Supplementary Material for
ScenePainter: Semantically Consistent Perpetual 3D Scene Generation with Concept Relation Alignment

The supplementary material is structured as follows:

- Section A illustrates the concrete algorithms of the concept relation construction and refinement stages, and the automatic pipeline to generate novel images with similar initial scene representation for visual diversity in the construction stage.
- Section B introduces the details on human preference quantitative evaluation.
- Section C presents additional qualitative results to facilitate a thorough comprehension of our method.

A. Algorithms

To clearly illustrate the computational process of our framework, we summarize the concept relation construction stage in Algorithm 1 and the concept relation refinement stage in Algorithm 2 respectively.

In addition, we propose an automatic pipeline to generate similar scene images for class-specific prior preservation in the concept relation construction stage as shown in Figure 1. To be specific, for a relation-concept handle pair $< v_i, r_{(i,j)}, v_j >$ with masks m_i and m_j, we keep the commonly masked region unchanged and generate scene description by Vision Language Model (VLM) as the text prompt for the native outpainting model to fill in the remaining partial image. Then we obtain the novel scene image and feed it to the VLM with the iterative query "$Is\ \ there\ \ any\ \ X_i\ \ in\ \ this\ \ image\ ?$" where $X_i \in \{X_1, ..., X_T\}$ is an wanted effect specified by text, such as "distortion", "outpainting border", or "blurry effect". If any unwanted effect is detected, we repeat the outpainting process and visual validation. In this way, novel scene views with various spatial layouts and object appearances can significantly enrich the diversity in the final generated 3D view sequences.

Algorithm 1 Concept Relation Construction

Input: Initial scene image I_0
Output: Constructed SceneConceptGraph G_0 and customized text-to-image model $\bar{\mathcal{M}}_0$

$\mathcal{V}, m \leftarrow$ Define_by_users(I_0) ▷ Initialize concept set with masks
$\mathcal{R} \leftarrow$ Connect_relation$(\mathcal{V})$ ▷ Initialize relation set
$G \leftarrow < \mathcal{V}, \mathcal{R} >$ ▷ Initialize SceneConceptGraph
$\mathcal{M} \leftarrow$ Load_pre-trained_model() ▷ Initialize text-to-imgae generation model
$p\_thrsh \leftarrow 0.5$ ▷ Probability threshold of using generated similar scene image
for $< v_i, r_{(i,j)}, v_j > \in G$ **do**
  $prompt \leftarrow < v_i, r_{(i,j)}, v_j >$ ▷ Take concept-relation pair as text prompt
  $m_u \leftarrow m_i \cup m_j$ ▷ Union mask of two concepts
  Sample $p \sim$ Uniform$(0, 1)$
  if $p > p\_thrsh$ **then**
    $I_{gen} \leftarrow \mathcal{M}(I_0, prompt)$
    $\mathcal{L} \leftarrow$ Calculate_loss(I_0, I_{gen}, m_u)
  else
    $I_s \leftarrow$ Generate_novel_scene_image$(I_0, prompt)$
    $I_{gen} \leftarrow \mathcal{M}(I_s, prompt)$
    $\mathcal{L} \leftarrow$ Calculate_loss(I_s, I_{gen}, m_u)
  end if
  $G \leftarrow$ Optimize_prompt$(\mathcal{L}, G)$
  $\mathcal{M} \leftarrow$ Optimize_model$(\mathcal{L}, \mathcal{M})$
end for
$G_0 \leftarrow G$ ▷ Final optimized SceneConceptGraph
$\bar{\mathcal{M}}_0 \leftarrow \mathcal{M}$ ▷ Final optimized text-to-image model

B. User Study Details

For the quantitative evaluation of the 3D views generation task, we recruit 100 volunteers to carry out the survey anonymously. We present the generated visual sequences from our method and other baseline side by side with random positioning. Each questionnaire contains 6 scene comparison pairs with 3 for each of two baselines, randomly chosen from total 50 scene comparison pairs.

C. Additional Results

We present a concrete 3D views generation example to illustrate how to initially construct and later refine the SceneConceptGraph in Figure 2. There, we define v_0 for

Algorithm 2 Concept Relation Refinement

Input: Initial scene image I_0
Additional User Input: Total generation length T with camera pose C_i and text prompt T_i for each timestamp.
Output: Generated 3D view sequence $\mathcal{I}$

$\mathcal{M}_0 \leftarrow BLD(\bar{\mathcal{M}}_0)$ ▷ Blended Latent Diffusion conversion
$t \leftarrow 0$ ▷ Initial timestamp
while $t < T$ **do**
  $D_t \leftarrow \text{DepthEstimation}(I_t)$
  $P_t \leftarrow \text{Unproject}(I_t, D_t)$
  $S_{t+1} \leftarrow S_t \cup P_t$
  $\bar{I}_{t+1}, m_{t+1} \leftarrow \text{Render}(S_{t+1}, C_{t+1})$
  $I_{t+1} \leftarrow \mathcal{M}_t(\bar{I}_{t+1}, m_{t+1}, T_{t+1}, G_t)$ ▷ Outpainting Process
  $v_n \leftarrow$ Get_changed_ or_added_concept() ▷ Refinement Process
  $\bar{G}_{t+1} \leftarrow$ Adjust_or_add_concept-relation_pair(G_t)
  $prompt \leftarrow < v_0, r_{(0,n)}, v_n >$ ▷ Take the concept-relation pair
  $m_u \leftarrow m_0 \cup m_n$
  $I_{gen} = \mathcal{M}_t(\bar{I}_{t+1}, m_{t+1}, prompt)$
  $\mathcal{L} \leftarrow \text{Calculate\_loss}(I_{t+1}, I_{gen}, m_u)$
  $G_{t+1} \leftarrow \text{Optimize\_prompt}(\mathcal{L}, \bar{G}_{t+1})$
  $\mathcal{M}_{t+1} \leftarrow \text{Optimize\_model}(\mathcal{L}, \mathcal{M}_t)$
end while
$\mathcal{I} \leftarrow \{I_1, ..., I_T\}$ ▷ Final generated 3D view sequences

the whole scene, v_1 for one tree, v_2 for another tree, v_3 for the tree area, v_4 for the grassland and later v_5 for the upcoming stream with SceneConceptGraph updated. The number and the content of concepts is determined by users depending on which users expect to maintain in the following generated scenes, and the number of relations are determined by the concept levels. Typically, no more than ten concepts could well represent the scene and the design of hierarchical SceneConceptGraph makes sure that the number of relations is controllable. After defining the scene concepts in the example, we initially obtain 5 relations according to the structure of SceneConceptGraph and later 6 relations in total. We train the text-to-image model on each relation-concept pair with the prompt "$A\ photo\ of\ < v_i, r_{(i,j)}, v_j >$" and we guide the outpainting model to generate views with defined relation-concept pairs. Besides, we can also guide the outpainting model with the prompt containing plausible descriptions on existing concepts or new concepts for desired diversity and conduct test-time finetuning to further define the scene representation. We present concrete SceneConceptGraph visualization, extended views with large perspective shift, and text prompt instructions for scene diversity and controllability in Figure 2.

Furthermore, we show additional 3D views generation results in Figure 3, long-range scene results in Figure 4 and the generated 3D view sequences used to present 3D representation results in the main paper in Figure 5. It can be shown that even when faced with different types of scenes and longer generation duration, our method can still be able to overcome the semantic drift issue and create consistent and visually diverse 3D views. Moreover, we take the 3D views as keyframes to generate long-term 3D videos leveraging pre-trained image-to-video models, and present the videos and 3D construction results of generated 3D views in the attached demo video, which shows the potential application of our method in long-term video generation and VR/AR scenarios.

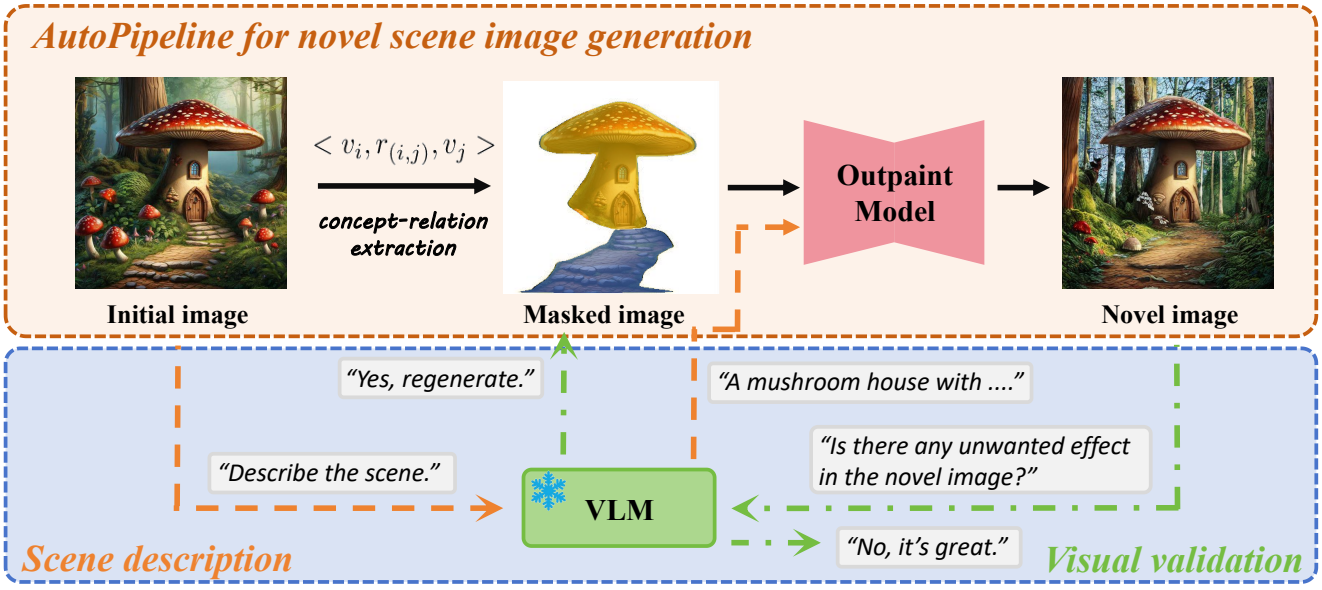

Figure 1. **Automatic pipeline for novel scene images generation.** We use orange dashed to denote the scene description process and green dash-dotted line to represent the visual validation process.

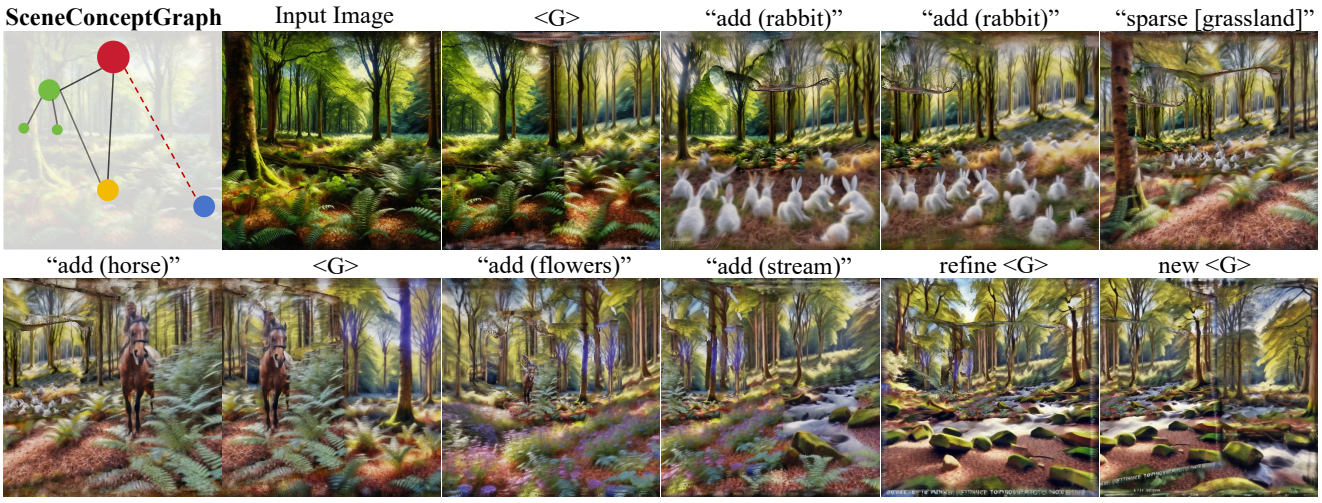

Figure 2. **A concrete 3D views generation example with SceneConceptGraph visualization and text prompt instructions.** We use square brackets to denote defined concepts, parentheses to denote new objects, and angle brackets with G to indicate all defined concept-relation pairs in the scene.

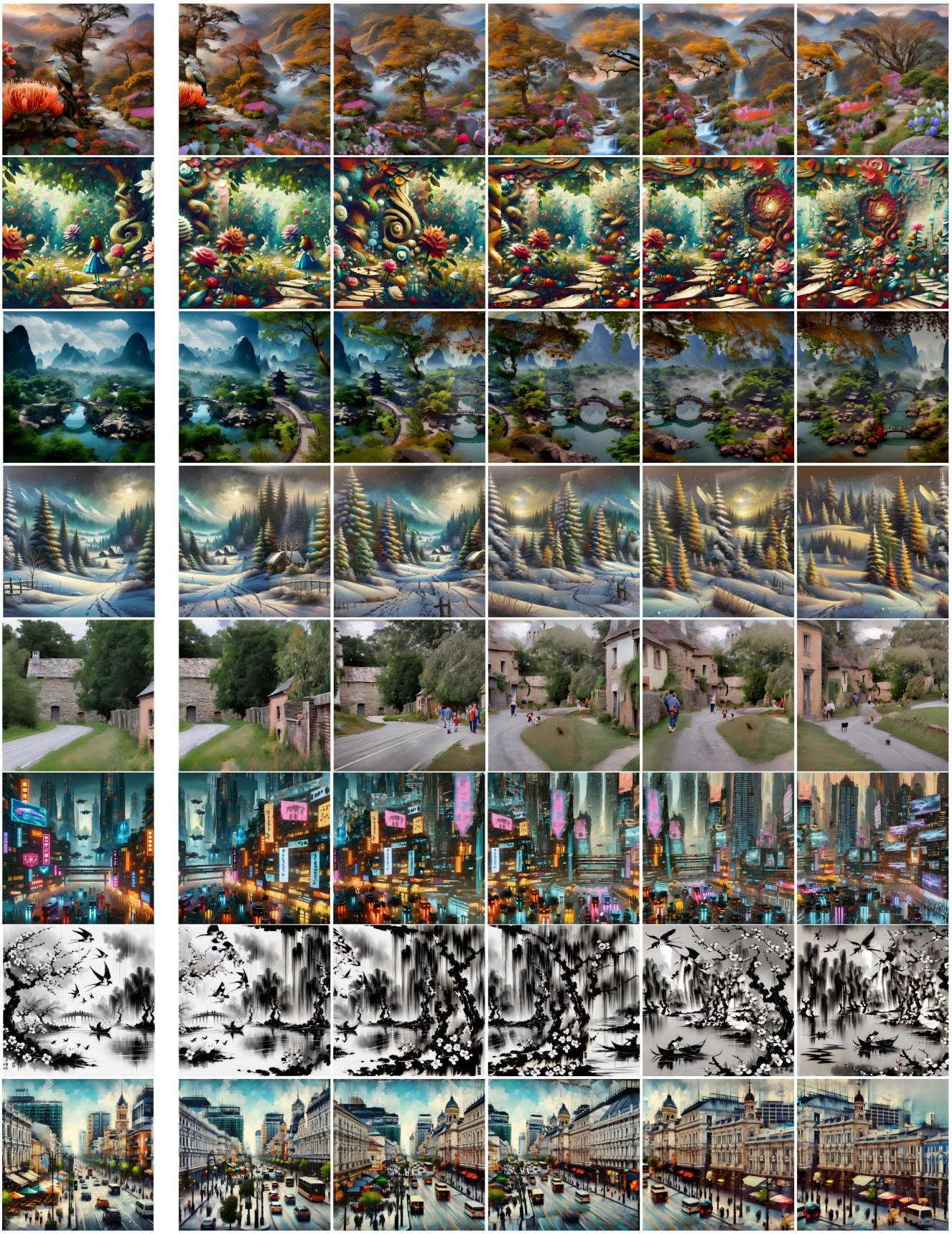

Figure 3. **Additional qualitative results of our generated 3D view sequences.** ScenePainter generates diverse yet coherent 3D scenes based on various scene representation (e.g., nature, village, city, ink painting, cyberpunk, or fantasy).

Figure 4. **Long-range qualitative results with 15 scene views.** We are able to maintain semantically consistent results under extremely large viewpoint shifts, which previous works could not achieve.

Figure 5. **Generated 3D view sequences for 3D representation.** We use the same first view and fixed camera path for evaluation, and the blue dashed box represents the first frame.