Multi-modal and Multi-scale Spatial Environment Understanding for Immersive Visual Text-to-Speech

Rui Liu[1*], **Shuwei He**[1], **Yifan Hu**[1], **Haizhou Li**[2, 3]

[1]Inner Mongolia University, China
[2]Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
liurui_imu@163.com, shuwei_he@163.com, hyfwalker@163.com, haizhouli@cuhk.edu.cn

A. Detailed Experimental Settings

We list the detailed model setup of M^{2}SE-VTTS. The detailed experimental settings are shown in Table 1.

Hyperparameter		M^{2}SE-VTTS
Multi-modal and Multi-scale scheme	Feature Dimension	512
	RGB Detector Attention Heads	4
	Attention Dropout	0.1
	Top$_k$	140
	Local-aware Attention Heads	4
	Semantic-guild Attention Heads	4
TTS-Encoder	Pre-net Layers	3
	Pre-net Hidden	512
	Phoneme Embedding	512
	Encoder Layers	4
	Encoder Conv1d Kernel	9
	Conv1D Filter Size	1024
	Encoder Dropout	0.5
Variance Predictor	Conv1D Kernel	3
	Conv1D Filter Size	512
	Dropout	0.5
Denoiser	Diffusion Embedding	384
	Transformer Layers	5
	Transformer Hidden	384
	Attention Heads	12
	Position Embedding	384
	Scale/Shift Size	384
Total Number of Parameters		105.35M

Table 1: Detailed Model Setup of M^{2}SE-VTTS.

B. Case Study

To more intuitively illustrate the outcomes of the Local Spatial Understanding, we present some visual instances. These include both the RGB and Depth images of the same scene, the spatial environment prompt, the environment captions acquired with Gemini Pro Vision, and the Top$_k$ patch regions identified in the RGB space and selected from the Depth space based on the aforementioned environment caption, where the patches identified are marked in red.

As shown in Fig. 1, Gemini first accurately describes the spatial information of the current environment, such as the relative positioning of the picture and the counter with respect to the speaker's location in the first example. In addition, the Local Spatial Understanding can effectively select semantic-aware patches that are relevant to the captions. Specifically, based on the speaker elements identified in each caption, the patches that correspond to the speakers' locations are selected. In the first example, the patches corresponding to the semantic information " the counter " are accurately identified. After conducting experimental comparisons, we selected the top 140 patches from a total of 256 as the final result, with each patch having a size of 16×16 pixels.

*Corrposending Author.

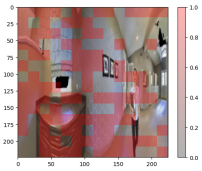
RGB Image

Observe this panoramic image and briefly describe its content. Identify the objects in the image in one to two sentences, focusing only on key information and avoiding descriptive words.

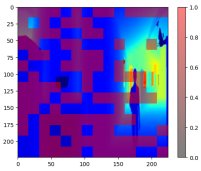
Spatial Environment Prompt

The image shows a long hallway with a woman walking down it. There are pictures on the wall to her left and a counter to her right.

Environment Captions



Top_k Patches of the RGB Space(K=140)



Top_k Patches of the Depth Space(K=140)

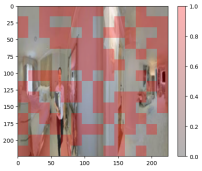
RGB Image

Observe this panoramic image and briefly describe its content. Identify the objects in the image in one to two sentences, focusing only on key information and avoiding descriptive words.

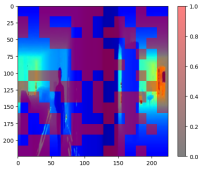
Spatial Environment Prompt

The image shows a woman walking in a house. The woman is wearing a black suit. The house is large and has a modern design. The woman is walking through a hallway. There are rooms on either side of the hallway. The rooms are all decorated in a modern style.

Environment Captions



Top_k Patches of the RGB Space(K=140)



Top_k Patches of the Depth Space(K=140)

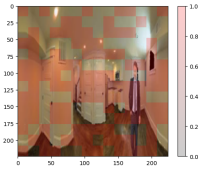
RGB Image

Observe this panoramic image and briefly describe its content. Identify the objects in the image in one to two sentences, focusing only on key information and avoiding descriptive words.

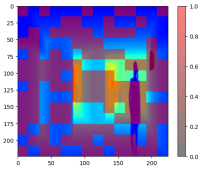
Spatial Environment Prompt

The image shows a man in a suit standing in a room. The room has white walls, brown wooden floors, and a white ceiling. There is a staircase leading up to the left, a door on the right, and a kitchen in the back.

Environment Captions



Top_k Patches of the RGB Space(K=140)



Top_k Patches of the Depth Space(K=140)

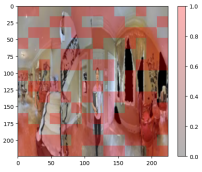
RGB Image

Observe this panoramic image and briefly describe its content. Identify the objects in the image in one to two sentences, focusing only on key information and avoiding descriptive words.

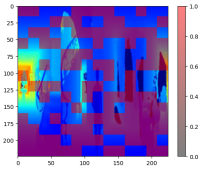
Spatial Environment Prompt

The image shows a woman walking in a large room. There is a staircase, a sofa, and a dining table in the room.

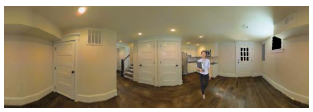
Environment Captions



Top_k Patches of the RGB Space(K=140)



Top_k Patches of the Depth Space(K=140)



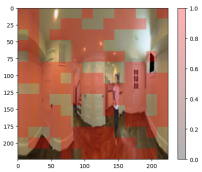
RGB Image

Observe this panoramic image and briefly describe its content. Identify the objects in the image in one to two sentences, focusing only on key information and avoiding descriptive words.

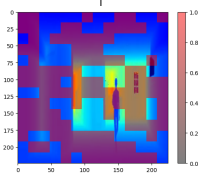
Spatial Environment Prompt

The image shows a mostly empty basement with wood floors and white walls. There is a woman standing in the center in front of the camera. There are two doors on the left, one on the right, and a staircase leading up on the left.

Environment Captions



Top_k Patches of the RGB Space(K=140)



Top_k Patches of the Depth Space(K=140)

Figure 1: Results of the Case Study