper. Yanzeng Li contributes this work during his internship at Tencent Inc.

References

[Baraheem *et al.*, 2023] Samah Saeed Baraheem, Trung-Nghia Le, and Tam V Nguyen. Image synthesis: a review

of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865, 2023.

[Benítez *et al.*, 2024] Trista M Benítez, Yueyuan Xu, J Donald Boudreau, Alfred Wei Chieh Kow, Fernando Bello, Le Van Phuoc, Xiaofei Wang, Xiaodong Sun, Gilberto Ka-Kit Leung, Yanyan Lan, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. *Journal of the American Medical Informatics Association*, page ocad252, 2024.

[Chen *et al.*, 2023] Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*, 2023.

[Demner-Fushman *et al.*, 2016] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[Felix and Simon, 2019] Heidi M. Felix and Leslie V. Simon. Types of standardized patients and their recruitment in medical simulation. 2019.

[Fensel *et al.*, 2020] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10, 2020.

[Gaba, 2007] David M Gaba. The future vision of simulation in healthcare. *Simulation in Healthcare*, 2(2):126–135, 2007.

[Glatard *et al.*, 2012] Tristan Glatard, Carole Lartizien, Bernard Gibaud, Rafael Ferreira Da Silva, Germain Forestier, Frédéric Cervenansky, Martino Alessandrini, Hugues Benoit-Cattin, Olivier Bernard, Sorina Camarasu-Pop, et al. A virtual imaging platform for multi-modality medical image simulation. *IEEE transactions on medical imaging*, 32(1):110–118, 2012.

[Gyrard *et al.*, 2018] Amelie Gyrard, Manas Gaur, Saeedeh Shekarpour, Krishnaprasad Thirunarayan, and Amit Sheth. Personalized health knowledge graph. In *CEUR workshop proceedings*, volume 2317. NIH Public Access, 2018.

[Hillier *et al.*, 2020] Maureen Hillier, Tony L Williams, and Tiffani Chidume. Standardization of standardized patient training in medical simulation. 2020.

[Holderried *et al.*, 2024] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, 10(1):e53961, 2024.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Li *et al.*, 2024a] Yaneng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*, 2024.

[Li *et al.*, 2024b] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, and Yingfang Zhang et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[Mesquita *et al.*, 2010] Alessandra R Mesquita, Divaldo P Lyra Jr, Giselle C Brito, Blcie J Balisa-Rocha, Patrcia M Aguiar, and Abilio C de Almeida Neto. Developing communication skills in pharmacy: a systematic review of the use of simulated patient methods. *Patient education and counseling*, 78(2):143–148, 2010.

[Pan *et al.*, 2023] Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 68(10):105004, 2023.

[Sanko *et al.*, 2013] Jill S Sanko, Ilya Shekhter, Richard R Kyle Jr, Stephen Di Benedetto, and David J Birnbach. Establishing a convention for acting in healthcare simulation: merging art and science. *Simulation in Healthcare*, 8(4):215–220, 2013.

[Wang *et al.*, 2021] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of applied clinical medical physics*, 22(1):11–36, 2021.

[Yang *et al.*, 2023a] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[Yang *et al.*, 2023b] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Zhou et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[Ziv *et al.*, 2006] Amitai Ziv, Paul Root Wolpe, Stephen D Small, and Shimon Glick. Simulation-based medical education: an ethical imperative. *Simulation in Healthcare*, 1(4):252–256, 2006.