Interact360: Interactive Identity-driven Text to 360° Panorama Generation

Zeyu Cai[1], Zhelong Huang[3], Xu Zheng[1], Yexin Liu[1], Chao Liu[1], Zeyu Wang[1,2], Lin Wang[1,2]
[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Hong Kong University of Science and Technology
[3] University of Science and Technology of China

***Abstract*—360° images offer an immersive and realistic visual experience for the emerging field, such as virtual tourism, particularly when users dream of recording their sweet moments without physically visiting a place. In this case, it is valuable to generate the 360° scene of a place (*e.g.*, seaside or mountainside) while allowing the users to generate their portraits given a taken face photo, which can be naturally harmonized with the generated 360° scene and their garment can be freely changed. In light of this, we propose a novel Interactive Identity-driven 360° Panorama Generation (Interact360) approach that produces vivid 360° panoramas featuring human portraits based on user inputs and commands, which encompass both the user's identity and the scene text description. Interact360 consists of three interactive components: prompt refinement module, identity-driven 2D portrait generation module, and 360° panorama generation module. Our approach enables users to modify and iterate on each component according to their individualized and customized requirements. We conduct both objective and subjective analysis to evaluate the effectiveness of our approach. The quantitative and qualitative results for the portrait image generation demonstrate the superiority of our method. Moreover, user studies provide empirical evidence of our method's effectiveness, demonstrating its strong interactivity and capacity to meet user requirements during the generation process, ultimately yielding satisfactory results for users. Our method highlights the potential benefits of panorama generation, particularly in addressing the personalized and customized needs of users.**

***Index Terms*—360 Scene Generation, Interactive AI, Text-to-Image Generation**

I. Introduction

Virtual reality (VR), acclaimed for its immersive interactivity, is increasingly leveraged in virtual tourism, providing sustainable, detailed exploration of diverse locales. This is particularly advantageous for regions with limited tourism infrastructure or environmental sensitivities [1]–[4]. In this domain, 360° images have become increasingly popular [5]–[9], offering immersive and realistic visual experiences that allow users to virtually revisit cherished moments. Advances include generating 360° scenes, such as seascapes or mountainscapes, allowing users to seamlessly integrate personalized portraits based on their photos. The ability to alter garments in these portraits further enhances customization and personalization.

To address these needs, we introduce a novel interactive identity-driven 360° panorama generation approach, dubbed **Interact360**. As shown in Fig. 1, Interact360 produces vivid 360° panoramas of a place (*e.g.*, seaside or mountainside)

Fig. 1: We introduce **Interact360**, an interactive text-to-image generative framework designed to create vivid 360° panoramas featuring human portraits, taking into account user inputs and commands that include both the user's identity and the scene description for virtual tourism.

featuring the portraits based on user inputs (*e.g.*, a taken face photo) and commands, encompassing both the user's identity and the scene text description. Our approach allows users to modify and iterate on each component according to their individualized and customized requirements. Specifically, Interact360 comprises three interactive components, including **1)** the prompt refinement module, **2)** identity-driven 2D portrait generation module, and **3)** 360° panorama generation module.

Specifically, for the prompt refinement module, we divide and refine the user's input text description into two parts—portrait and scene descriptions—to facilitate subsequent iterations of user interactions with the assistance of large language models (LLMs), such as ChatGPT [10]. The identity-driven 2D portrait generation module then combines the input identity image with the portrait description to generate a portrait image with varying garments while preserving the input identity by designing an identity-driven adapter. This adapter involves a combination of the input identity image and portrait description, resulting in the production of a portrait image that accurately corresponds to the desired garment and scene. Lastly, since Stable Diffusion [11] cannot be directly applied to generate 360° panorama, the 360° panorama generation module employs the Parameter Efficient Fine-tuning (PEFT) method [12] and Low-Rank Adaptation

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00141

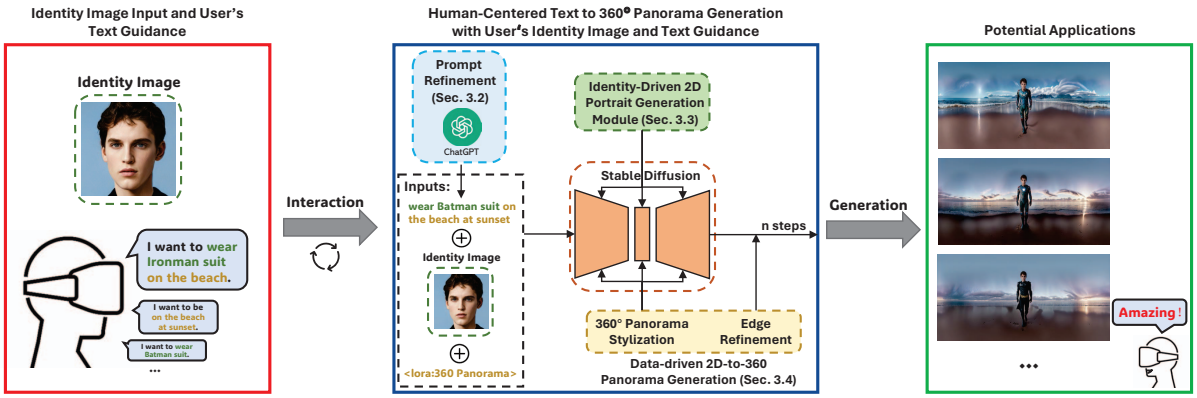

Fig. 2: An overview of our proposed Interact360 framework.

(LoRA) [13] to generate high-quality and realistic panoramas of various scenes, such as seaside or mountainside, while maintaining edge consistency to preserve the integrity of the entire surrounding generation.

We conduct both objective and subjective analyses to evaluate the effectiveness of our approach. The quantitative and qualitative results for the portrait image generation demonstrate the superiority of our method. Also, intensive user studies provide empirical evidence of our approach's effectiveness, demonstrating its strong interactivity and capacity to meet user requirements during the generation process, ultimately yielding satisfactory results for users. Our method highlights the potential benefits of panorama generation for 360° images, particularly in addressing the personalized and customized needs for virtual tourism.

In summary, the main contributions of our work are as follows: (**I**) We propose a novel interactive identity-driven text-to-panorama generation approach that produces vivid 360° panoramas featuring human portraits based on user inputs and commands in an interactive manner. (**II**) We present an identity-driven portrait generation module capable of producing high-quality images of the input identity that accurately match the text description for the garment. Additionally, we introduce a 360° panorama generation module that is capable of generating high-quality and realistic panoramas. (**III**) We conduct both quantitative and qualitative analyses to evaluate the effectiveness of our approach. (**IV**) We further conduct user studies to verify our approach's effectiveness, demonstrating its strong interactivity and capacity to meet user's requirements during the generation process for virtual tourism.

II. Related Works

360 Virtual Reality and Virtual Tourism Virtual tourism, increasingly recognized as a viable alternative to conventional tourism for locations like museums and historical landmarks, offers immersive experiences through 360° imagery and video, negating the need for physical travel [14] This modality enables users to create personalized portraits, including garment customization, in virtual scenes, enhancing user engagement. The transformative potential of this technology spans various sectors, such as tourism, education, and entertainment, prompting extensive research into more sophisticated 360° VR systems and their diverse applications [15], [16].

Text to Image Generation (T2I) T2I generation, aiming to create high-fidelity images from textual descriptions, initially employed Generative Adversarial Networks (GAN) [17]. Recent advances have scaled data and model sizes, with leading techniques being auto-regressive [18]–[21] or diffusion-based [11], [22]–[24]. A significant breakthrough, DALL-E, showcased the zero-shot capabilities of auto-regressive models, while diffusion models like DALL-E 2 [25] have also demonstrated impressive results. Methods such as Imagen [24], DALL-E2 [25], and Stable Diffusion [11] have achieved advanced semantic generation.

However, these methods primarily generate 2D images with limited Field-of-View (FoV), not suitable for 360° panoramas. The challenge in 360° panorama generation lies in the complex omnidirectional scene (180° × 360°) and integrating user portraits. Text2Light [26], an early effort in zero-shot text-driven panorama generation, focuses only on scene imagery, not editable portrait integration. Addressing these issues, we extend the Stable Diffusion model [11] to 360° panoramas, adapting 2D generative models to the 360° domain with conditioned inputs including user identity and scene descriptions.

Identity-driven 2D Portrait Generation Image generation has evolved to encompass both implicit conditions like style and mode, and explicit conditions such as scene graphs [27], [28]. Hybrid condition approaches, combining layout and style with text, have emerged in recent advancements [17], [29]. Diffusion models now accommodate diverse inputs, including sketches [30], semantic masks [31], text [11], [24], and more [32]–[34], enhancing versatility in conditioned image generation. Recent test-time fine-tuning methods allow for personalized image generation from individual images [35]–[39]. Subject-Diffusion, for instance, facilitates subject-driven image generation with just a single reference image for personalized single- or multi-subject generation across domains [40]. Building on this, our method introduces identity-driven image generation for identity-driven panoramas, aiming to produce

Prompt Disentanglement: … MUST use the following format:
portrait_description: [‘the clothes I wear’]
scene_description: [‘the scene I stay’]
description: ‘description of the clothes I wear and the scene I stay’

User Prompt: I want to wear Batman suit on the beach at sunset
Generated Response:
portrait _description: [‘wear Batman suit’]
scene_description: [‘on the beach at sunset’]
description: ‘wear Batman suit on the beach at sunset’

(a) Prompt Disentanglement

Portrait Iteration: … MUST use the following format:
portrait_description: [‘the clothes I wear’]
scene_description: [‘the scene I stay’]
description: ‘description of the clothes change’

User Prompt: I want to wear Spiderman suit
Generated Response:
portrait _description: [‘wear Spiderman suit’]
scene_description: [‘on the beach at sunset’]
description: ‘Spiderman suit’

(b) Portrait Iteration

Scene Iteration: … MUST use the following format:
portrait_description: [‘the clothes I wear’]
scene_description: [‘the scene I stay’]
description: ‘description of the scene change’

User Prompt: I want to be on the beach at noon
Generated Response:
portrait _description: [‘wear Spiderman suit’]
scene_description: [‘on the beach at noon’]
description: ‘on the beach at noon’

(c) Scene Iteration

Fig. 3: The process of the prompt disentanglement in the proposed prompt refinement module.

clear, realistic portraits aligned with textual descriptions while maintaining the subject’s identity.

III. The Proposed Approach

A. Overview

Fig.2 depicts Interact360, a deep-learning framework for creating high-quality, text and identity image-driven panoramas. Initially, using LLMs like ChatGPT [10], Interact360 splits and refines input text into portrait and scene descriptions (Sec.III-B). It then generates identity-specific 2D portraits and 360° panoramas (Sec.III-C and Sec.III-D). Enhanced by Stable Diffusion (Rombach et al., 2022), Interact360 ensures realistic panorama integrity with edge refinement (Sec.III-D). SectionIII-E details the training process, focusing on a large dataset of text and identity images.

B. Prompt Refinement Module

Interact360 takes as input the textual description and an identity image (face image) from the user. The following is a detailed procedure of our prompt refinement module:
(a) The module first divides and refines the text description into two segments: scene generation and portrait generation, as illustrated in Fig. 3. This segmentation enables better adjustment of the textual description, allowing the model to generate more accurate and realistic images.
(b) To enhance user-friendliness, our prompt refinement module eliminates the need for users to separately describe the scene and the portrait. Instead, users provide an all-encompassing image description, with the module automatically segmenting it into scene and portrait descriptions.
(c) After the initial generation of a 360-degree panorama, users can further modify and edit the generated scene or portrait. This iterative process is interactive and facilitated by ChatGPT [10] with specialized functionalities in collaboration with the user. As depicted in Fig. 1, the user can refine the generated panorama by providing further instructions, such as ”*I want to wear a Batman suit.*” to obtain a new panorama that fulfills their desired specifications. This iterative process enables users to customize and personalize their panoramas according to their individual preferences and requirements, resulting in more satisfactory and personalized outputs.

C. Identity-driven 2D Portrait and Scene Generation

Our identity-driven 2D portrait generation module, depicted in Fig. 4, is designed to produce high-quality images matching the user’s text description while maintaining the specified identity. This module alters the portrait’s attire and background while preserving identity characteristics.

The module’s core, an identity-driven adapter (Fig. 4), integrates the input identity image and text description to generate images reflecting the desired outfit and setting. Comprising patch-MLP and identity-driven adapters, it processes image patch features from the encoder, then combines these with text prompt embeddings. Utilizing self-attention, detailed in [13], the adapter ensures fidelity to both identity and textual input.

1) Architecture Details: **Input Image and Text Preprocessing.** We propose a streamlined Textual Inversion approach, utilizing only the Patch Feature of input images for Stable Diffusion. This method, diverging from existing works [40], [41], enhances efficiency by eliminating the need for image description reproduction.

Interact360 employs Retinaface [42] for precise detection and cropping of the human face, I^{face}, from the input image I. This cropped face, combined with the text description, refines the image generation process, enabling Interact360 to produce more accurate and lifelike portraits.
Patch Feature Learning. To ensure the model effectively learns the details of the input cropped face image, we utilize the CLIP [43] image encoder to extract the patch feature f_p from the input face image I^{face}. However, since Stable Diffusion [11] has not been trained with the input face image, it is necessary to pre-process the patch features with a learnable Patch MLP module:

$$\hat{f}_p = \text{Patch\_MLP}(f_p) \tag{1}$$

where $\hat{f}_p$ is the processed patch feature. This allows for a better understanding of the details of identity images, given that Stable Diffusion [11] is originally trained on large-scale image text pairs.

We then feed the patch feature $\hat{f}_p$ into the Stable Diffusion [11] to incorporate identity information during generation. Since the UNet of pre-trained Stable Diffusion [11] consists

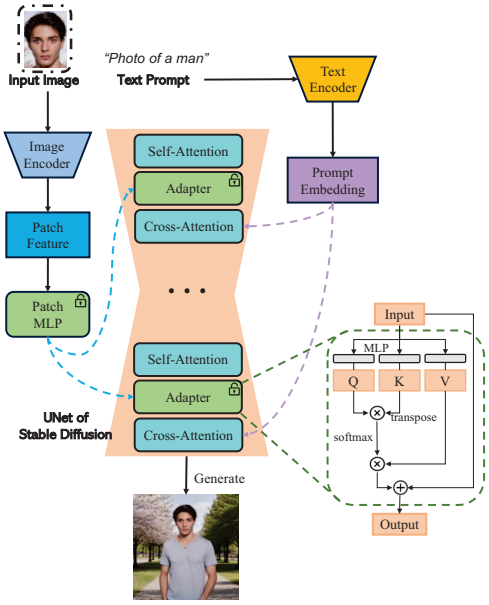

Fig. 4: The structure of our identity-driven module.

of several Transformer blocks, we adopt a similar approach as previous works [33], [40], [41]. Specifically, we insert a learnable identity-driven adapter layer between the self-attention and cross-attention layers in each Transformer block, as shown in Fig. 4, which is formulated as:

$$Z := Z + \beta \cdot \tanh(\gamma) \cdot \mathbf{S}([Z, \hat{f}_p]), \tag{2}$$

where Z is the output of the self-attention layer, β is a constant that balances the importance of the identity-driven adapter layer, γ is a learnable scalar initialized to 0, and $\mathbf{S}$ is the self-attention operator whose input is the concatenation of Z and patch feature $\hat{f}_p$, which can be formulated as:

$$\hat{Z} = \mathbf{Attention}(Q, K, V) = \mathbf{Softmax}(\frac{QK^T}{\sqrt{\hat{d}}}) \cdot V \tag{3}$$

where $Q = W^Q[Z, \hat{f}_p], K = W^k[Z, \hat{f}_p], V = W^V[Z, \hat{f}_p]$ are three projections of the input, $\hat{d}$ is the rescale dimension and $\hat{Z}$ is the output of self-attention. During UNet model training, we selectively activate the key and value layers of cross-attention and identity-driven adapter layers while freezing the remaining layers. This approach enables the model to focus more on learnable layers.

D. Data-driven 2D-to-360 Panorama Generation

360 Panorama Stylization: Having generated the 2D portrait and scene, we now transition to 2D-to-360 panorama generation. While the traditional Stable Diffusion model [11] excels in 2D image generation from text, it is not inherently equipped for 360° panoramas. To overcome this, we employ Parameter Efficient Fine-tuning (PEFT) through Low-Rank Adaptation (LoRA) [13], enabling the adaptation of Stable Diffusion for 360° panorama creation.

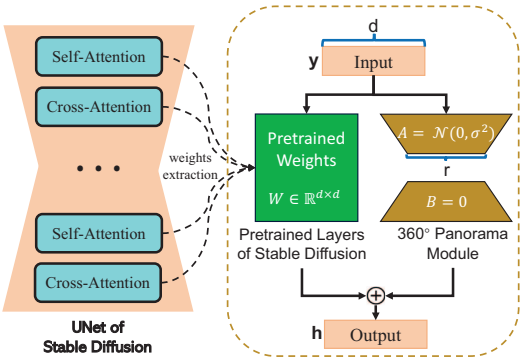

Fig. 5: The structure of data-driven 2D-to-360 panorama generation module.

Stable Diffusion, a Vision-Language Model (VLM), incorporates numerous dense layers with full-rank weight matrices [11]. LoRA shows that for downstream tasks, effective learning can occur by training only a smaller subspace of these layers [13]. We utilize a low-rank decomposition to represent the pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ as $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. During training, W_0 remains unchanged while A and B are trainable, with both W_0 and $\Delta W = BA$ contributing to the output:

$$h = W_0 y + \Delta W y = W_0 y + BAy, \tag{4}$$

Here, y denotes the model's input, and BA functions as an adjustable module for task-specific learning, like 360° panorama generation. The module's structure is shown in Fig. 5, where A and B are initially set to random Gaussian and zero values, respectively. This initialization aids training stability and allows gradient updates. Starting with $B = 0$ minimizes early-stage instability in the 360° panorama module, enhancing final output quality.

Edge Refinement: Although we can obtain the 360° panorama through the 360° panorama stylization, generating vivid panoramas requires meeting strict indicators, with the alignment of the two sides of the generated image being of utmost importance. As initially designed for generating general 2D images, Stable Diffusion [11] struggles to meet the specifications of 360° panoramas through learning alone, even with fine-tuning using the PEFT method, which is not enough for panorama generation. Therefore, we propose an edge refinement module to achieve edge alignment in panoramas.

The proposed edge refinement enhances the accuracy and effectiveness of the Interact360 model by ensuring that the generated panoramas are visually consistent and realistic, with smooth transitions between different parts of the panorama. The alignment of the left and right ends of 360° panoramic images is the key to ensuring the continuity of 360° view. Since different scenes have different components, it is necessary to maintain the consistency of the components on both sides. Inspired by Multi-Diffusion [36], We design an edge refinement module, which uses the sliding window to

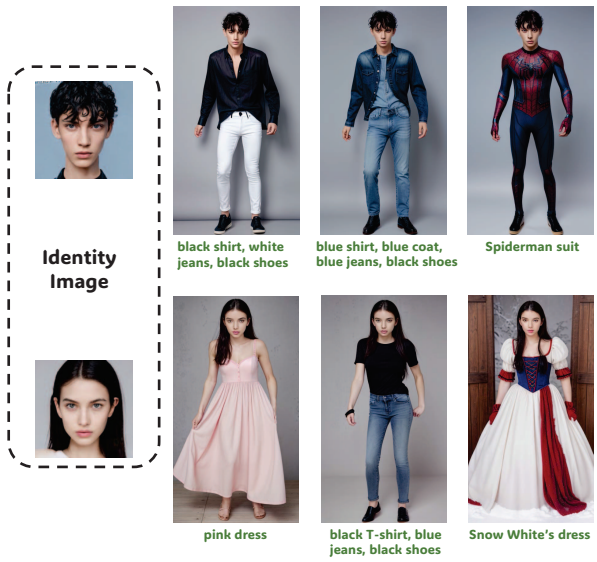

Fig. 6: Experimental results with varying garment description.

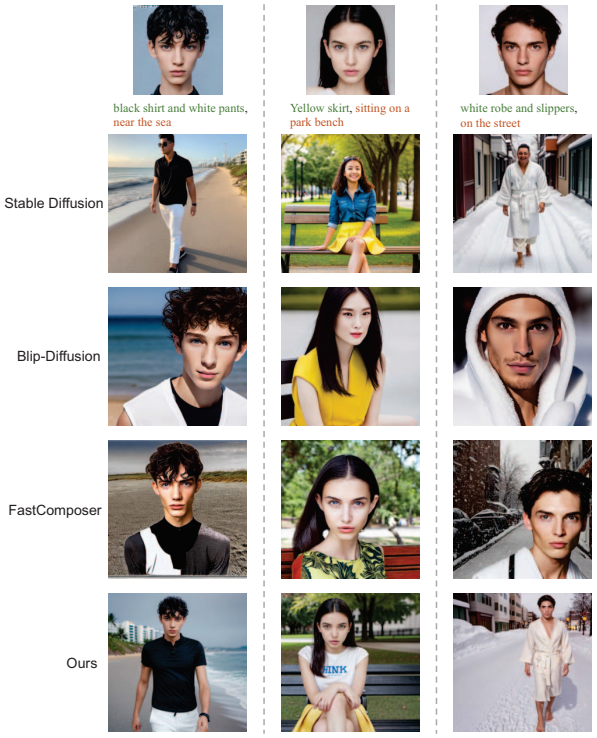

Fig. 7: Comparison of different methods for identity-driven 2D portrait and scene generation.

continuously slide on both sides of the semantic latent, and finally achieve the semantic consistency of the edge of the image by calculating the weighted average values.

E. Model Training and Inference

Model Training: Due to the specificity of our undertaking, existing datasets are not directly applicable for model training. Nonetheless, our identity-driven portrait generation module and 360° panorama approach are designed for flexibility and modularity. This allows for distinct training of the portrait generation component with the identity-driven adapter and the 360° panorama generation using the Scene Adapter. In training the identity-driven portrait generation, the module is integrated into Stable Diffusion as the sole learnable segment [11]. We employ data augmentation $\mathcal{A}$ to generate varied identity image instances I_s^{face}, using original images as the baseline for accuracy. The loss function $\mathbf{L}_h$ to optimize the model:

$$\mathbf{L}_h = \mathbb{E}_{z,t,y^h,I_s^{face},\epsilon\in\mathbf{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, y^h, I_s^{face})\|_2^2\right] \quad (5)$$

where y^h is the textual embedding that specifically describes the portrait, z_t is the latent noisy image at time step t, ϵ is the latent noise to predict, and ϵ_θ is the noise prediction model with parameters θ.

For 360° panorama generation training, the learnable data-driven 2D-to-360° panorama generation module works with pre-trained Stable Diffusion [11]. Given y^p, the textual embedding of the provided 360° panorama scene, similar to $\mathbf{L}_h$, the loss function $\mathbf{L}_p$ is:

$$\mathbf{L}_p = \mathbb{E}_{z,t,y^p,\epsilon\in\mathbf{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, y^p)\|_2^2\right] \quad (6)$$

In order to distinguish 360° panoramas from general images, we add a special tag <lora:360 Panorama> to the text description of 360° panoramas. Only the proposed two modules are trained in the above two training processes, which fully maintains the generation capability of the original Stable Diffusion [11].

Model Inference: While the identity-driven 2D portrait and scene generation module and the 2D-to-360° panorama generation module undergo separate training, their processes maintain the original parameters of Stable Diffusion intact [11]. Consequently, these modules can be seamlessly integrated during model inference, enabling the creation of 360° panoramas that include identity-preserving portraits.

IV. Experiments and Evaluation

Dataset Our empirical analysis utilizes two distinct datasets: a human-based dataset for identity-driven portrait generation and a 360° panorama dataset. The former is derived from the FFHQ dataset [44], enhanced with identity-augmented image-text pairs. Captions were generated using BLIP-2 [45], and identity regions were extracted via Retinaface [42]. The 360° panorama dataset, comprising approximately 2000 captioned images, was sourced from [46]. **Implementation Details** We base our model on Stable Diffusion v1-5 [11], complemented by OpenAI's clip-vit-large-patch14 vision model for visual identity encoding [43]. Training focuses on the Patch MLP and Adapter within the portrait and panorama generation modules, keeping other model components static.

"a beautiful garden with flowers"

"a snow mountain at sunset"

Fig. 8: Generated 360° panoramic scenes.

TABLE I: Quantitative comparison with different methods.

Method	CLIP-Score↑	Face Similarity↑	Mean↑
Stable Diffusion [11]	33.11	3.28	17.69
Blip-Diffusion [45]	27.22	23.58	25.4
FastComposer [47]	27.58	33.89	30.73
Interact360 (Ours)	31.68	31.99	**31.83**

A. Experimental Results

Fig. 6 showcases the proficiency of Interact360 in producing high-quality portraits from user inputs, affirming its capability to handle diverse gender representations. These outcomes underline the model's robustness and fidelity in generating realistic portraits. Further visual comparisons are available in Fig. 7. Fig. 8 demonstrates our 360° panorama module's efficacy in creating panoramas from textual scene descriptions. To evaluate Interact360's adaptability, we varied garment or scene descriptions in inputs while maintaining other factors constant. As depicted in Fig.9 (a&b), the model reliably produced accurate and aesthetically pleasing garments, capturing nuanced details like color (**black T-shirt**, **blue jeans**) and style (**Spiderman suit**, **Superman suit**). Similarly, Fig.9 (c&d)illustrates the model's consistent and detailed scene generation, such as differing lighting and objects (**sunset**, **noon**, **lake**, **garden**).

These results validate the efficacy of Interact360 in generating authentic and detailed 360-degree images, offering transformative potential in online shopping and virtual tourism. The model enhances user experience by providing immersive previews of garments and realistic virtual exploration of diverse scenes and environments.

B. Experimental Settings

We conduct a quantitative comparison of our proposed identity-driven module with other methods on the test samples. The comparison is based on two indicators: CLIP-Score, which reflects the degree of matching between the generated image and the text description, and Face Cosine Similarity, which reflects the similarity between the generated portrait and the input face image. Since our method does not require additional time-consuming fine-tuning for any input face, we compare it with other methods that also do not require fine-tuning, including Stable Diffusion [11], Blip-Diffusion [45], and FastComposer [47]. We utilize OpenCLIP [48] to calculate CLIP-Score and MTCNN [49] with AdaFace [50] to calculate Face Cosine Similarity.

1) Quantitative Analysis: Table. I presents the quantitative comparison results of our method with other methods. Stable Diffusion [11], which does not consider identity information, has the highest CLIP-Score, indicating that its generation results match the input description best. However, as expected, Stable Diffusion [11] does not preserve identity information. Our proposed method achieves good performance on both CLIP-Score and Face Similarity, although its Face Similarity score is slightly lower than that of FastComposer [47]. Notably, our method generates more natural results than FastComposer [47], as demonstrated in Fig. 7.

C. User Study

A series of user studies was conducted to evaluate the robustness and applicability of our proposed methodology. The studies centered on participants' subjective assessments of its effectiveness.

1) Participants: Twenty-four individuals participated in the study. The majority, 70.83%, were aged 18-24, with 29.17% in the 25-34 age range. Gender distribution was 79.17% male and 20.83% female, and 91.67% had prior VR experience.

2) Task: The study included three tasks designed to assess different aspects of our method:

- **Human Garment Change Task:** Participants provided textual descriptions of preferred garments, which our system used to optimize prompts and generate images combining the user's facial image with the garment style. Participants rated these images.
- **360-Degree Background Image Manipulation Task:** Participants specified desired scenes, and our system generated corresponding background images. These images were then evaluated and scored by the participants.
- **Simultaneous Manipulation of Human and Background Images Task:** Participants detailed preferences for both garment and background settings. The system produced composite images based on these inputs, which participants subsequently assessed and scored.

3) Measurement: After completing tasks, participants rated our framework on a 7-point Likert scale, covering its utility in VR scenarios, immersion enhancement through text-generated images, preferences for image intuition versus detail, and text-image correspondence. They also evaluated the aesthetic

Fig. 9: Experimental results: (a,b): with varying **scene text description**. (c,d): with varying **garment text description**.

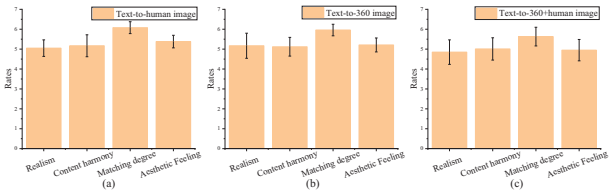

Fig. 10: The responses from participants were evaluated using a 7-point Likert scale.

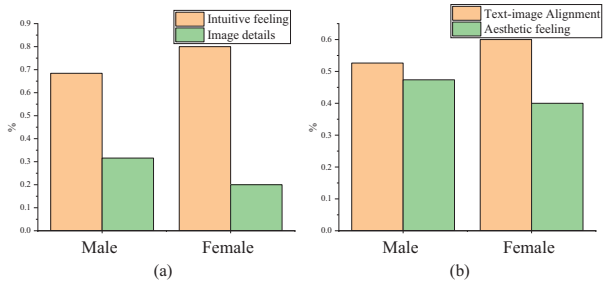

Fig. 11: The responses from participants were evaluated using a 7-point Likert scale.

appeal, realism, content harmony, manipulability, and textual congruence of generated images. Feedback on the overall aesthetic experience and emotional responses during image generation was also collected.

4) Results: This section analyzes participants' feedback on their experience with our methodology, highlighting insights from all three tasks.

Perceived Success of Text-to-Human Image Generation. Participants rated the realism of our text-to-human image generation highly, averaging 5.05, and consistently above 4. Content harmony received higher mean scores than realism but with greater score variance, indicating varied perceptions of consistency. Our method's text-image matching, especially in garment styles, scored high with less variance, demonstrating alignment proficiency. Ratings on aesthetic appeal often surpassed 5, indicating a favorable reception of the images' visual quality.

Perceived Success of Text-to-360° Image Generation. The perceived realism in generating 360° background images using our method was higher than that of text-to-human images, with an average score of 5.17. However, content harmony scores were lower, likely due to the expansive viewing angle of 360° images, which may render local details less noticeable and cause overall distortion. Additionally, the score for text-image matching slightly decreased from 6.08 to 5.96, possibly because some details in the 360-degree images might not have fully aligned with participants' expectations.

Perceived Success of Text-to-360° and Human Image Generation. Of the three tasks, the simultaneous alteration of character appearance and 360° background received the lowest scores: 4.85 for realism, 5.01 for content harmony, 5.63 for text-image matching, and 4.95 for aesthetic perception. This relatively lower evaluation may stem from the complexities in rendering and integrating various objects within 360° scenes. The intricate task of achieving high realism while modifying both the character and the background likely contributed to these reduced scores across the evaluation metrics.

V. Limitations and Future Work

The future work for our project includes expanding user study scope for more diverse insights, enhancing image realism to approach the quality of actual photographs, broadening scene generation with deterministic algorithms for predictable environments, and implementing local change capabilities for targeted scene modifications. These improvements aim to enhance user experience and system utility.

VI. Conclusion

In this paper, we presented an interactive framework for generating 360° panoramic images from texts. Given the scene and human texts described by users, our method can automatically refine the prompt of the user's text description, generate an identity-driven portrait, and generate a 360° panoramic image. To assess the framework's effectiveness, we conducted

comprehensive evaluations, encompassing qualitative, quantitative, and user studies. The findings indicate that our approach holds significant promise in potential applications, *e.g.*, virtual tourism.

REFERENCES

[1] W. Wei, M. A. Baker, and I. Onder, "All without leaving home: building a conceptual model of virtual tourism experiences," *International Journal of Contemporary Hospitality Management*, vol. 35, no. 4, pp. 1284–1303, 2023.

[2] H. Go and M. Kang, "Metaverse tourism for sustainable tourism development: Tourism agenda 2030," *Tourism Review*, vol. 78, no. 2, pp. 381–394, 2023.

[3] L. Gamidullaeva, A. Finogeev, M. Kataev, and L. Bulysheva, "A design concept for a tourism recommender system for regional development," *Algorithms*, vol. 16, no. 1, p. 58, 2023.

[4] X. Chen and Z.-f. Cheng, "The impact of environment-friendly short videos on consumers' low-carbon tourism behavioral intention: A communicative ecology theory perspective," *Frontiers in Psychology*, vol. 14, p. 1137716, 2023.

[5] S. Verma, L. Warrier, B. Bolia, and S. Mehta, "Past, present, and future of virtual tourism-a literature review," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100085, 2022.

[6] X. Wu and I. K. W. Lai, "The use of 360-degree virtual tours to promote mountain walking tourism: stimulus–organism–response model," *Information Technology & Tourism*, vol. 24, no. 1, pp. 85–107, 2022.

[7] X. Zheng, J. Zhu, Y. Liu, Z. Cao, C. Fu, and L. Wang, "Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1285–1295, 2023.

[8] X. Zheng, T. Pan, Y. Luo, and L. Wang, "Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18687–18698, 2023.

[9] X. Zheng, P. Zhou, A. Vasilakos, and L. Wang, "Semantics, distortion, and style matter: Towards source-free uda for panoramic segmentation," *arXiv preprint arXiv:2403.12505*, 2024.

[10] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[12] Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, and N. Collier, "On the effectiveness of parameter-efficient fine-tuning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 12799–12807, 2023.

[13] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[14] T. Zhang and A. Hacikara, "Virtual tourism and consumer wellbeing: A critical review, practices, and new perspectives," *Handbook of Tourism and Quality-of-Life Research II: Enhancing the Lives of Tourists, Residents of Host Communities and Service Providers*, pp. 545–557, 2023.

[15] C. Groth, J.-P. Tauscher, N. Heesen, M. Hattenbach, S. Castillo, and M. Magnor, "Omnidirectional galvanic vestibular stimulation in virtual reality," *IEEE transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2234–2244, 2022.

[16] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360 images," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2003–2013, 2022.

[17] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

[18] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *European Conference on Computer Vision*, pp. 89–106, Springer, 2022.

[19] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *European conference on computer vision*, pp. 720–736, Springer, 2022.

[20] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[21] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16890–16902, 2022.

[22] D. Xing and A. Tzes, "Synthetic aerial dataset for uav detection via text-to-image diffusion models," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, pp. 51–52, IEEE, 2023.

[23] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10157–10166, 2023.

[24] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[26] Z. Chen, G. Wang, and Z. Liu, "Text2light: Zero-shot text-driven hdr panorama generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022.

[27] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1219–1228, 2018.

[28] Y. Lyu, X. Zheng, and L. Wang, "Image anything: Towards reasoning-coherent and training-free multi-modal image generation," *arXiv preprint arXiv:2401.17664*, 2024.

[29] J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Text-to-image generation grounded by fine-grained user attention," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 237–246, 2021.

[30] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

[31] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," 2023.

[32] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," *arXiv preprint arXiv:2302.09778*, 2023.

[33] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

[34] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[35] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[36] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

[37] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," *arXiv preprint arXiv:2303.11305*, 2023.

[38] Z. Fei, M. Fan, and J. Huang, "Gradient-free textual inversion," *arXiv preprint arXiv:2304.05818*, 2023.

[39] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation," *arXiv preprint arXiv:2305.03374*, 2023.

[40] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," *arXiv preprint arXiv:2307.11410*, 2023.

[41] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," *arXiv preprint arXiv:2304.03411*, 2023.
[42] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.
[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
[44] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.
[45] D. Li, J. Li, and S. C. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *arXiv preprint arXiv:2305.14720*, 2023.
[46] https://pixexid.com/search/360-panorama.
[47] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, "Fastcomposer: Tuning-free multi-subject image generation with localized attention," *arXiv preprint arXiv:2305.10431*, 2023.
[48] https://github.com/mlfoundations/open_clip.
[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[50] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.