FlexWorld：Progressively Expanding 3D Scenes for Flexible-View Exploration

Luxi Chen

Project page: https://ml-gsai.github.io/FlexWorld
Github: https://github.com/ML-GSAI/FlexWorld

Motivation

- We hope to generate 3D scenes that support more flexible viewpoints.
- Single-pass generation is highly challenging; broader scenes can be generated through multiple iterations.
- Since iteration is required, the model must be able to acquire current scene information to determine the content of the next generation.

↓

- A video-to-video model that supports large camera variation!

Design of video model

- Fine-tuning the video model, under a given camera trajectory, it takes the incomplete video rendered from the current coarse scene and outputs a repaired high-quality video (i.e., video-to-video).

Input

FlexWorld

Design of video model

- Under large camera variation, existing models fail to effectively complete the scene.

ViewCrafter

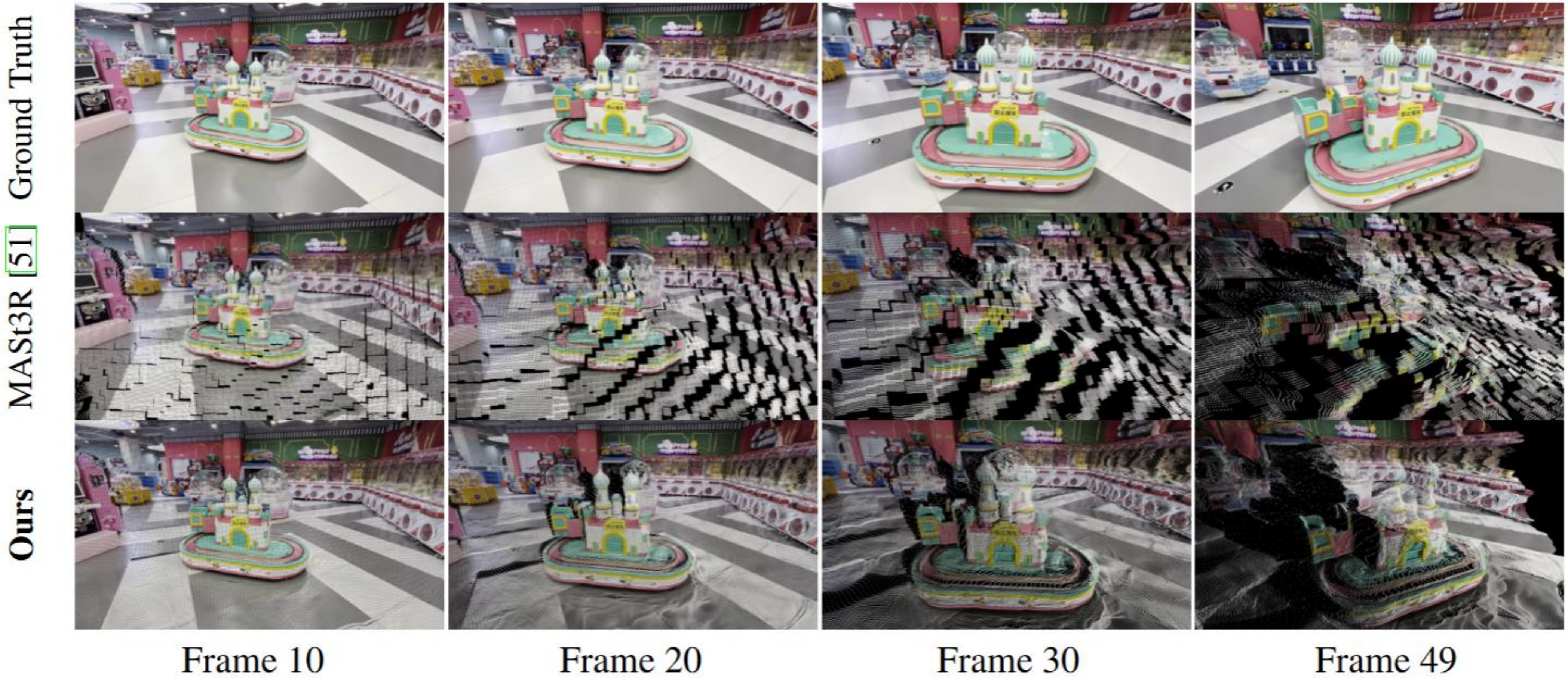
See3D

Design of video-to-video

- Utilize a more powerful base model，CogVideoX-5B-I2V
  - Replace the original model's image condition directly with a video condition.
  - Specifically, the video condition is compressed via a VAE and then concatenated with the denoising latent variables.

Design of video-to-video

- An improved strategy for creating training video pairs.



3D scene construction -- scene initialization

- Video-to-video translation requires a 3D scene as an intermediary.
- Starting from a single image input, we utilize a dense stereo model (DUSt3R) to obtain the point cloud corresponding to this image. This point cloud is then converted into 3DGS to serve as our scene initialization.

3D scene construction -- novel view synthesis

- How can a scene with only a frontal view be transformed into a 360-degree scene?
- A key limitation of V2V is that if the rendered 3D scene lacks substantial 3D content, the completed content may become inconsistent with the input camera trajectory.
- The proposed solution is to first move the camera backward to expand the scene, then sequentially rotate it 180 degrees to the left and 180 degrees to the right.

3D scene construction -- novel view synthesis

3D scene construction -- scene integration

- With the novel view, we still need to convert them into 3D content.
- We select m keyframes from the generated videos and use DUSt3R to simultaneously estimate the depth D_i for them along with the reference image. The depth $\widehat{D}_0$ of the reference image is known and is used to estimate the scale factor.
- The new 3D content is incorporated according to the masks to avoid duplication.

$$\tilde{D}_i = \text{Depth-align}\left(\frac{\text{Median}(D_0))}{\text{Median}(\hat{D}_0))} \cdot \hat{D}_i, D_i, M_i\right),$$

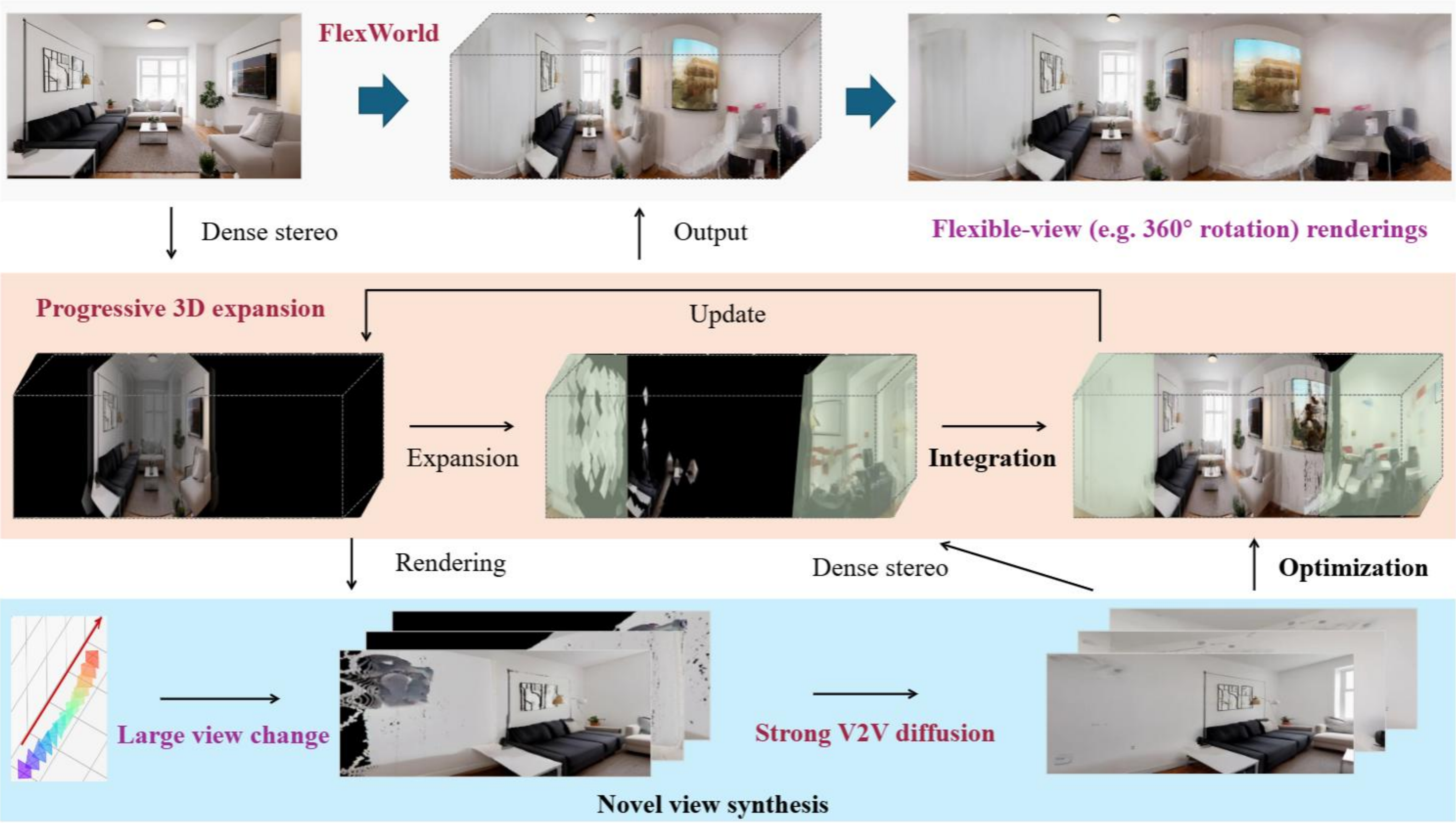
$$\mathcal{P}_i = \{\tilde{D}_i(u,v)E_i^{-1}K^{-1} \cdot (u,v,1)^T | M_i(u,v) = 1\},$$

3D scene construction -- scene optimization

- Upon obtaining the point cloud P_i, we convert it into 3DGS and merge it into the original scene. Then, all video frames are treated as ground truth to perform a comprehensive 3DGS optimization of the entire scene.

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\mathrm{SSIM}} \mathcal{L}_{\mathrm{SSIM}} + \lambda_{\mathrm{LPIPS}} \mathcal{L}_{\mathrm{LPIPS}}.$$

Overview of FlexWorld



Experiments

FlexWorld

Part I Main results

Thank you!