MajutsuCity: Language-driven Aesthetic-adaptive City Generation with Controllable 3D Assets and Layouts

Supplementary Material

Overview

Due to the strict page limit of the main submission, we provide additional implementation details and extensive qualitative results in this supplementary material to support the findings presented in the paper. The content is organized as follows:

- **Section 1: Related Work.** We present an extended review and discussion of the relevant literature, specifically focusing on Layout Generation, Image-to-3D Generation, and City Scene Generation. This section highlights key technical inspirations that inform the design of our framework.
- **Section 2: Evaluation Metrics.** We provide detailed definitions and evaluation protocols for all metrics used in our experiments, including an additional explanation of the Layout Generation metrics and a comprehensive description of our proposed VLM-based assessment (AQS and RDR).
- **Section 3: Additional Qualitative Results.** This section presents extensive visual results to demonstrate the versatility of our framework, including (1) prompt templates used in the Scene Design module, (2) examples of text-aligned layout generation, (3) diverse city generation results across different aesthetic styles, (4) galleries of 3D assets, seamless materials, and skyboxes from MajutsuDataset, (5) comparisons of seamless materials generated by Qwen-Image and our fine-tuned Qwen-Image variant, (6) prompt templates used for the VLM-based evaluation, and (7) the user interface employed for evaluating generated city scenarios.
- **Section 4: Limitations and Future Work.** We discuss the potential limitations of the current MajutsuCity framework and outline several promising future research directions.

1. Related Work

1.1. Layout Generation

Compared to general scene layouts [9, 16], urban layouts exhibit significantly higher complexity due to the richer semantic categories and irregular geometric topologies present in real-world cities. While vector-based layout generation methods such as BlockPlanner [28] and GlobalMapper [10] provide a structured formulation, they often suffer from limited semantic representation and struggle to model complex, fine-grained spatial patterns.

In contrast, recent mask-based generation methods [4, 5, 19, 27] offer better scalability and geometric fidelity, capturing detailed spatial boundaries while maintaining efficient inference. However, these methods typically lack intuitive user control, specifically in expressing high-level design intent through natural language.

To address this limitation, we train a language-guided urban layout generation model that aligns fine-grained spatial mask synthesis with user-provided textual descriptions, effectively bridging the gap between high-fidelity mask generation and user-intent controllability.

1.2. Image-to-3D Generation

Recent advances in image generation have provided a new direction for 3D content creation, fundamentally reshaping the synthesis of high-quality 3D assets. Early methods were often limited by low-fidelity geometry and low-resolution textures [15, 20, 21, 23], but current 3D generation frameworks increasingly leverage powerful 2D visual priors to improve the consistency of geometry and textures. In particular, many state-of-the-art works employ visual foundation model embeddings such as DINOv2 [22] to extract rich semantic and textural representations, coupled with Vision Transformer (ViT) [6] architectures to guide and regularize 3D geometry generation.

The rapid evolution of a series of SOTA and open-source models (e.g., Trellis [26], Step1X-3D [18], and Hunyuan3D 2.0 / 2.1 [14, 31], as well as commercial tools (e.g., Tripo, Rodin [25], Meshy, Hunyuan3D 2.5 [17], and Hitem3D) have demonstrated that photorealistic and high-detail 3D assets are now not only feasible but increasingly reliable.

Motivated by these advances, this work explores a principled integration of advanced 3D asset generation models as core components into a city-scale pipeline, enabling large-scale, controllable, and aesthetic-adaptive 3D city generation within a unified, reproducible framework.

1.3. City Scene Generation

In urban scene generation, existing methods such as InfiniCity [19], CityDreamer [27], and Persistent Nature [2] have demonstrated the ability to generate large-scale 3D scenes. However, they often rely on implicit or neural representations, resulting in two critical bottlenecks: (1) geometric artifacts and multi-view inconsistency, which stem from the inherent ambiguity of implicit fields, and (2) the absence of explicit, editable object-level structures, making them unsuitable for downstream applications that require precise in-

teraction, editing, and simulation compatibility.

On the other hand, Procedural Content Generation (PCG)-based techniques can produce highly structured cities [7, 30, 32], but they follow a fundamentally Retrieve-and-Place paradigm. As a result, the diversity and expressiveness of the generated scenes are strictly limited by the scale, style coverage, and quality of the predefined asset libraries, restricting their ability to generalize to novel or stylistically distinctive demands.

Recent advanced works in indoor scene generation have successfully validated a new path: combining the powerful priors of 2D vision models with on-demand 3D object generation to achieve object-level, controllable scene synthesis [3, 13, 29]. Inspired by this, we aim to scale this object-centric generative paradigm from the indoor level to the macroscopic urban level, enabling a unified framework that is not only controllable and editable, but also adaptive to diverse aesthetic styles, and addressing the core limitations of prior approaches in urban scene generation.

2. Metrics

2.1. Layout Generation Metrics

Fréchet Inception Distance (FID) [12]. Evaluates the distribution similarity between generated images and real images in the Inception-v3 feature space. It is calculated as:

$$\text{FID} = ||\mu_r - \mu_g||_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \quad (1)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) represent the mean and covariance of the real and generated feature distributions, respectively. Lower FID scores indicate higher visual realism and closer proximity to the real data distribution.

Kernel Inception Distance (KID) [1]. Also measures distribution similarity in the feature space, but is more robust to small sample sizes. It computes the squared Maximum Mean Discrepancy (MMD) between feature representations using a polynomial kernel k:

$$\begin{aligned}\text{KID} = \mathbb{E}_{x,x'\sim P_g}[k(x,x')] + \mathbb{E}_{y,y'\sim P_r}[k(y,y')] \\ -2\mathbb{E}_{x\sim P_g, y\sim P_r}[k(x,y)]\end{aligned} \quad (2)$$

where P_g and P_r denote the generated and real distributions. Unlike FID, KID is an unbiased estimator, making it particularly suitable for datasets with fewer samples.

Inception Score (IS) [24]. Assesses the quality and diversity of the generated images.It calculates the KL divergence between the conditional class distribution $p(y|x)$ and the marginal class distribution $p(y)$:

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim P_g}[D_{KL}(p(y|x)||p(y))]\right) \quad (3)$$

A higher IS indicates that the model generates clear, distinct objects (low entropy for $p(y|x)$) while maintaining diversity across all classes (high entropy for $p(y)$).

2.2. Scene Generation Metrics

To comprehensively and rigorously evaluate the quality of the generated scenes, we define four core evaluation dimensions:

- **Structural and View Consistency (SVC):** Evaluates the geometric soundness and multi-view coherence of the generated scene.
- **Scene Richness and Complexity (SRC):** Measures the diversity, density, and structural complexity of scene elements.
- **Material and Texture Fidelity (MTF):** Assesses the realism and detail fidelity of object materials and textures.
- **Lighting and Atmosphere (LA):** Evaluates the plausibility of lighting, color harmony, and the resulting sense of immersion.

These dimensions form the basis of our two GPT-driven evaluation protocols: **Absolute Quantitative Scoring (AQS)** and **Relative Dimension Ranking (RDR)**. As shown in Figure 16 and Figure 17, we present the detailed prompts used for both AQS and RDR, which explicitly outline the precise definitions and key assessment criteria for each of these four core dimensions. To mitigate the subjective bias inherent in absolute scoring, our **Relative Dimension Ranking (RDR)** employs the TrueSkill ranking system [11]. TrueSkill models the skill of each method as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, where μ represents the average skill (score) and σ represents the uncertainty.

In our pairwise comparison setting, for a given dimension (e.g., SVC), let the skill of method A be $s_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, and method B be $s_B \sim \mathcal{N}(\mu_B, \sigma_B^2)$. The performance difference is modeled as $d = s_A - s_B \sim \mathcal{N}(\mu_A - \mu_B, \sigma_A^2 + \sigma_B^2 + 2\beta^2)$, where β represents the inherent noise in performance.

The probability that method A outperforms method B is given by:

$$P(A > B) = \Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2 + 2\beta^2}}\right) \quad (4)$$

where Φ is the cumulative distribution function of the standard normal distribution.

After observing a comparison outcome (e.g., A wins), the posterior skill distributions for both methods are updated using Bayesian inference to minimize the prediction error for future comparisons. We initialize all methods with $\mu = 25$ and $\sigma = 25/3$, and iteratively update these parameters based on the pairwise results from both human and GPT evaluators until convergence. The final reported RDR score is the converged mean skill μ.

3. Additional Qualitative Results

- **Figure 1:** Illustration of the analytical prompt template employed by the Scene Design module. This template

allows the model to interpret user intent and decompose it into instructional text for subsequent generation stages.
- **Figure 2:** Qualitative results of text-guided urban layout generation produced by our trained layout generation model.
- **Figure 3**: Comparison of text-guided scene generation quantitative results between Majutsucity and Syncity[8].
- **Figure 4:** Representative examples demonstrating the fundamental capabilities and workflow of the MajutsuAgent.
- **Figures 5-9:** Visualizations of 3D model assets within the MajutsuDataset, generated by five distinct commercial frameworks.
- **Figures 10-13:** Excerpts from the MajutsuDataset featuring seamless texture maps and skybox environments.
- **Figure 14-15:** Comparative analysis of Qwen-Image generation results before and after fine-tuning on our dataset, validating the utility and efficacy of the proposed data.
- **Figures 16-17:** The prompt templates utilized for the AQS (Automated Quality Score) and RDR (Reference Deviation Rate) evaluation protocols.
- **Figures 18-19:** The user interface of the evaluation platform employed for conducting the AQS and RDR assessments.

4. Limitations and Futurk Work

Despite the significant advancements that MajutsuCity introduces to controllable urban generation, several limitations remain.

- **Dependency on Prompt Logic Consistency:** A critical limitation lies in the sensitivity of our Layout Generation module to the logical coherence of the input prompts. Since our pipeline is strictly hierarchical, if the initial Scene Design stage generates contradictory spatial instructions or geometrically impossible layouts, these errors propagate downstream. This can result in implausible road networks or building placements that negatively impact the final scene assembly, as the subsequent modules faithfully follow the flawed structural guidance.
- **Complex Geometry Collisions:** While our layout-guided approach significantly reduces object intersection compared to scatter-based methods, minor collision artifacts may still occur in extremely high-density areas. This is particularly evident when generating assets with highly irregular footprints or overhanging structures that extend beyond their 2D semantic masks.
- **Inference Latency:** Due to the multi-stage nature of our pipeline—which involves sequential diffusion processes for layouts, height maps, and individual 3D assets—the total inference time for a city-scale scene is higher than end-to-end neural rendering approaches. This currently limits the framework's applicability in real-time generation scenarios without pre-computation.
- **Shape Controllability:** Although we introduced two schemes to strengthen the constraints on the building shape, it is difficult to ensure high fidelity while maintaining geometric rationality when faced with highly complex or irregular building topologies.

4.1. Future Work

Building upon the current framework, our future research will focus on the following directions:

- **Enhanced Spatial Reasoning:** To address the dependency on prompt logic, we plan to integrate a dedicated "Spatial Reasoning & Verification" module powered by advanced reasoning models (e.g., Chain-of-Thought prompting). This module will pre-validate user instructions and automatically resolve logical conflicts before layout generation begins, ensuring structural plausibility.
- **Multi-view Building Generation:** While our current approach relies on single-view inputs for shape control, we plan to incorporate multi-view consistency constraints in future work. This extension aims to better preserve the global geometric structure while simultaneously enhancing the fidelity of local details.
- **Physics-Aware Assembly:** We aim to incorporate a lightweight physics engine during the Scene Generation phase. By applying rigid body dynamics and collision detection algorithms, we can automatically adjust the placement and orientation of assets to resolve overlaps, ensuring a physically consistent environment.
- **Dynamic Urban Environments:** Moving beyond static scenes, we intend to extend MajutsuCity to support dynamic elements. This includes procedural traffic simulation, pedestrian flow modeling, and time-varying weather systems, transforming the generated static cities into living, breathing digital twins suitable for interactive applications and autonomous driving simulations.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2

[2] Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, and Noah Snavely. Persistent nature: A generative model of unbounded 3d worlds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20863–20874, 2023. 1

[3] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. *arXiv preprint arXiv:2410.07408*, 2024. 2

[4] Jie Deng, Wenhao Chai, Junsheng Huang, Zhonghan Zhao, Qixuan Huang, Mingyan Gao, Jianshu Guo, Shengyu Hao, Wenhao Hu, Jenq-Neng Hwang, et al. Citycraft: A real crafter for 3d city generation. *arXiv preprint arXiv:2406.04983*, 2024. 1

[5] Jie Deng, Wenhao Chai, Jianshu Guo, Qixuan Huang, Junsheng Huang, Wenhao Hu, Shengyu Hao, Jenq-Neng Hwang, and Gaoang Wang. Citygen: Infinite and controllable city layout generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1995–2005, 2025. 1

[6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] Yinglin Duan, Zhengxia Zou, Tongwei Gu, Wei Jia, Zhan Zhao, Luyi Xu, Xinzhu Liu, Yenan Lin, Hao Jiang, Kang Chen, et al. Latticeworld: A multimodal large language model-empowered framework for interactive complex world generation. *arXiv preprint arXiv:2509.05263*, 2025. 2

[8] Paul Engstler, Aleksandar Shtedritski, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Syncity: Training-free generation of 3d worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27585–27595, 2025. 3, 8

[9] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1004–1014, 2021. 1

[10] Liu He and Daniel Aliaga. Globalmapper: Arbitrary-shaped urban layout generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 454–464, 2023. 1

[11] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[13] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23646–23657, 2025. 2

[14] Team Hunyuan3D, Bowen Zhang, Chunchao Guo, Haolin Liu, Hongyu Yan, Huiwen Shi, Jingwei Huang, Junlin Yu, Kunhong Li, Penghao Wang, et al. Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. *arXiv preprint arXiv:2509.21245*, 2025. 1

[15] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1

[16] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9895–9904, 2019. 1

[17] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025. 1

[18] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 1

[19] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22808–22818, 2023. 1

[20] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 1

[21] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[23] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 2

[25] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong

Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 1

[26] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 1

[27] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9666–9675, 2024. 1

[28] Linning Xu, Yuanbo Xiangli, Anyi Rao, Nanxuan Zhao, Bo Dai, Ziwei Liu, and Dahua Lin. Blockplanner: City block generation with vectorized graph representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5077–5086, 2021. 1

[29] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics (TOG)*, 44(4):1–19, 2025. 2

[30] Shougao Zhang, Mengqi Zhou, Yuxi Wang, Chuanchen Luo, Rongyu Wang, Yiwei Li, Zhaoxiang Zhang, and Junran Peng. Cityx: Controllable procedural content generation for unbounded 3d cities. *arXiv preprint arXiv:2407.17572*, 2024. 2

[31] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 1

[32] Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng, and Zhaoxiang Zhang. Scenex: Procedural controllable large-scale scene generation. *arXiv preprint arXiv:2403.15698*, 2024. 2

Scene Design prompt

ROLE

You are an expert 3D Asset and Scene Prompt Engineer. Your task is to analyze a user's scene description, infer its single, core visual style/theme, and then apply that theme consistently to generate ten highly specific sub-prompts for a 3D/game development pipeline.

TASK

Given a short user scene description, produce ten concise sub-prompts: ***layout, building, tree, altree, lamp, ground, grass, road, water,*** and ***sky***. Output English only.

STYLE & THEME INFERENCE

You must first read the <USER_SCENE> and infer its single, dominant visual style, theme, or aesthetic (e.g., "Ghibli anime," "spooky Halloween," "medieval fantasy," "photorealistic modern," "cyberpunk"). This single inferred style must be used to generate all ten sub-prompts to ensure visual consistency.

HARD CONSTRAINTS & FORMATTING BY FIELD

You must follow these templates exactly, filling in the content based on your inferred style and the <USER_SCENE> description.

1. **Layout**
   - **Task**: Infer a top-down, semantic map description based on the user's scene.
   - **Content**: Describe road networks (e.g., "radial road network (red)"), building clusters and patterns (e.g., "rectangular buildings (yellow) clustered in linear patterns"), and large zones for vegetation/water (e.g., "large, contiguous vegetation patch (green)... bordering a large water body (blue)").
2. **Building:**
   - **Task:** Generate a prompt to edit a base image into the inferred style.
   - **Format:** Must start with Edit the input image only; keep the existing camera angle and composition. Convert the building into a single, [inferred_style_and_subject, e.g., 'spooky haunted building in a Halloween style'].
   - **Content:** Describe 2-3 key materials (e.g., "Spooky materials (dark weathered wood, crumbling stone, wrought iron)") and 3-5 rich textures (e.g., "Rich textures (spider webs, carved pumpkin faces, creeping vines, eerie fog, bat silhouettes, cracked facades)").
   - **Suffix:** Must end with the exact text: , Pure white background (#FFFFFF), no ground/floor/base, no surrounding objects, uniform bright lighting, no background shadow. Ultra-clean silhouette, crisp edges, high detail.
3. **Tree:**
   - **Task:** Generate an isolated asset prompt for a standard tree.
   - **Format:** Must start with A ISOMETRIC [inferred_attributes, e.g., 'autumn tall'] tree designed in the [inferred_style, e.g., 'style of the Ghibli anime in Japan'].
   - **Suffix:** Must end with the exact text: , with white background and without ground plane, no ground, no terrain, isolated object
4. **Altree:**
   - **Task:** Generate a contextualized asset prompt for a tree with a base.
   - **Format:** Must start with A ISOMETRIC [inferred_attributes, e.g., 'autumn roadside tall'] tree.
   - **Content:** Must include a description of its base or immediate surroundings (e.g., surrounded by a small circular stone border at its base, with some flowers).
   - **Style:** Must include the inferred style (e.g., designed in the Ghibli anime in Japan).
   - **Suffix:** Must end with the exact text: , on a white background.
5. **Lamp:**
   - **Task:** Generate an isolated asset prompt for a street lamp.
   - **Format:** Must start with A ISOMETRIC street lamp designed in the [inferred_style, e.g., 'style of the Ghibli anime in Japan'].
   - **Suffix:** Must end with the exact text: , with white background and without ground plane.
9. **ground, grass, road, water:**
   - **Task:** Generate PBR texture prompts.
   - **Format:** Must start with A seamless, tileable PBR texture of [inferred_style_and_subject, e.g., 'medieval stone ground' or 'old, cracked asphalt road']:
   - **Content:** Provide 3-5 descriptive clauses about its appearance, wear, finish, and roughness (e.g., "large, roughly cut, grey stone bricks... weathered with age... patches of moss... matte finish, high roughness").
   - **Suffix:** Must end with the exact text: ; seamless tileable PBR texture, uniform lighting, no perspective, no seams.
10. **Skybox:**
   - **Task:** Generate a 360-degree skybox prompt.
   - **Format:** Must start with A [inferred_style_and_description, e.g., 'realistic, clear expanse of a medieval-era sunset'] sky,.
   - **Content:** Describe the atmospheric details, color gradients (horizon-to-zenith), and any features (e.g., "bathed in soft, muted golden hues... grades smoothly from a faint, clear amber... no clouds are present").
   - **Suffix:** Must end with the exact text: — rendered in [inferred_style_tag, e.g., 'photorealistic HDRI']; skybox equirectangular 360° (2:1) texture.

Input <USER_SCENE>

OUTPUT FORMAT

Return ONLY a JSON object (no extra text before or after):

```
{
  "layout": "<Generated layout description>",
  "building": "<Generated building edit prompt>",
  "tree": "<Generated isolated tree prompt>",
  "altree": "<Generated contextualized tree prompt>",
  "lamp": "<Generated isolated lamp prompt>",
  "ground": "<Generated ground texture prompt>",
  "grass": "<Generated grass texture prompt>",
  "road": "<Generated road texture prompt>",
  "water": "<Generated water texture prompt>",
  "sky": "<Generated skybox prompt>"
}
```

Figure 1. The analysis prompt template used in the Scene Design module

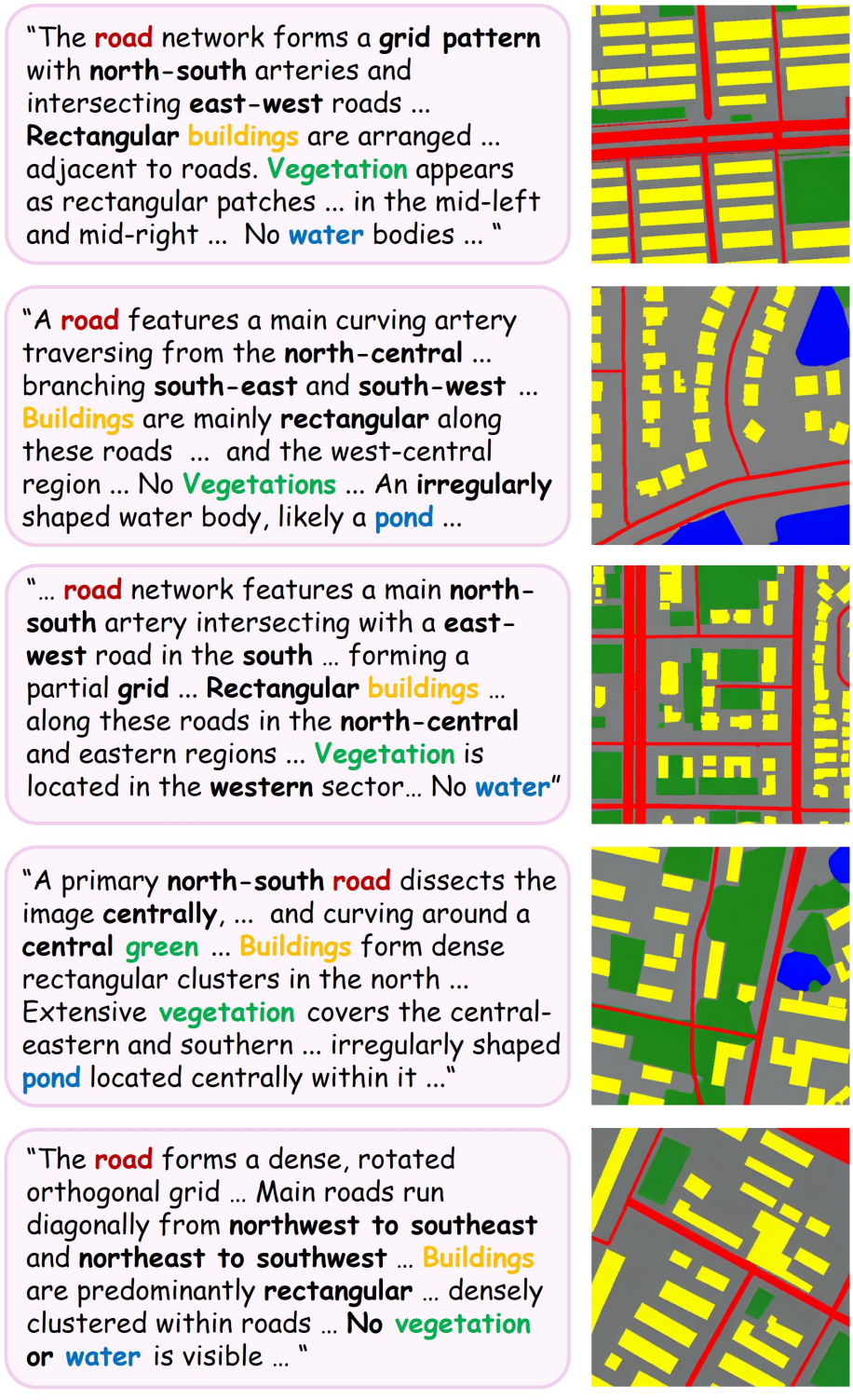


Figure 2. Example of text-guided city layout generation results

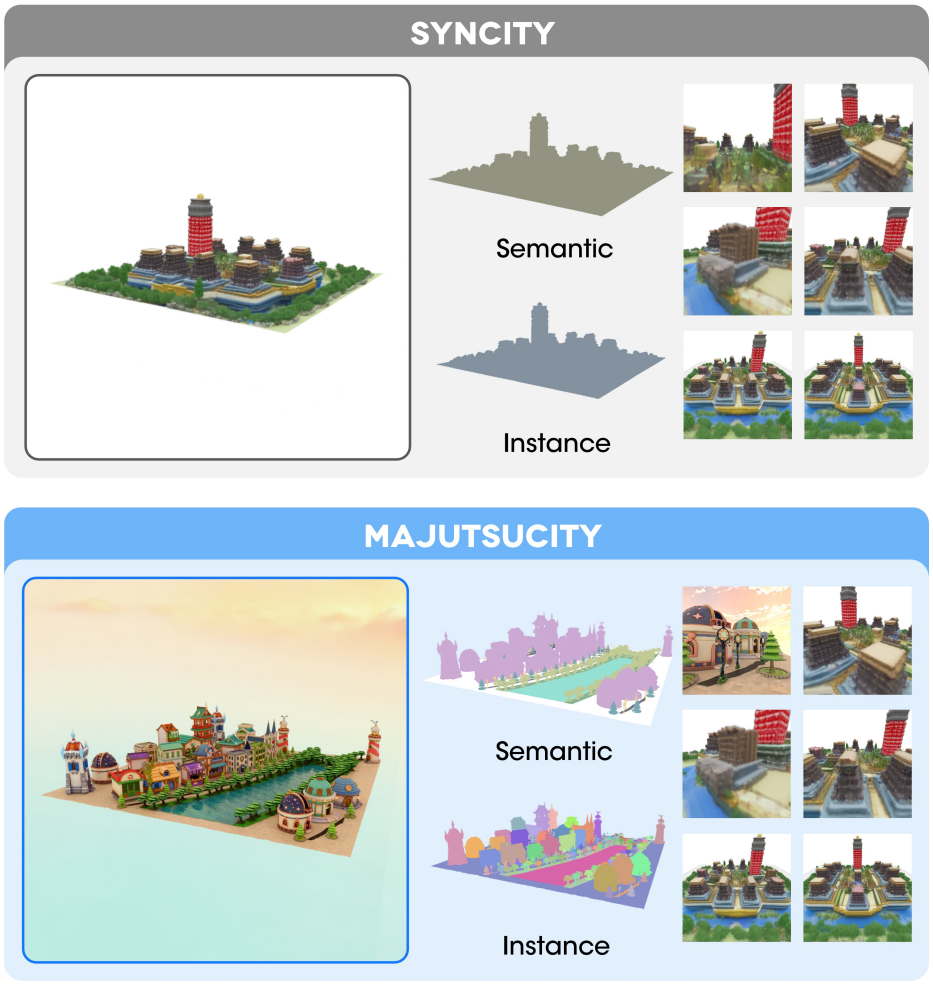


Figure 3. Compared with SynCity[8]

Figure 4. Examples of MajutsuAgent, includes five operations (Edit, Add, Delete, Move, Replace).

Figure 5. 3D building models generated by Hunyuan3D from MajutsuDataset

Figure 6. 3D building models generated by Meshy from MajutsuDataset

Figure 7. 3D building models generated by Hitem3D from MajutsuDataset

Figure 8. 3D building models generated by Tripo3D from MajutsuDataset

Figure 9. 3D building models generated by Hyper3D from MajutsuDataset

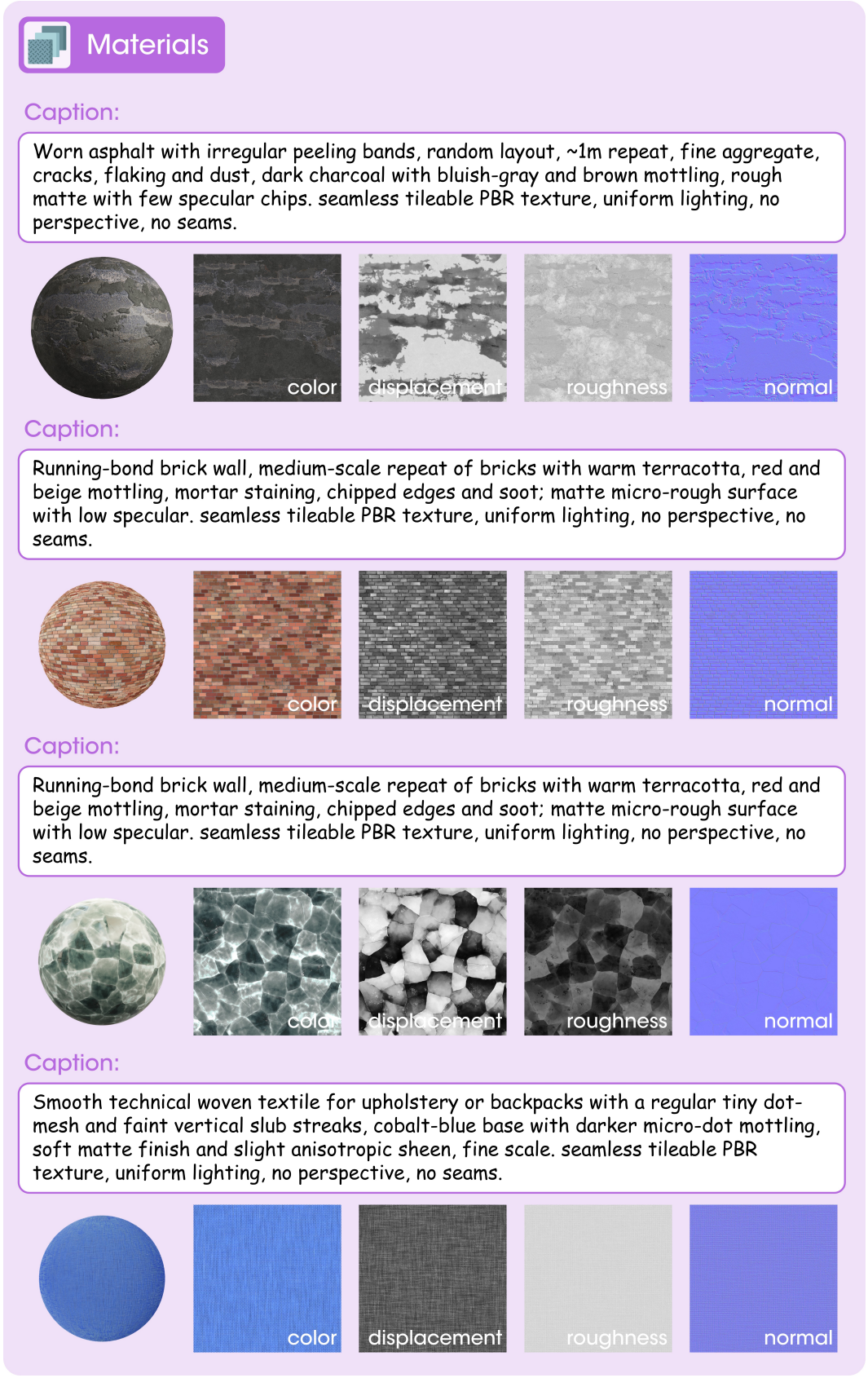


Figure 10. Material examples from MajutsuDataset

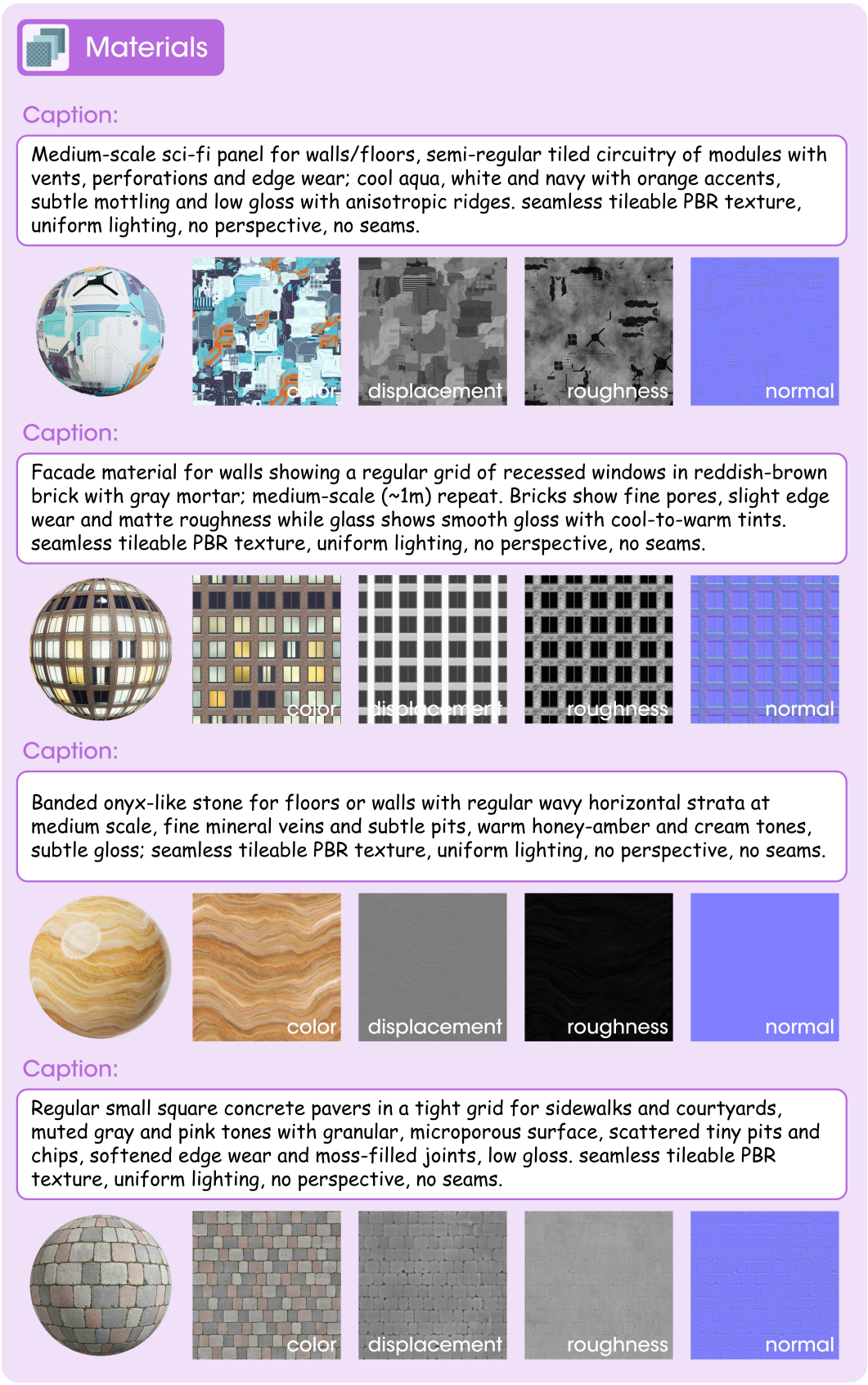


Figure 11. Material examples from MajutsuDataset

Skybox

Caption:

A wide, low horizon dominated by a continuous dark stratiform cloud bank stretches across the panorama, forming a nearly unbroken silhouette with a single broad breach at center where an intense sunlit cavity punctures the deck; above that breach compact, congested cumulus turrets and broken mid-level puffs give way to thin, sheared cirrus and small convective fragments radiating outward, creating layered cloud morphology with roughly sixty percent low-deck coverage and scattered higher elements; the color gradient moves from deep ultramarine at the zenith through cobalt and desaturated sky-blue into a concentrated warm band of ochre, peach and rose around the central glow, while the undersides of clouds carry violet and magenta inflections over a very dark navy lower hemisphere; lighting is strongly backlit with the sun obscured by cloud (sun not directly visible), producing hard rim lighting on cloud edges, soft ambient fill and subtle crepuscular scattering that suggests calm, post-storm clearing and high visibility toward the illuminated gap, no night features visible — rendered in [painterly digital matte]; skybox equirectangular 360° (2:1) texture.

Caption:

A nearly flat, mirror-like open-water horizon dominates the panorama with a low diffuse mist line and a thin, broken stratiform band at the distance that produces a pronounced vertical specular bloom centered on the horizon where the sun is obscured by cloud; the upper sky is sculpted by a large, luminous cirrostratus/altostratus mass with dense, soft-edged core and sweeping, high-altitude cirrus filaments and wisps radiating outward, while a few thin low stratiform fragments ride the horizon; the color grades from a deep cobalt blue at the zenith through saturated cerulean and pale cyan into a warm, desaturated lemon-ivory glow around the bright center, with visible airlight softening contrast; overall lighting is diffuse and backlit with subtle rim illumination on cloud edges and a calm specular reflection on the water, sun not directly visible, no nocturnal features present, overall weather mood calm and post-dawn clear — rendered in [photorealistic HDRI]; skybox equirectangular 360° (2:1) texture.

Caption:

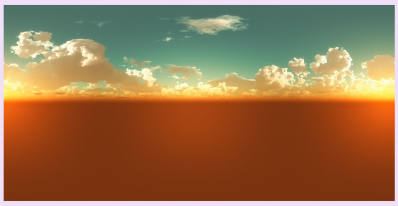
A clean, uninterrupted flat horizon with a near-perfect mirror reflection below suggests an open, glassy sea or perfectly reflective plane with no landforms or skyline visible, the distant horizon punctuated only by a low, thin cloud bank and the vertical specular streak of a low sun; cloud morphology is a broken deck of mid- and low-level cumulus and stratocumulus clusters with larger puffy cores interspersed with extensive fine altocumulus/stratocumulus "mackerel sky" texture and a few high, wispy cirrus flecks, overall coverage roughly half to two-thirds of the dome and layered in depth; the zenith transitions from deep cobalt-blue through vivid cyan toward a concentrated warm amber and pale yellow band around the sun, then washes into soft rose and mauve toward the horizon and its mirrored counterpart, with the lower hemisphere graduating to saturated violet and indigo in the nadir; lighting is a low, near-central sunrise/sunset producing soft diffuse backlight and gentle rim illumination on cloud undersides with faint crepuscular diffusion and no visible halo or lens flare; no night features such as stars, Milky Way, aurora, or city domes are present, and the overall weather mood is calm and tranquil with high atmospheric clarity after or before a calm period — rendered in [photorealistic HDRI]; skybox equirectangular 360° (2:1) texture.

Figure 12. Skybox examples from MajutsuDataset

Skybox

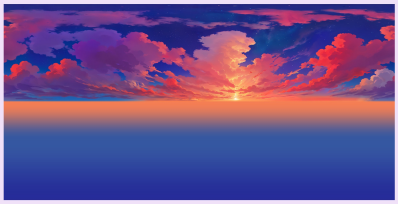
Caption:

A continuous, featureless low-lying orange aerosol layer forms a perfectly flat horizon across the panorama, obscuring ground detail and giving the impression of an endless open plain or sea; above that thin haze a broken band of large cauliflower cumulus and several congestus towers march around the horizon with crisp, sunlit tops and softer, shadowed bases, punctuated by a few thin high cirrus streaks closer to zenith, overall cloud coverage moderate and concentrated near the horizon with isolated puffs lofted higher; the sky shifts from a muted teal-green at zenith through cyan into a bright amber and intense orange at the horizon, showing strong scattering where warm light bleeds upward; the sun itself is not visible but strong low-angle warm backlighting from near-horizon glows at the lateral edges produces pronounced rim lighting and high-contrast cloud shadows, suggesting late-afternoon/early-sunset calm with reduced horizontal visibility in the haze and no night features present — rendered in [photorealistic HDRI CGI]; skybox equirectangular 360° (2:1) texture.



Caption:

Flat, uninterrupted ocean horizon bisects the panorama in a calm, mirrorlike expanse with a low, centered sun resting on the rim and a narrow band of specular warmth reflecting across the sea; dominating the sky are large, billowing cumulus congestus and compact stratocumulus banks that fan outward from the sun into towering vertical columns and flattened bases, interleaved with thinner high veils and wispy cirrus streaks toward the zenith, creating a broken, multi-layered cloud deck; the clouds display saturated rim lighting — molten gold and tangerine at sunward faces transitioning to magenta, violet and deep fuchsia in the shadowed lobes; the sky color grades from a deep indigo-cobalt studded with a scattering of stars and a faint Milky Way-like band at zenith down through ultramarine and cerulean to a glowing peach and amber band immediately above the horizon, producing hard rim light on cloud edges, crepuscular rays and a clear, tranquil sunset-to-twilight mood with no moon visible — rendered in [stylized digital painting]; skybox equirectangular 360° (2:1) texture.



Caption:

A wide, open-ocean horizon dominates the panorama, the sea rendered as a flat, mirror-smooth plane with virtually no surface texture and a narrow, centralized solar glow sitting at low elevation just below the horizon; above it a dramatic, symmetric arch of large puffy cumulus and congestus-like masses stretches left to right like a crown, many with sharply rim-lit pink-orange tops and cooler lavender bases, interspersed with elongated cirrus streaks and small detached cumulus fragments giving roughly 30–40% mid-to-upper cloud coverage, while the sky color grades from deep indigo at the zenith dotted with faint star points, through cobalt and bright turquoise, into a concentrated amber–coral band at the horizon; lighting is warm, low-angle backlight that produces soft diffuse illumination and strong colored rim light on cloud edges, the sun itself not visible, no precipitation or virga evident, overall mood calm and tranquil at twilight-dawn transition with clear atmospheric visibility — rendered in [stylized digital painting]; skybox equirectangular 360° (2:1) texture.

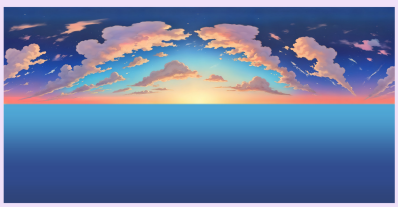


Figure 13. Skybox examples from MajutsuDataset

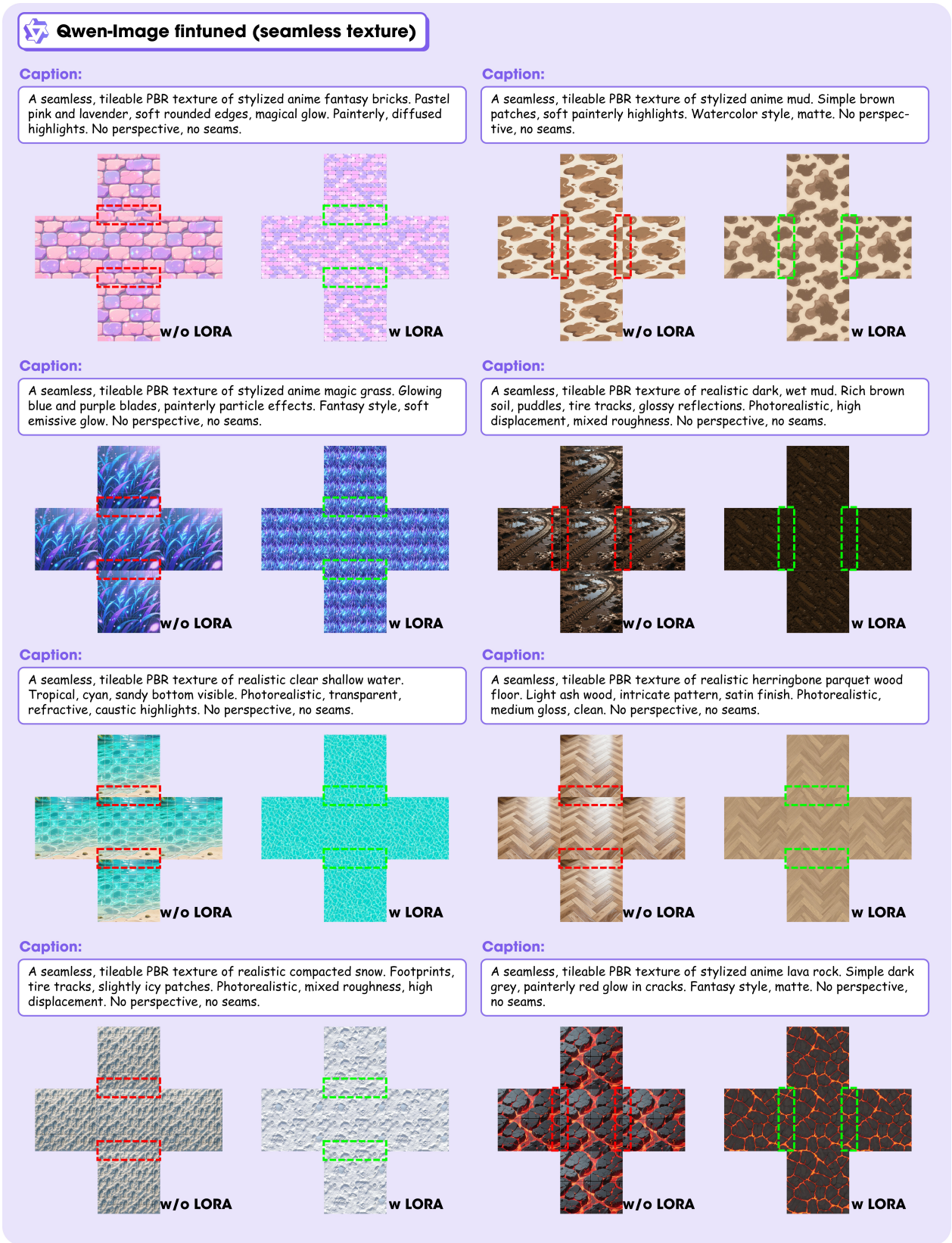


Figure 14. Comparison of the Texture generated models from the original Qwen-Image model and the Qwen-Image model fine-tuned using MajutsuDataset-Materials.

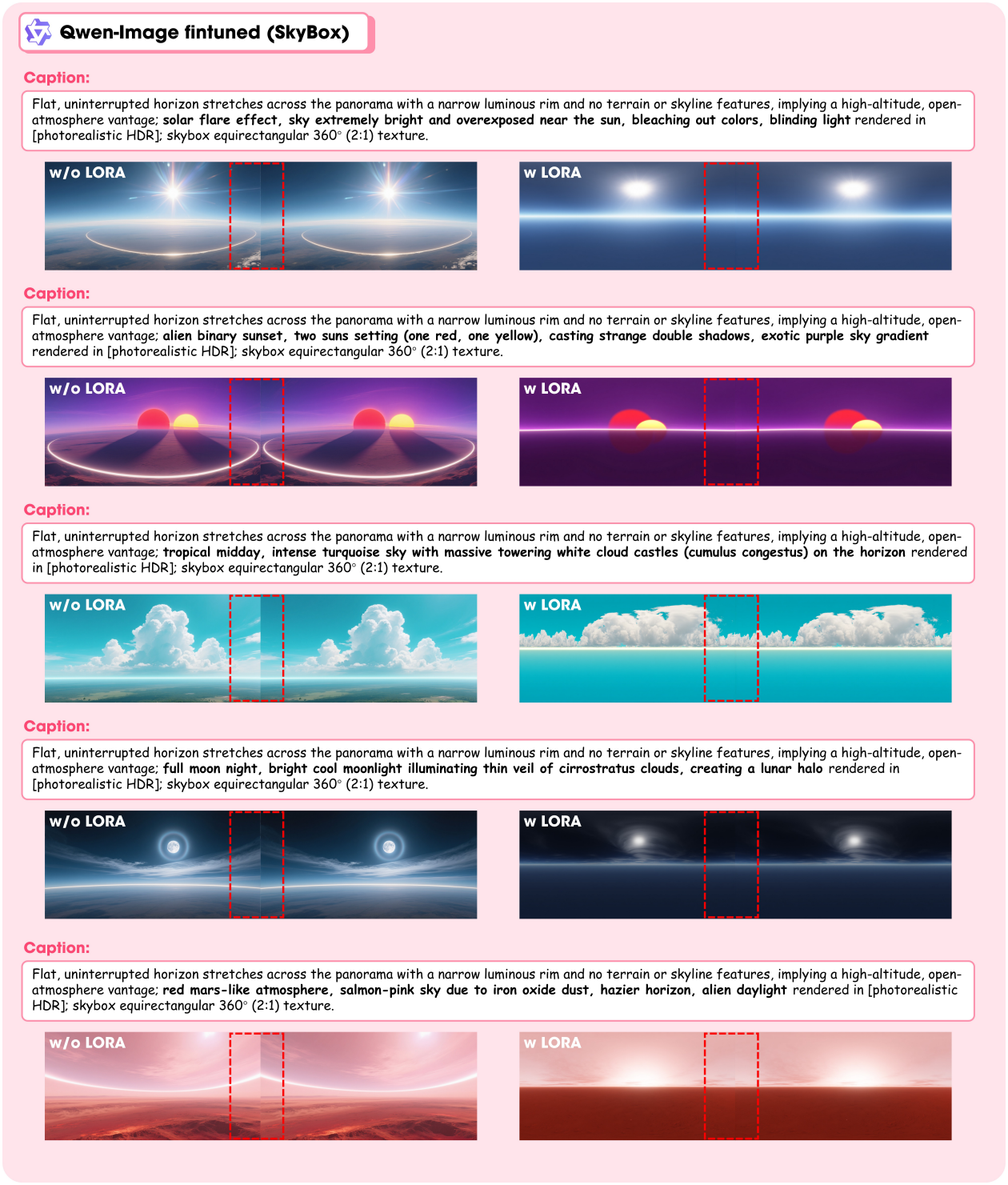


Figure 15. Comparison of the SkyBox generated models from the original Qwen-Image model and the Qwen-Image model fine-tuned using MajutsuDataset-Materials.

GPT-5 AQS prompt

You are an expert 3D rendering and scene composition analyst. Your task is to evaluate the provided set of images, which represent different views of the same 3D scene. You must assess the scene's overall quality based on the four detailed metrics below. Provide a score from 1 (worst) to 10 (best) for each.

METRICS:

1. Structural and View Consistency: Score 1-10. Assesses physical plausibility AND absolute cross-view consistency.

- **Structural Plausibility:** Are geometries complete, closed, and physically sound (e.g., no floating building parts, unsupported cantilevers, or bizarre object intersections)? Is the perspective and scale (e.g., door-to-building ratio) correct?
- **Cross-View Consistency:** Does the scene represent a single, static 3D space?
  - **Geometric Consistency:** Does the structure seen from one view (e.g., a building's silhouette, window placement) perfectly match the same structure seen from another view?
  - **Appearance Consistency:** Does the material, texture, and color of an object (e.g., a specific wall) remain identical across all views where it is visible?

---

2. Scene Richness and Complexity: Score 1-10. Assesses the density, diversity, and thoughtful arrangement of scene content.

- **Asset Variety:** Does the scene contain a wide variety of unique assets (e.g., different building styles, street furniture like lamps)? Or is it clearly reliant on a few repeated models?
- **Content Density & Detail:** Does the scene feel full and lived-in? Are there sufficient "secondary" and "background" elements to fill the space, or does it feel empty and sparse?
- **Layout & Semantics:** Is the scene layout logical and believable (e.g., sidewalks, roads, and green spaces are arranged sensibly)? Does the composition feel deliberate, or like a random collection of assets?

---

3. Material and Texture Fidelity: Score 1-10. Assesses the quality, realism, and physical properties of all surfaces.

- **Resolution & Clarity:** Are textures sharp and high-resolution, even at a medium distance? Or are they blurry, low-resolution, or show compression artifacts?
- **Texture Artifacts:** Are there obvious and distracting repeating patterns (tiling), visible seams, or UV mapping errors (stretching/pinching)?
- **Realism & Scale:** Do materials (e.g., brick, glass, asphalt, foliage) look believable? Is the texture scale (e.g., the size of a brick, the grain of wood) correct relative to the object it's on?
- **Physical Properties (PBR):** Do materials exhibit correct physical light interaction? (e.g., glass is transparent and reflective; metal has sharp specular highlights; concrete is rough and diffuse).

---

4. Lighting and Atmosphere: Score 1-10. Assesses the lighting realism, color harmony, and overall "immersiveness".

- **Lighting & Shadows:** Is there a clear, motivated light source (e.g., the sun)? Are all shadows cast in a consistent, correct direction? Do shadows have realistic softness (penumbras) and contact shadows?
- **Global Illumination & Rendering:** Does the scene feel "flat," or does it show evidence of Global Illumination (GI)? (e.g., are shaded areas realistically lit by bounced light, or are they pitch black? Is there subtle color bleeding?).
- **Color & Mood:** Is the color palette harmonious, aesthetic, and realistic? Does it successfully establish a specific mood or time of day (e.g., bright noon, warm sunset)? Or are colors unnatural, oversaturated, or washed out?
- **Atmosphere & Depth:** Is there a sense of depth (e.g., via atmospheric perspective, fog, or depth of field)? Or do all objects have the same clarity, making the scene feel like a flat cutout?

Answer:

Your response **MUST** be only the scores in this exact format, with no extra text or explanations:

- **Structural_and_View_Consistency:** [score]
- **Scene_Richness_and_Complexity:** [score]
- **Material_and_Texture_Fidelity:** [score]
- **Lighting_and_Atmosphere:** [score]

Figure 16. AQS evaluation prompt template

GPT-5 RDR prompt

You are an expert 3D scene analyst performing a blind A/B test. You will be shown two images, "IMAGE A" and "IMAGE B", which represent two different 3D scenes.
Your task is to compare them **ONLY** on the following metric: **METRIC[i] (i=1,2,3,4)**

METRICS:

1. **Structural and View Consistency:** Assesses physical plausibility AND absolute cross-view consistency.
- **Structural Plausibility:** Are geometries complete, closed, and physically sound (e.g., no floating building parts, unsupported cantilevers, or bizarre object intersections)? Is the perspective and scale (e.g., door-to-building ratio) correct?
- **Cross-View Consistency:** Does the scene represent a single, static 3D space?
  - **Geometric Consistency:** Does the structure seen from one view (e.g., a building's silhouette, window placement) perfectly match the same structure seen from another view?
  - **Appearance Consistency:** Does the material, texture, and color of an object (e.g., a specific wall) remain identical across all views where it is visible?

---

2. **Scene Richness and Complexity:** Assesses the density, diversity, and thoughtful arrangement of scene content.
- **Asset Variety:** Does the scene contain a wide variety of unique assets (e.g., different building styles, street furniture like lamps)? Or is it clearly reliant on a few repeated models?
- **Content Density & Detail:** Does the scene feel full and lived-in? Are there sufficient "secondary" and "background" elements to fill the space, or does it feel empty and sparse?
- **Layout & Semantics:** Is the scene layout logical and believable (e.g., sidewalks, roads, and green spaces are arranged sensibly)? Does the composition feel deliberate, or like a random collection of assets?

---

3. **Material and Texture Fidelity**: Assesses the quality, realism, and physical properties of all surfaces.
- **Resolution & Clarity:** Are textures sharp and high-resolution, even at a medium distance? Or are they blurry, low-resolution, or show compression artifacts?
- **Texture Artifacts:** Are there obvious and distracting repeating patterns (tiling), visible seams, or UV mapping errors (stretching/pinching)?
- **Realism & Scale:** Do materials (e.g., brick, glass, asphalt, foliage) look believable? Is the texture scale (e.g., the size of a brick, the grain of wood) correct relative to the object it's on?
- **Physical Properties (PBR):** Do materials exhibit correct physical light interaction? (e.g., glass is transparent and reflective; metal has sharp specular highlights; concrete is rough and diffuse).

---

4. **Lighting and Atmosphere**: Assesses the lighting realism, color harmony, and overall "immersiveness".
- **Lighting & Shadows:** Is there a clear, motivated light source (e.g., the sun)? Are all shadows cast in a consistent, correct direction? Do shadows have realistic softness (penumbras) and contact shadows?
- **Global Illumination & Rendering:** Does the scene feel "flat," or does it show evidence of Global Illumination (GI)? (e.g., are shaded areas realistically lit by bounced light, or are they pitch black? Is there subtle color bleeding?).
- **Color & Mood:** Is the color palette harmonious, aesthetic, and realistic? Does it successfully establish a specific mood or time of day (e.g., bright noon, warm sunset)? Or are colors unnatural, oversaturated, or washed out?
- **Atmosphere & Depth:** Is there a sense of depth (e.g., via atmospheric perspective, fog, or depth of field)? Or do all objects have the same clarity, making the scene feel like a flat cutout?

Answer:

Which image is **BETTER** according to this single metric? Your response **MUST** be only a single letter: "A" or "B", and nothing else. If they are of identical quality, make your best guess. Do not declare a tie.
- **"A":** IMAGE A is better.
- **"B":** IMAGE B is better.

Figure 17. RDR evaluation prompt template

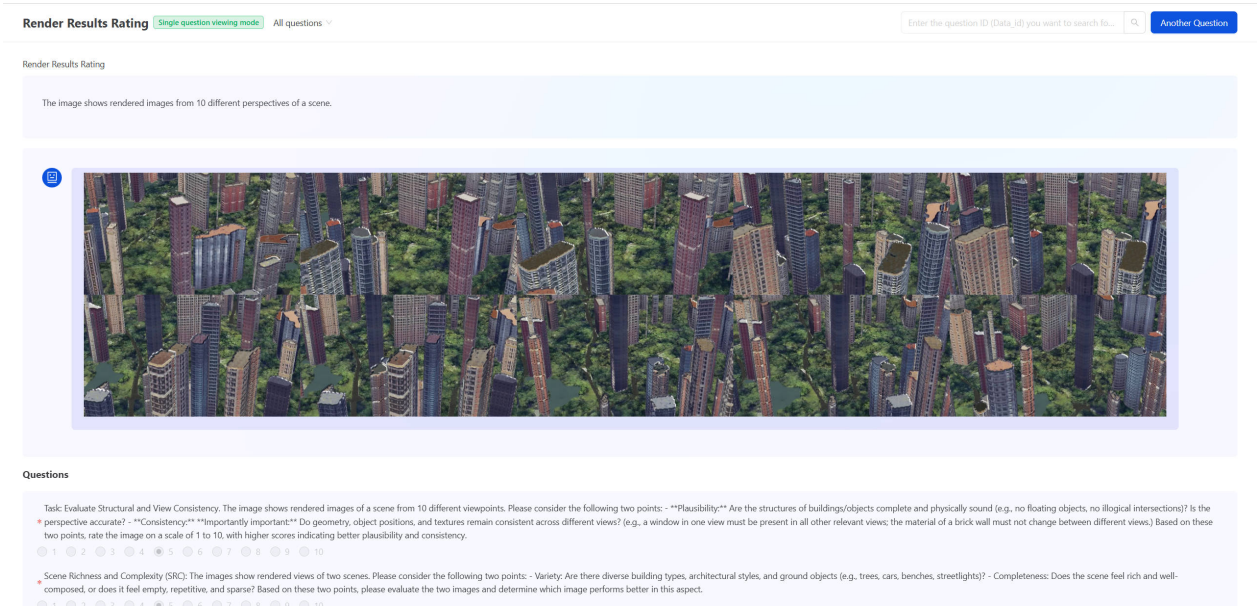


Figure 18. User AQS evaluation platform interface

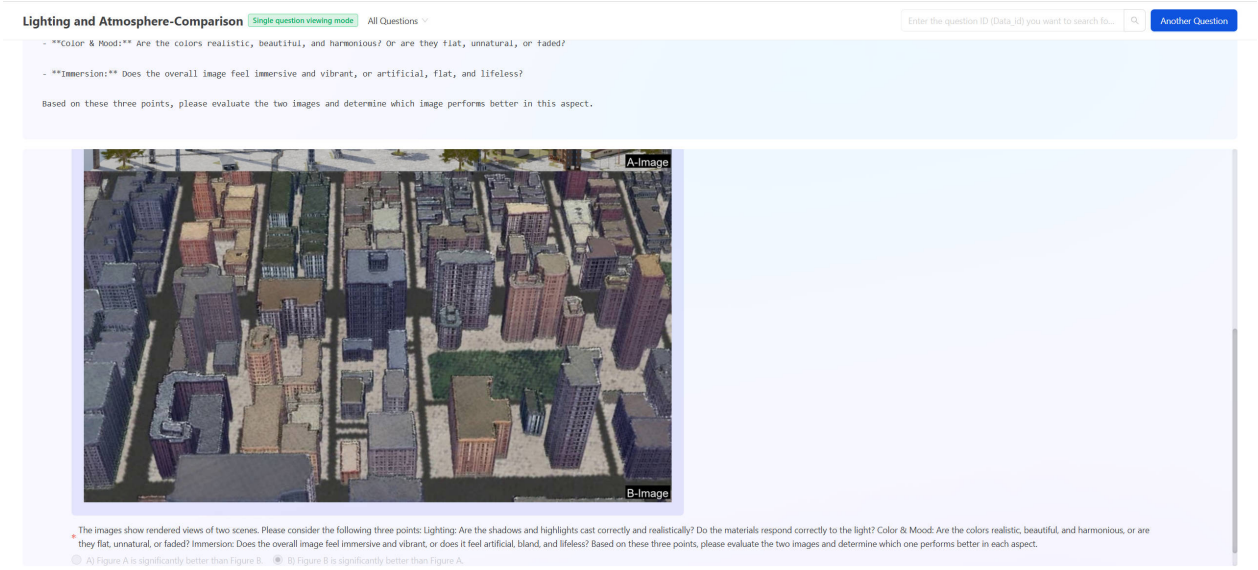


Figure 19. User RDR evaluation platform interface