Incremental VistaDream: Incremental Expansion of Field of View for Single-view Scene Reconstruction

Mina Matoba
Department of Integrated Information
Aoyama Gakuin University
Sagamihara, Japan
c5624217@aoyama.jp

Tobe Yoshito
Department of Integrated Information
Aoyama Gakuin University
Sagamihara, Japan
y.tobe@rcl-aoyama.jp

Kazuhiko Sumi
Department of Integrated Information
Aoyama Gakuin University
Sagamihara, Japan
sumi@it.aoyama.ac.jp

***Abstract*— We propose Incremental VistaDream, a single-view scene reconstruction method that represents a 3D scene as a set of Gaussian distributions. Starting from a 60° field-of-view (FoV) image, we progressively expand the horizontal FoV to 120° by applying four 30% expansions on each side (eight steps in total). After each step, the expanded image is converted into 3D Gaussian distributions via monocular depth estimation, and the resulting 3D scenes composed of these Gaussian distributions are then sequentially registered and merged. For evaluation, we compare only the newly synthesized regions against ground-truth panoramas, focusing on semantic consistency. Across 11 scenes, Incremental VistaDream improves SSIM by 31% and reduces LPIPS by 18% over VistaDream.**

Keywords*— *Computer Vision, Computer Graphics, Single-View Scene Reconstruction, Novel View Synthesis, Gaussian Splatting

I. Introduction

Recent progress in single-view scene reconstruction and novel view synthesis (NVS) [1] aims to generate high-fidelity 3D scenes from sparse observations or even a single image. Traditionally, scenes were modeled as textured geometry, whereas advances in NVS reconstruct the light field from multi-view images to synthesize novel viewpoints without manual modeling. These developments enable high-quality image generation and broaden downstream applications.

In particular, Neural Radiance Fields (NeRF) [2] estimate radiance and density along rays from a large number of input images, achieving high-quality free-viewpoint rendering. Nevertheless, NeRF requires long training times for complete 3D scenes and assumes multi-view images, which limits its applicability. In contrast, 3D Gaussian Splatting (3DGS) [3] represents a scene as a set of anisotropic 3D Gaussians parameterized by position, scale, and color, and directly splats these 3D Gaussian distributions onto the image plane. This enables much faster rendering compared to NeRF and yields high-quality results with shorter training. However, 3DGS still depends on multi-view input and is not applicable when only a single image is available.

To address this limitation, VistaDream [4] was proposed as a method for generating a 3D Gaussian Field (3D GF) from a single image. A 3D GF denotes a scene modeled as a set of anisotropic 3D Gaussians that can be directly rendered by splatting. VistaDream combines diffusion-based outpainting and depth estimation to complete the outer regions of the input image and first builds a 3D Global Scaffold (Scaffold), where a Scaffold represents a 3D scene as a set of 3D Gaussian distributions. Specifically, the method expands 45% of the surrounding area of the input image in one step (outpainting) [5], and then constructs a coarse Scaffold from RGB-D images containing both color and depth. To ensure geometric consistency, Multiview Consistency Sampling (MCS) [4] is applied across images rendered from different viewpoints. Finally, the results are refined and integrated into a coherent 3D GF. However, this one-shot expansion is limited to a 90° field of view (FoV), which is insufficient to cover the approximately 120° horizontal FoV of VR head-mounted displays (HMDs). Furthermore, large one-shot expansions are prone to semantic drift, a phenomenon in generative models where scene semantics deviate from the intended structure (e.g., buildings turning into trees).

In this work, we propose Incremental VistaDream, a method designed to achieve a 120° FoV while mitigating semantic drift from a single image. Our approach progressively expands the input image by 30% in both directions, guided by user-defined textual prompts that constrain the semantics of the generated regions. The resulting expanded images are converted into Scaffolds, which are subsequently aligned and merged using GaussReg [7]. GaussReg is a registration technique for aligning multiple Scaffolds, suppressing geometric misalignment and enabling continuous scene representations. In summary, Incremental VistaDream introduces (i) progressive outpainting for 120° FoV expansion and (ii) integration of multiple Scaffolds via GaussReg. We evaluate Incremental VistaDream against the one-shot expansion baseline.

The remainder of this paper is organized as follows: Section II reviews related work, Section III presents our approach, Section IV reports experimental evaluations, and Section V concludes the paper.

II. RELATED WORK

In this section, we review existing studies on diffusion models, large-scale Vision-Language Models (VLMs), and Scaffold integration techniques, in order to clarify the positioning of our approach.

A. Vision-Language Models (VLMs)

Vision-Language Models (VLMs) integrate visual and natural language processing, enabling tasks such as captioning, reasoning, and guided generation. They can generate detailed textual descriptions conditioned on input images, and such descriptions can assist the generative process, improving the quality of RGB-D inpainting and image expansion. In our

979-8-3315-7896-1/26/$31.00 ©2026 IEEE

approach, VLM-guided textual prompts are used to stabilize semantic consistency during progressive outpainting; in the experiments, we instantiate the VLM with LLaVA [6].

B. Diffusion Models

Diffusion models are generative approaches that learn to add and remove noise in a progressive manner, enabling high-quality image generation and inpainting. Recent work has combined text-conditioned generation and depth estimation, extending their applicability to NVS from a single image. While most existing studies emphasize consistency between input and generated images, they often fail to maintain cross-view coherence, leading to artifacts such as structural inconsistency or semantic drift across viewpoints. VistaDream addresses this issue by incorporating Multiview Consistency Sampling (MCS), which enforces cross-view consistency during the reverse diffusion process and stabilizes multi-angle generation.

C. Integration Methods and the Role of GaussReg

When integrating multiple Scaffolds, it is essential to align their positions and shapes, making registration techniques a critical component. A classical approach is Iterative Closest Point (ICP), which iteratively extracts correspondences and estimates rigid transformations through rotation and translation. However, ICP suffers from sensitivity to initialization and susceptibility to local minima. In contrast, GaussReg introduces a probabilistic formulation by treating Scaffolds as distributions and directly aligning them in distribution space. This distribution-level registration avoids explicit point correspondences, providing robustness against initialization issues inherent in ICP and enabling stable integration of successively generated Scaffolds. Consequently, GaussReg is considered effective for suppressing semantic drift and achieving coherent wide-FoV reconstruction.

III. PROPOSED METHOD

In this section, we describe our approach for expanding a single input image to a 120° FoV and generating a high-fidelity 3D GF. In Section III-A, we explain the progressive outpainting strategy that expands the input image while maintaining semantic consistency. Section III-B then describes how the resulting Scaffolds are sequentially aligned and integrated. Finally, Section III-C presents our integration strategy with GaussReg, which suppresses redundancy and ensures coherent wide-FoV reconstruction.

A. Outpainting

Fig. 1 shows the overview of Incremental VistaDream. Given an input image, we first employ LLaVA to generate textual descriptions. The descriptions are reused as prompts for progressive outpainting, preserving scene semantics and reducing artifacts. To preserve the semantic context of the input, each expansion step is limited to 30%.

Let the FoV of the input image be denoted as FoV_{in}, and the target FoV as FoV_{out}. The input image is represented as I_0, with height H and width W. We define the outpainting operation with a 30% extension as $OP_{0.3}(\circ)$, and cropping operators that preserve the width W while extracting an $H \times W$ region from the left or right of the expanded image as C_L and C_R, respectively.

Let I_{2k-1} and I_{2k} denote the k-th left and right expansions, respectively ($k = 1, 2, 3, \dots, N$).

```
Initialization: $J_0 \leftarrow OP_{0.3}(I_0)$
                $I_1 \leftarrow C_L(J_0)$
                $I_2 \leftarrow C_R(J_0)$
$for\ k$: $1 \sim N$:
    $J_{2k-1} \leftarrow OP_{0.3}(I_{2k-1})$
    $J_{2k} \leftarrow OP_{0.3}(I_{2k})$
    $I_{2k+1} \leftarrow C_L(J_{2k-1})$
    $I_{2k+2} \leftarrow C_R(J_{2k})$
```

To achieve FoV_{out} = 120°, we set $N = 4$, yielding I_k $(k = 0, \dots, 8)$. Each image I_k is processed by 3DGS, producing a coarse Scaffold denoted as $S_k = GS\ (I_k)$. Since each S_k contains positional errors arising from monocular depth estimation, we sequentially align them using GaussReg for registration (see Section II-C); ICP is used only as a baseline for comparison.

A higher expansion ratio may cause semantic inconsistencies such as distorted object shapes or discontinuities at region boundaries, as well as geometric inconsistencies such as deformation of straight structures. Conversely, setting the ratio too small would require a large number of expansion steps to reach the target FoV, resulting in increased computation time and resource consumption due to repeated novel view prediction and image generation. Therefore, a 30% expansion was selected as a balance, ensuring semantic consistency while maintaining computational efficiency. Furthermore, in our design, four progressive expansions are applied in each horizontal direction, resulting in a total of eight steps that achieve an FoV of approximately 120°. This design balances semantic stability and computational cost; a full ablation of expansion ratios is left as future work. In Fig. 1, the progressive outpainting is explicitly indicated as multiple iterations (×4 per side), yielding a total of 9 expanded images and thus 9 Scaffolds before registration.

To ensure fairness, we adopt VistaDream's Multiview Consistency Sampling (MCS) as-is and apply identical settings to both methods based on the public implementation.

B. Sequential Integration of Scaffolds

The sequence of 9 Scaffolds S_k $(k = 0, \dots, 8)$ obtained from the expanded images is integrated in the order of generation. Since each Scaffold is produced independently from monocular depth estimation, small errors in position or shape are unavoidable. If these errors are not corrected, they accumulate and lead to noticeable distortions in the final wide-FoV reconstruction. For this reason, careful step-by-step registration and merging are required.

First, we estimate a coarse registration using GeoTransformer [8]. This method is chosen because it can

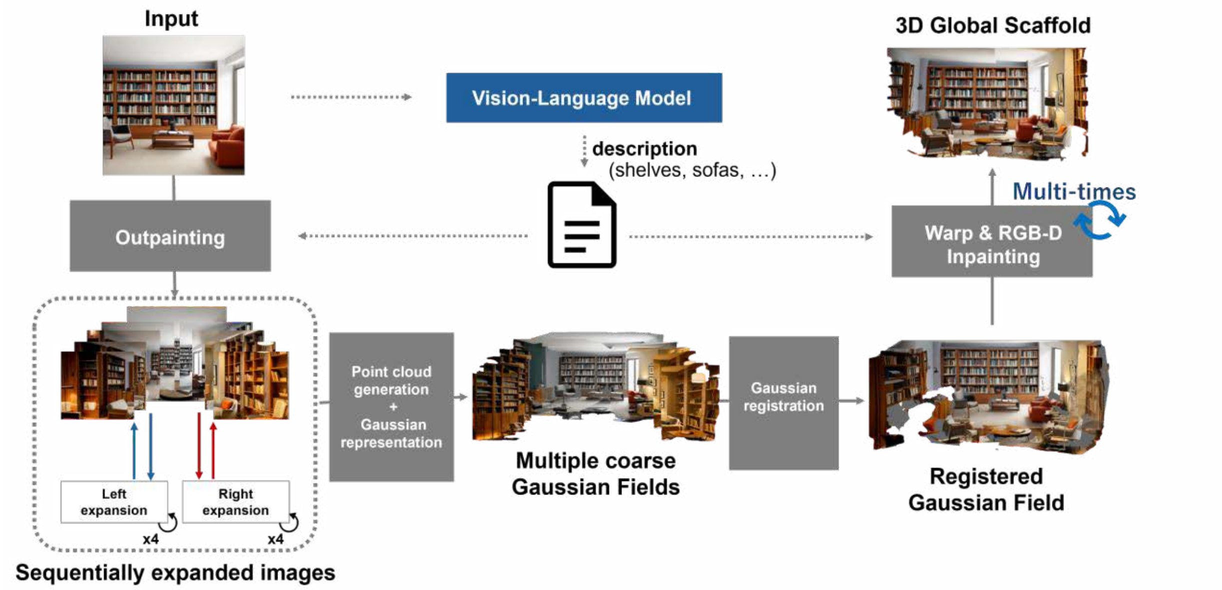

Fig. 1. System overview of Incremental VistaDream. The pipeline progressively expands a single image to a 120° FoV via VLM-guided outpainting with textual prompts, generates a 3D GF from each expansion, and aligns them with GaussReg to obtain a registered wide-FoV 3D GF. For consistency selection, we apply Multiview Consistency Sampling (MCS) with the same settings as VistaDream, updating the 3D GF only with renderings that satisfy cross-view consistency.

capture both the overall structure of the scene and the local details of the Gaussian distributions. By considering information from a wide area at once, GeoTransformer provides a more reliable initial guess of the relative pose between Scaffolds. Without this step, the following refinements would be more likely to fail or converge to incorrect registrations.

Next, the registration is refined with GaussReg [7], which directly matches Gaussian distributions. Point-to-plane ICP [9] is included only as an ablation baseline; it is not part of our main pipeline. Instead of only matching points directly, in point-to-plane ICP, each point is adjusted to lie close to the counterpart surface of the other Scaffold. This is especially effective in scenes with many flat or structured areas, such as roads, walls, or building facades. Using point-to-plane ICP reduces registration errors more efficiently and achieves higher accuracy. We iterate the registration until convergence, defined as translation < 1 mm and rotation < 0.05°.

After the registration step, the overlapping parts of the Scaffolds are merged. We merge overlaps via weighted averaging and radius-based fusion. Weighted averaging means that Gaussians with more reliable depth or appearance are given stronger influence in the merged result. This reduces the effect of uncertain or noisy components. Radius-based fusion removes redundant components by combining those that are very close to each other. This not only reduces memory but also avoids excessively dense clusters of points in overlapping regions, resulting in a smoother and cleaner Scaffold.

After these steps, we perform pose-graph optimization [10] to minimize accumulated drift. Pose-graph optimization is a global refinement method that adjusts all camera poses jointly based on pairwise registration constraints. In this approach, each viewpoint is represented as a node, and each estimated transformation is represented as a connection between nodes. By adjusting all nodes together, small errors from individual steps are spread out and corrected globally. As a result, the final integrated 3D GF is more consistent and stable.

Through these stages—initial registration with GeoTransformer, local refinement with point-to-plane ICP, careful merging of overlapping components, and global adjustment with pose-graph optimization—we obtain an integrated 3D GF that maintains both semantic and geometric consistency. This sequential integration is essential for expanding from a single narrow-FoV image to a coherent wide-FoV 3D reconstruction.

C. Integration Strategy with GaussReg

In this study, we employed GaussReg [7] to integrate the sequentially generated Scaffolds. The use of GaussReg is important because each Scaffold produced from the progressive expansion contains slight positional shifts and inconsistencies. If these are not corrected, the final integrated scene would suffer from visible misalignments or unnatural overlaps. GaussReg treats each Scaffold as a distribution and performs registration only through rotation and translation. This approach avoids the need for establishing explicit point-to-point correspondences, which can be unreliable when the views differ greatly, and thus simplifies the integration process. By aligning distributions directly, GaussReg achieves stable global registration of successive Scaffolds.

However, since GaussReg only considers global rotation and translation, it cannot correct local deformations within the Scaffolds. As a result, small distortions may remain after integration. To reduce this problem, we introduce an additional strategy that detects overlapping regions between successive Scaffolds and removes redundant elements. The detection is based on positional closeness and appearance similarity. Specifically, when the centers of two Gaussian components are closer than a predefined distance and their appearance, such as mean color, is sufficiently similar, they are considered duplicates. These duplicates are removed from the later-generated Scaffold during the integration process.

This policy is necessary because successive expansions inevitably generate overlapping areas, and if such redundancies are not removed, the overlaps can cause visual inconsistencies at the boundaries, such as double edges or unnatural density. By eliminating duplicates, the integrated 3D GF becomes cleaner, with fewer redundant elements, and the transition between regions becomes smoother. In this way, GaussReg provides the global registration, while the overlap-removal strategy ensures local consistency. Together, they enable rigid registration and produce a more coherent and visually plausible integrated 3D GF.

IV. EXPERIMENTS

In this section, we present the evaluation methodology and results. All experiments were conducted on a workstation equipped with an NVIDIA RTX 6000 Ada Generation GPU (49 GB memory), 128 GB RAM, and Ubuntu 22.04 with CUDA 12.4. Processing one scene required 15–20 min. Fig. 2 compares ground-truth panoramas, VistaDream, and Incremental VistaDream across 11 scenes (a–k). We conducted qualitative and quantitative evaluations: the former examines visual consistency, and the latter compares Incremental VistaDream with VistaDream in the image space using SSIM [11] and LPIPS [12] on the expanded regions only.

A. *Qualitative Evaluation*

We qualitatively compare panoramic images including the expanded regions. Fig. 2 shows the comparison among ground truth (GT) images, the baseline (VistaDream), and our method (Incremental VistaDream). Scenes (a–h) are outdoor panoramas, while (i)–(k) are indoor environments. The following issues were observed with VistaDream: In images (b), (f), (g), (i), and (j), semantic drift was observed, where additional spaces or non-existent objects (blue boxes) were generated. The expanded regions (red boxes) appeared darker than the central regions, causing illumination discontinuities. Significant geometric inconsistencies were also observed; for example, in Scene (e), the straight structure of the road (red boxes) was distorted. In Scene (h), VistaDream not only caused unnatural darkening at the image boundaries but also generated structures resembling parts of the input image in a manner inconsistent with the background. In contrast, Incremental VistaDream successfully avoided such semantic drift and preserved both illumination consistency and semantic plausibility across the expanded regions. These results reveal that semantic drift and illumination inconsistencies occur simultaneously. In contrast, Incremental VistaDream exhibited the following properties: No new rooms or spurious structures were generated after expansion, and the integrity of the original scene was preserved. Brightness and color tones remained highly consistent with the central view, with almost no darkening or unnatural tonal variations. Geometric continuity, such as the straightness of roads and buildings and the arrangement of furniture, was maintained. These results demonstrate that progressive outpainting with sequential registration suppresses semantic drift and maintains illumination and geometric consistency.

B. *Quantitative Evaluation*

For quantitative evaluation, we focus exclusively on the expanded regions and employ SSIM and LPIPS. SSIM assesses structural similarity between a prediction and a ground truth photograph, thus being sensitive to blur, contrast, and local rendering fidelity. However, SSIM alone is not sufficient to judge semantic drift. We therefore add LPIPS, which measures perceptual/semantic similarity using deep features and is known to correlate with human judgments. In this study, we emphasize relative improvements under identical conditions rather than absolute scores. Table I reports per-scene SSIM and LPIPS. On average across 11 scenes, VistaDream achieved SSIM = 0.318 and LPIPS = 0.612, whereas Incremental VistaDream achieved SSIM = 0.416 and LPIPS = 0.502. This corresponds to a +0.098 (+31%) improvement in SSIM and a −0.110 (-18%) reduction in LPIPS. To verify statistical significance, we conducted a one-sided paired t-test across 11 scene pairs. The improvement in SSIM yielded $\mathrm{t}(10) = 6.37$, $\mathrm{p} = 4.1 \times 10^{-5}$, and the reduction in LPIPS yielded $\mathrm{t}(10) = -5.81, \mathrm{p} = 8.5 \times 10^{-5}$. Both results are highly significant ($\mathrm{p} < 0.001$). Notably, Scene (j) achieved the highest SSIM value of 0.5837, while the same scene also achieved the lowest LPIPS value of 0.4160. Because the ground truth is a real photograph while the expansion is an imaginative completion, absolute pixel scores tend to be low across all methods. Nevertheless, consistent with the qualitative results (Fig. 2), Incremental VistaDream significantly reduces semantic drift near FoV boundaries.

C. *Failure Cases and Limitations*

Most failures arise during the progressive expansion (outpainting and monocular-depth-based 3D GF construction), especially in regions with a lack of structural visual features (e.g., uniform walls, sky, plain floors) and weak texture cues. In such low-evidence areas, semantic completion tends to overreach, yielding object duplication, shape distortion, and boundary discontinuities. Consistency selection (MCS) suppresses cross-view inconsistencies by filtering rendered views, but it does not guarantee semantic correctness; thus, hallucinations introduced during the expansion may persist. These failures stem not from the expansion strategy itself, but from the behavior of the generative model when visual information is limited or imbalanced. These failure cases suggest that insufficient constraints during the outpainting stage are the primary cause. In our pipeline, a text prompt is generated from the input image using a Vision-Language Model (VLM), and a text-conditioned diffusion model performs progressive outpainting. However, this setup can cause the generation to be overly influenced by dominant visual elements in the input, sometimes resulting in repeated strong colors or hallucinated structures when visual cues are limited, as shown in Fig. 3.

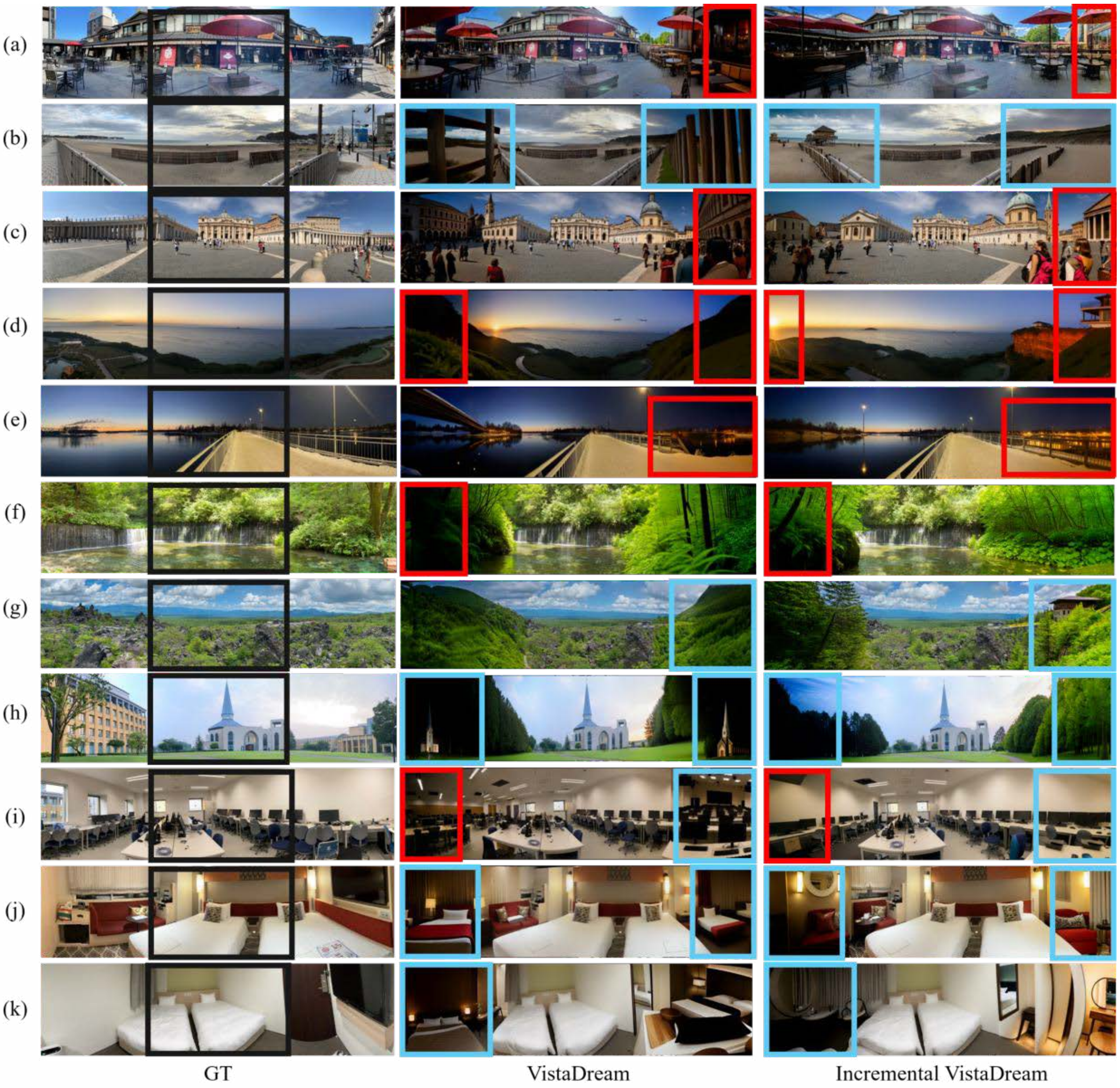

Fig. 2. Comparison across 11 scenes (Left: ground truth (GT), Center: VistaDream, Right: Incremental VistaDream). Red boxes indicate boundary artifacts, and blue boxes indicate semantic drift. VistaDream shows boundary artifacts (red) and semantic drift (blue), while Incremental VistaDream achieves a 120° FoV with minimal artifacts.

To address this issue, we consider it important to extract visual features from the input image and use them to constrain the generation during outpainting. Such feature-based control has the potential to suppress these failures and produce more stable expansions, even in regions with few visual cues.

First, when a scene contains areas with very strong colors, both methods tended to repeat and exaggerate those colors, sometimes forming unnatural duplicated patterns (Fig. 3(a)). As a result, neither the baseline nor Incremental VistaDream produced convincing outputs, and no clear advantage was observed.

Second, in scenes with few visual cues, such as wide water surfaces or snowy fields, both methods occasionally created objects that were not present in the real scene, resulting in unrealistic structures that blocked the intended FoV (Fig. 3(b)). These errors are likely caused by the model attempting to "fill in the blanks" by guessing the missing content, rather than grounding the generation in reliable evidence.

In future work, we will quantify uncertainty in low-cue regions and constrain generation with feature-level priors to systematically reduce these failures.

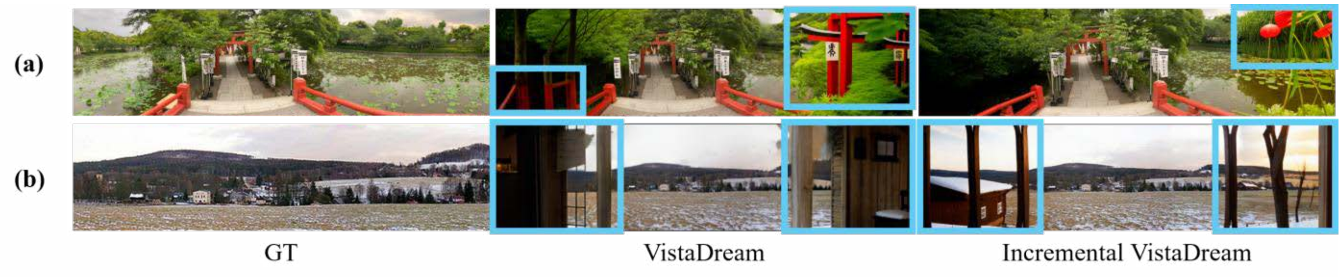

Fig. 3. Representative failure cases in wide-FoV expansion:
(a) **Strong-color repetition:** visually salient regions (e.g., bright red structures) are repeatedly synthesized, producing duplicated or exaggerated textures.
(b) **Hallucination in low-cue regions:** when structural visual features are lacking (e.g., snow or water surfaces), the model guesses missing content and generates non-existent structures.

TABLE I. PER-SCENE METRICS ON EXPANDED REGIONS (HIGHER IS BETTER / LOWER IS BETTER)

	SSIM↑		LPIPS↓	
	VistaDream	*Incremental VistaDream*	*VistaDream*	*Incremental VistaDream*
(a)	0. 2045	0. 2289	0. 6728	0. 6335
(b)	0. 1482	0. 2756	0. 6597	0. 5175
(c)	0. 2389	0. 2744	0. 5658	0. 5213
(d)	0. 4536	0. 5806	0. 6047	0. 5232
(e)	0. 3788	0. 4985	0. 5370	0. 4265
(f)	0. 1871	0. 2599	0. 7584	0. 5526
(g)	0. 2176	0. 4285	0. 6553	0. 4239
(h)	0. 4619	0. 5569	0. 5645	0. 4633
(i)	0. 2502	0. 3474	0. 5743	0. 5301
(j)	0. 5174	0. 5837	0. 5192	0. 4160
(k)	0. 4428	0. 5416	0. 6185	0. 5104

V. CONCLUSION

This work demonstrates that the proposed combination of progressive outpainting and sequential registration of Scaffolds enables the generation of panoramic images with a 120° FoV while effectively suppressing semantic drift. By gradually expanding the input image and then carefully integrating the resulting 3D representations, the method maintains both semantic consistency and geometric continuity, which are essential for high-quality 3D scene reconstruction. Compared to VistaDream, Incremental VistaDream achieved statistically significant improvements of +31% in SSIM and -18% in LPIPS, demonstrating clear advantages in both structural preservation and perceptual quality. These improvements confirm that stepwise expansion and robust integration directly contribute to more stable reconstructions, providing practical benefits for wide-FoV panoramic content suitable for high-quality 3D scene reconstruction.

As a future direction, we plan to incorporate feature-based constraints into the outpainting stage to further suppress failure cases and improve robustness in visually ambiguous regions.

REFERENCES

[1] R. Tucker and N. Snavely: Single-View View Synthesis with Multiplane Images, Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2020), pp. 551–560 (2020).

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Proc. Eur. Conf. Comput. Vis. (ECCV 2020), Lecture Notes in Computer Science, Vol. 12346, pp. 405–421 (2020).

[3] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis: 3D Gaussian Splatting for Real-Time Radiance Field Rendering, ACM Trans. Graph. 42 (4), Article 139, 1–14 (2023)

[4] H. Wang, Y. Liu, Z. Liu, W. Wang, Z. Dong, and B. Yang: VistaDream: Sampling Multiview-Consistent Images for Single-View Scene Reconstruction, arXiv:2410.16892 (2024)

[5] H. Wang et al. : VistaDream: Sampling Multiview-Consistent Images for Single-View Scene Reconstruction, GitHub repository, WHU-USI3DV/VistaDream, commit 9a743a9 (Oct. 23, 2024). Available: https://github.com/WHU-USI3DV/VistaDream (accessed Jan. 15, 2025)

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee: LLaVA: Large Language and Vision Assistant via Visual Instruction Tuning, NeurIPS (NeurIPS 2023, Oral), arXiv:2304.08485 (2023)

[7] J. Chang, Y. Xu, Y. Li, Y. Chen, and X. Han: GaussReg: Fast 3D Registration with Gaussian Splatting, Proc. Eur. Conf. Comput. Vis. (ECCV 2024), 407–423 (2024)

[8] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu: GeoTransformer: Unifying Alignment and Correspondence for Point Cloud Registration, IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 9806–9821 (2023)

[9] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. Image and Vision Computing, 10(3):145–155, 1992.

[10] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard: g2o: A General Framework for Graph Optimization, Proc. IEEE Int. Conf. on Robotics and Automation (ICRA 2011), pp. 3607-3613 (2011)

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli: Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Trans. Image Process., Vol. 13, No. 4, pp. 600-612 (2004)

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 586-595 (2018)