ImmerseGen: Agent-Guided Immersive World Generation with Alpha-Textured Proxies Supplementary Material

Jinyan Yuan, Bangbang Yang, Keke Wang, Panwang Pan, Lin Ma, Xuehai Zhang, Xiao Liu, Zhaopeng Cui, and Yuewen Ma

In this supplementary material, we describe more details of our method in Sec. 1. Furthermore, we present additional experiments in Sec. 2. More qualitative results can be found in our supplementary video.

1 Implementation Details

1.1 Regional Prompts for Specific Landscape Elements

To integrate specific landscape elements like lakes at desired locations, we employ regional prompts [1] to guide the texture generation process. For terrains containing water bodies, we generate a panoramic water mask by rendering the terrain's water regions. This mask enables targeted texture synthesis, where water-specific prompts are applied to designated areas while maintaining natural ground textures in the remaining regions. The resulting textures seamlessly blend water features with the surrounding terrain.

1.2 Details of World Modeling Agents

We show system prompt examples for the world modeling agents based on ChatGPT-4o, including the asset selector, asset designer, asset arranger, and immersive enhancer. The asset selector analyzes contextual features from the base world panorama to systematically identify appropriate proxies, and the asset designer further enhances these assets with enriched descriptions and contextually coherent details, ensuring both environmental relevance and visual harmony within the scene, as illustrated in Fig. 3. To achieve accurate arrangement, the coarse arranger determines appropriate grid cell placements for assets by analyzing terrain and ecological factors, while the fine arranger refines placement by selecting optimal sub-cell positions within each cell for greater spatial detail and ecological realism, as illustrated in Fig. 4. The immersive enhancer comprises both an effect agent that infers environmental conditions to define dynamic animation parameters and a sound agent that selects and mixes appropriate soundtracks to enhance the overall visual atmosphere, as illustrated in Fig. 5.

Similarly, the LLM agent used in base world generation selects suitable terrain and enhances the user's original prompt by adding appropriate details and stylistic descriptions, based on the list of available terrain and illustrative prompt examples in system prompt, ensuring contextual consistency and improved diversity in the generated world.

1.3 Design of Dynamic Shaders

We implement three dynamic shader effects to enhance the realism of our natural environments. These effects are exposed as functional parameters that can be added by our immersive agent. *Cloud Movement:* The cloud movement uses a flow map to define overall cloud movement direction, combined with a noise texture where the R channel stores low-frequency noise for large-scale disturbances and the G channel stores high-frequency noise for detailed variations, creating layered cloud dynamics. *Screen-Space Rain:* The rain effect uses a spindle-shaped volume covering the camera range, with three baked textures. The depth map stores raindrop depth information across three channels (R: 0-5m, G: 5-10m, B: 10-15m), while the alpha map defines raindrop shapes and transparency, and the normal map enables light refraction simulation. This is combined with a panoramic depth map for scene interaction, with UV scrolling controlling the falling speed of three raindrop layers. *Water Ripples:* The effect leverages a procedurally generated texture with four channels — the R channel controls ripple propagation distance, the G and B channels store X and Y-axis normal gradients respectively, and the alpha channel contains animation time offset. Four layers of ripples are combined with a decay function to create natural-looking water surface interactions.

1.4 Ambient Sound Synthesis

Our ambient sound system builds upon a curated library of natural soundtracks labeled with descriptive tags. During 3D scene generation, we employ GPT-4o [2] to analyze the panorama rendered from the complete world, and select the three most appropriate audio tracks that match the scene's visual elements and atmosphere. The VLM also determines suitable volume levels for each track based on their relative importance to the scene. To ensure seamless playback in the immersive experience, we process the selected tracks with smooth crossfade transitions between their endings and beginnings, enabling continuous looping without noticeable interruptions.

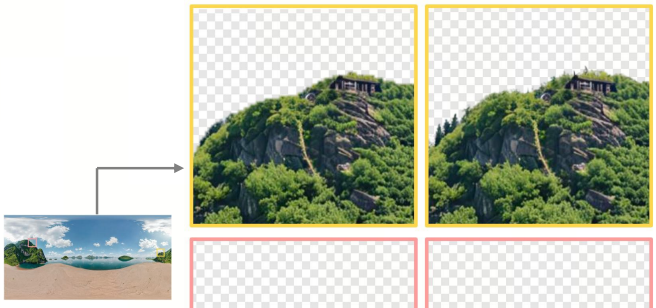
1.5 Bottom Map Enhancement with Repainting and Displacement

To improve both geometric and appearance detail in foreground explorable areas, we implement a bottom map refinement scheme through texture repainting and displacement mapping. Specifically, we create a dedicated UV map from a top-down perspective of the terrain, and refine it using image-to-image translation with ControlNet Tile model [4]. The refined texture is seamlessly blended back into the main terrain texture, ensuring smooth transitions while maintaining high-resolution details in explorable areas. For geometric enhancement with displacement, we estimate depth [3] from the top-down view and apply a high-pass filter to isolate fine-scale height variations, generating a displacement map that adds detailed rocky texturing to the terrain surface. This combined texture and displacement refinement significantly improves the visual fidelity of ground-level areas that users directly interact with.

1.6 Efficient Light Baking for Photorealism

To optimize performance while maintaining visual quality, we implement a panoramic light baking process. Specifically, we render a high-resolution panoramic shadow map of the entire scene. This map is then efficiently sampled using pre-calculated UV coordinates during runtime, eliminating the need for real-time lighting on VR devices. To this end, we export the entire environment to Unity using unlit materials while preserving photorealistic shadow effects.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx



(a) Panoramic Texture with Sky and Ground

(b) Original Matting (1K)

(c) Tile-based Matting (8K)

Fig. 1: We compare our tile-based matting with the baseline.

1.7 Execution Time for Scene Generation

The base world generation including terrain texture synthesis and projection takes about 3 minutes. The proxy asset arrangement process requires about 10 seconds per asset, and the layout generation (semantic grid-based analysis for hierarchical arrangement) of asset arranger takes about 1 minute. The immersive enhancements, including ambient sound integration and dynamic shader effects, are accomplished within 1 minute. The final post-processing stage, which includes panoramic light baking and scene asset export for game engines (Unity), requires 1-2 minutes.

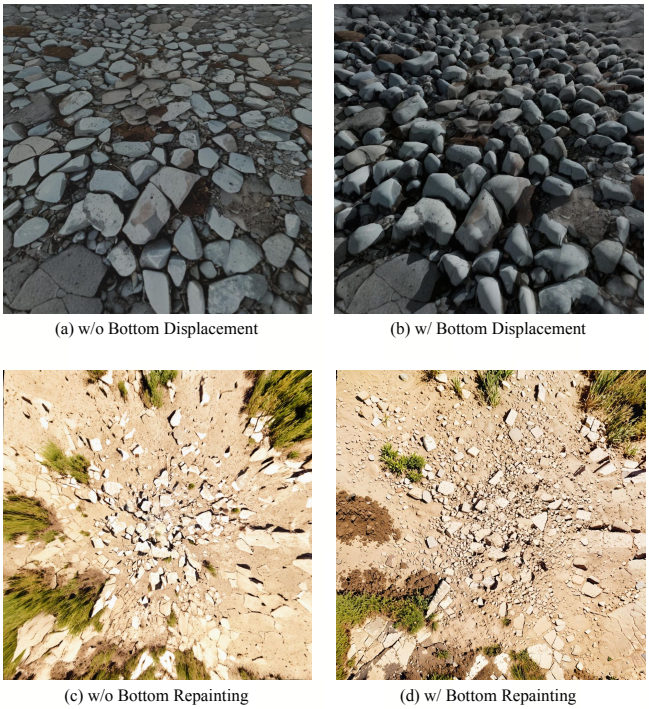
2 More Experiments

2.1 Improvement of Tile-based Matting

Our tile-based matting strategy significantly improves alpha matte details on large panoramic images. As shown in Fig. 1, this approach enables us to achieve crisp, detailed silhouettes of vegetation and structures against the sky, even when rendering low-poly terrain meshes from distant viewpoints. The enhanced alpha matte quality is particularly evident in the clear delineation of tree lines along mountain ridges and contours of man-made structures.

2.2 Improvement of Bottom Map Enhancement

We demonstrate the improvement in geometric details and texture quality achieved through the proposed bottom map displacement repainting in Fig.2. As shown in the first row of Fig.2, after performing displacement, the terrain exhibits enhanced geometric details with more pronounced rocky textures and surface variations. As shown in the second row of Fig.2, after repainting, the stretching and artifacts present in unobserved or distant areas are replaced with new content, notably enhancing the overall visual quality.



(a) w/o Bottom Displacement

(b) w/ Bottom Displacement

(c) w/o Bottom Repainting

(d) w/ Bottom Repainting

Fig. 2: We show the enhancement of bottom view before and after repainting and displacement.

Asset Selector

```
You are a professional 3D designer skilled in Geography, Botany, and Aesthetics.

Your task is to analyze a panorama image of a nature scene and determine suitable objects to add
from a provided JSON (objects list).

INPUT:
A list of objects in JSON format, where each element includes type, id, and description:
```json
{list}
```

OUTPUT:
A JSON array where each selected object is in the same format as the input.
- Includes a new "reason" field, explaining the rationale behind your selection choices.

FORMAT:
```json
[
  {{"type": "tree", "id": 1, "description": "pine tree', 'reason': '...'}},
 ...
]
```

INSTRUCTIONS:
- Analyze the panorama image to determine the season, soil, and terrain features.
- Select suitable objects based on the analysis.
- The exact number of objects to select will be specified by the user.
```

Asset Designer

```
You are a professional 3D designer specialized in modeling, visual aesthetics.

Your task involves analyzing a provided panoramic image and enhance relevant and visually fitting
details for each asset provided in the input JSON (objects list).

INPUT:
You will receive a panoramic image and a JSON list of assets/objects.

OUTPUT:
Return a JSON array containing one structured object per selected asset.
- Contains an enriched "description" field, detailing visual and aesthetic features based on your
analysis (e.g., size, color, texture, shape, season).
- Includes a new "reason" field, explaining the rationale behind your design choices (e.g., how the
enhancement contributes to cohesiveness or relevance within the scene).

EXAMPLE:
```json
[
  {"type": "tree", "id": 1, "description": "pine tree, autumn-season foliage, tall, conical shape,
warm orange-brown hues", "reason": "..."},
  ...
]
```

INSTRUCTION:
1. Analyze the provided panoramic image, looking carefully at visual and aesthetic properties such
as season, color tone, environmental context.

2. For each asset in the provided JSON list, determine the most suitable and visually consistent
details based on your image analysis.
```

Fig. 3: Prompt examples for Agent Selector and Designer.

Coarse Arranger

```
You are a professional 3D designer skilled in Geography, Botany, and Aesthetics.
Given a panorama of a nature scene, your task is to analyze the scene and determine suitable positions to plant some
objects.
The panorama is overlaid with green grids and labeled from 1 to 6 for rows and A to L for columns.
You are required to select cells in the panorama for each object.

INPUT:
A list of objects in JSON format, where each element includes type, id, and description.

OUTPUT:
A JSON array, containing the selected cells for each object.
Each element should contain the object information same as input, with added selected cells and reasons for choosing
each cell.

FORMAT:
    ```json
    [{{
        "id": 3,
        "type": 'leaves',
        "description": 'Fallen leaves, autumn',
        "cells": ['A6', 'C6', ...],
        "reasons": '...'
    }},...
    ]
    ```

GUIDELINES:
- Identify Features in the Scene:
    Analyze the terrain for flat or slightly sloped areas that can support plants.
    Detect existing vegetation to ensure a natural distribution of the new plants.
    Spot water sources such as lakes, rivers, or streams, as plants generally thrive near water.
    Avoid rocky areas and densely forested regions.
- Consider Distribution:
    Ensure plants have sufficient spacing for healthy growth.
    Create clusters in certain areas while maintaining sporadic spacing in others.
- Aesthetic and Ecological Factors:
    Enhance visual appeal by breaking symmetry.
    Consider biodiverse planting to support wildlife.
```

Fine Arranger

```
You are a professional 3D designer skilled in Geography, Botany, and Aesthetics.
Given some images of a nature scene, your task is to analyze the scene and determine suitable positions to plant
some objects.
The images are overlaid with green grids and labels. The green labels are image cell label and the small red labels
inside grids are sub-cell labels.
You are required to select sub-cells in the image for each object.

INPUT:
    An array of objects in JSON format, where each element includes type, id, and description and selected cell
labels.

OUTPUT:
    An array of JSON. Each element contains object information and selected cells and sub-cells.
    The cells is a nested JSON object, the key is the cell label corresponding to the input, and the value is a json
object with the selected sub-cell label as key and reasons to choose each sub-cell as value.

FORMAT:
        ```json
        [{
            "id": 3,
            "type": 'Small',
            "description": 'Fallen leaves, autumn',
            "cells": {'A6': {'F5': 'reason'
                             'A7': '...'},
                    'K5': {'H6': '...',
                             'B8': '...'},
                    ...},
        },
        ...]
        ```

GUIDELINES:
- Identify Features in the Scene:
    Analyze the terrain for flat or slightly sloped areas that can support plants.
    Detect existing vegetation to ensure a natural distribution of the new plants.
    Spot water sources such as lakes, rivers, or streams, as plants generally thrive near water.
    Avoid rocky areas and densely forested regions.
- Consider Distribution:
    Ensure plants have sufficient spacing for healthy growth.
    Create clusters in certain areas while maintaining sporadic spacing in others.
- Aesthetic and Ecological Factors:
    Enhance visual appeal by breaking symmetry.
    Consider biodiverse planting to support wildlife.
```

Fig. 4: Prompt examples of Agent Arranger (including Coarse Arranger and Fine Arranger).

Effect Agent

```
You are a professional 3D designer with expertise in geography and meteorology. Given the provided scene panoramic
image, first analyze and determine the environmental conditions amd then provide optimal parameter values for
realistic nature animation effects (including water, clouds, and rain) that best represent the observed
environmental conditions.
Refer to the following parameter descriptions of effects:
[
    ["Rain_Speed",
     "Scroll speed of the rain texture over time",
     "[0.0, 10.0]",
     "Provide clearly reasoned and realistic suggestions based on observation"],

    ["Water_RippleTiling",
     "Scales UV coordinates to control ripple density. Lower values produce fewer, larger ripples.",
     "[0.01, 10.0]",
     "1.0: standard ripple scale\n0.5: double ripple size (fewer but larger ripples)\n2.0: dense, smaller ripples"],
     ["Bird_Density",
     "Controls the density of birds visible in the scene.",
     "[0.0, 1.0]",
     "0.0: no birds\n0.25: occasional birds\n0.5: moderate bird presence\n1.0: dense flocking birds"],
     ...
]
OUTPUT:
Present your answer strictly in the following structured JSON format:
Suggested parameters as nested JSON, where each parameter includes a "value" and a textual "reason" clearly
explaining your parameter choice.

INSTRUCTION:
1. You should determin the parameter by inferring the enviroment instead of simply identify visible elements since
the image contains only basic terrain.

EXAMPLE:
```json
{{
        "Water_RippleTiling": {{
            "value": 1.3,
            "reason": "..."
        }},
        ...
}}
```
```

Sound Agent

```
You are a professional scene audio designer. Your task is to select suitable ambient soundtracks from an existing material
library to mix audio that matches the given scene image.

You can choose one or multiple audio files from the existing library for mixing.
For instance, if a single audio file suffices for the scene, you can use just that. However, if the scene is more complex,
multiple audio files will be needed for mixing.

Below is our audio library, where the file names represent the potential audio content.

{
AUDIO_LIST_PLACEHOLDER
}

Now, based on the provided panoramic image, you need to first describe the scene, and then give a corresponding audio mixing
plan (including audio file names and volume information) and provide explanations or reasons.

Note:
1. Usually, 1-3 tracks are enough for a scene. Too many repetitive bird calls might be confusing.
2. The audio file names need to match the file names in the audio library, otherwise, the corresponding audio files cannot be
found.

Example of output:
```
{
    "scene_description": "This scene is a forest scene, containing elements such as trees, water, etc., with an overall
atmosphere of tranquility and mystery.",
    "audio": [
        {
            "filename": "night_wind_in_forest_V2.wav",
            "volume": 0.5,
            "descriptions": "This audio features the sound of wind in a jungle at night, suitable for conveying the tranquility
and mystery of the given jungle scene."
        },
        {
            "filename": "rainy_V2.wav",
            "volume": 0.5,
            "descriptions": "This audio features the sound of rain, suitable for conveying the dampness and chill of the given
scene."
        }
    ]
}
```
```

Fig. 5: Prompt examples of Effect Agent and Sound Agent for immersion enhancement.

REFERENCES

[1] 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. *CVPR*, 2023. 1

[2] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. Accessed: 2024-10-01. 1

[3] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 1

[4] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. 1