A. L-MAGIC Prompts

In Sec. 3.2 we have briefly described how to use language models in L-MAGIC. Here, we provide more details on our prompt design when applying language models.

For line 2 of Alg. 1, we ask the following two questions to the BLIP-2 model ($\mathcal{L}_v(\cdot)$):

Q1$_{\text{BLIP}}$ Question: What is this place (describe with fewer than 5 words)? Answer:

Q2$_{\text{BLIP}}$ Question: Describe the foreground and background in detail and separately? Answer:

These two questions make the model output scene-level coarse and fine descriptions without focusing on centralized objects, which is beneficial for inferring the global scene layout at line 3. The final $d_{\mathcal{I}}$ is the answers of both questions.

To obtain scene layout descriptions d_{360} of individual views, we ask the following question to ChatGPT ($\mathcal{L}(\cdot)$) at line 3:

Q1$_{\text{GPT}}$ Given a scene with <answer of **Q1$_{\text{BLIP}}$**>, where in font of us we see <answer of **Q2$_{\text{BLIP}}$**>. Generate 6 rotated views to describe what else you see in this place, where the camera of each view rotates 60 degrees to the right (you dont need to describe the original view, *i.e.*, the first view of the 6 views you need to describe is the view with 60 degree rotation angle). Dont involve redundant details, just describe the content of each view. Also don't repeat the same object in different views. Don't refer to previously generated views. Generate concise (< 10 words) and diverse contents for each view. Each sentence starts with: View xxx(view number, from 1-6): We see...

As mentioned in Sec. 3.4, ChatGPT sometimes cannot fully follow the format request in **Q1$_{\text{GPT}}$**, which makes automatic prompt generation fail. To avoid this catastrophic failure, we check whether the output of **Q1$_{\text{GPT}}$** has the required number of lines (6), and whether each line starts from '*View XXX (line number): We see*'. We re-run line 3 if any of the condition is violated. This ensures that ChatGPT understands our question and satisfies all our format requests.

To remove object-level information at line 4, we ask:

Q2$_{\text{GPT}}$ Modify the sentence: <answer of **Q1$_{\text{BLIP}}$**> so that we remove all the objects from the description (*e.g.*, 'a bedroom with a bed' would become 'a bedroom'. Do not change the sentence if the description is only an object). Just output the modified sentence.

To adaptively judge whether we should avoid repeated objects, we ask the following two questions at line 5

Q3$_{\text{GPT}}$ Given a scene with <answer of **Q1$_{\text{BLIP}}$**>, where in font of us we see <answer of **Q2$_{\text{BLIP}}$**>. What would be the two major foreground objects that we see? Use two lines to describe them where each line is in the format of "We see: xxx (one object, dont describe details, just one word for the object. Start from the most possible object. Don't mention background objects like things on the wall, ceiling or floor.)"

Q4$_{\text{GPT}}$ Do we often see multiple <each object in the answer of **Q3$_{\text{GPT}}$**> in a scene with <answer of **Q1$_{\text{BLIP}}$**>? Just say 'yes' or 'no' with all lower case letters.

The final d_{repeat} is the set of objects in **Q3$_{\text{GPT}}$** such that the corresponding answer of **Q4$_{\text{GPT}}$** is 'no'.

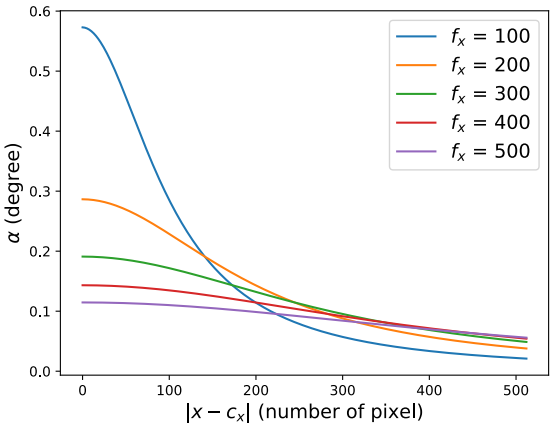


Figure 9. **Angular distance change w.r.t. the pixel location.** We can see that the angular distance α is larger for centered pixels (small $|x - c_x|$, where $|x - c_x|$ represents the horizontal distance from pixel x to the image center c_x) for different focal length values f_x. This phenomenon causes the blurry warping mentioned in Sec. 3.3.

(a) w/o super-resolution. (b) With super-resolution

Figure 10. **Warping with and without super-resolution.** (a) The phenomenon in Fig. 9 causes blurry warping. (b) With super-resolution, we can significantly enhance the sharpness of the warped image.

B. Blur During Warping

As mentioned in Sec. 3.3, adjacent pixels at the center of an image have a large angular distance than the ones at the

side of an image, which causes the blurry warped image. The cause of this issue lies in the construction process of an image. Specifically, let x be the horizontal coordinates of a pixel at the image plane, and let f_x and c_x be respectively the focal length and the principal location (camera center at the image plane) of the camera on the horizontal direction. Then, the horizontal angular distance between the camera rays of x_1 and $x_1 + 1$ is

$$\alpha = |\arctan(\frac{|x+1-c_x|}{f_x}) - \arctan(\frac{|x-c_x|}{f_x})| \quad (2)$$

Fig. 9 shows the change of value α w.r.t. $|x - c_x|$, where large $|x - c_x|$ means x is at the side of an image, and small $|x - c_x|$ means x is at the center of an image. We can see that the angular distance α is larger for centered pixels regardless of the focal length f_x. Hence, within the same angle, there are more pixels on the side of an image than at the center of an image. This means that when warping the center region of an image to another view, we require interpolation since more pixels are created in the corresponding warped region. This phenomenon causes the blurry warping mentioned in Sec. 3.3, see Fig. 10 for an example.

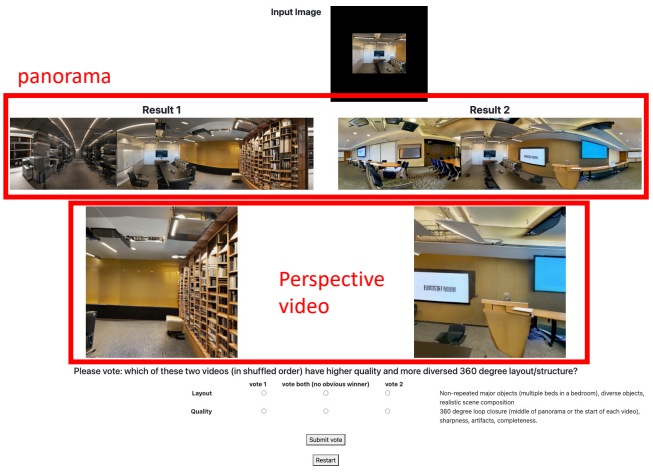
C. Text Inputs for Text-to-panorama

In order to evaluate the performance of different algorithms on in-the-wild inputs, we ask ChatGPT to generate 20 random scene descriptions (10 indoor and 10 outdoor) in the main experiment of text-to-panorama (Sec. 4.2). The resulting text prompts are:

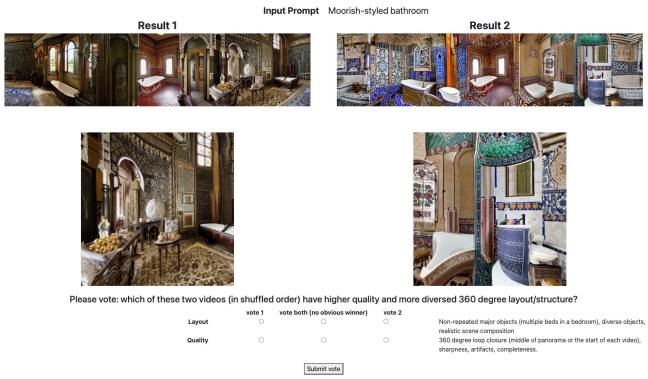
1. Autumn maple forest path.
2. Tropical beach at sunset.
3. Snowy mountain peak view.
4. Tuscan vineyard in summer.
5. Desert under starlit sky.
6. Sakura blossom park, Kyoto.
7. Rustic Provencal lavender fields.
8. Underwater coral reef scene.
9. Ancient Mayan jungle ruins.
10. Manhattan skyline at night.
11. Victorian-era library.
12. Rustic Italian kitchen.
13. Minimalist Scandinavian bedroom.
14. Moorish-styled bathroom.
15. Vintage record store interior.
16. Luxurious Hollywood dressing room.
17. Industrial loft-style office.
18. Art Deco hotel lobby.
19. Japanese Zen meditation room.
20. Modern living room with a sofa and a TV.

D. Voting Web Page

As mentioned in Sec. 4.1, we use a voting web page during human evaluation. Fig. 11 shows the example web page



(a) Image-to-panorama.



(b) Text-to-panorama

Figure 11. **The voting web page for human evaluations.** (a) The web page for image-to-panorama. The outer black region of the input image is the missing region when expanding the field of view. (b) The web page for text-to-panorama. In each voting, we show the panoramas and the rendered perspective videos for two methods. For each criterion, we not only allow to vote for one of the results but also allow to vote for both results when there is no obvious winner.

for both image-to-panorama and text-to-panorama. In each voting, we show for each method a panorama and the perspective video rendered from the panorama so that the user can use the panorama to clearly see the 360 degree layout and loop closure, and use the perspective video to see the rendering quality. Besides voting for one of the two results, we also allow to vote for both results when there is no obvious winner for a certain criterion.

E. Ablation visualizations

In Sec. 4.3, we conducted ablation studies and reported quantitative results. Here we further show visualizations of the ablation experiment in Fig. 12. Consistent with the quantitative results, changing L-MAGIC components hurts

Figure 12. **Visualization of ablation results.** Consistent with the quantitative result in Sec. 4.3, removing individual components of L-MAGIC hurt the performance. A zoom-in comparison is recommended for *No SR* and *No smooth*. In *No SR*, the panorama is less sharp even though the image resolution is the same with L-MAGIC. In *No smooth*, there were two unnatural black lines (zoom in to the bounding boxes) caused by the non-smooth fusion, which we do not observe in the full L-MAGIC method.

the visual quality of the output panorama.

F. Bias in Quantitative Metrics

As mentioned in Sec. 4.3, the Inception Score (IS) sometimes cannot fully reflect the preference from human evaluations. Fig. 13 shows "adversarial" examples where the panorama has poor quality and multi-view coherence yet has a higher inception score compared to the result with better human evaluation preference. This shows the importance of human evaluations in the experiment.

G. Video generation

When generating video frames with pure camera rotation, we follow the strategy of Sec. 3.1, project the panorama to a unit sphere, and render each frame according to the rotation matrix and camera intrinsics of the frame.

To generate immersive videos with camera translations, we apply depth-based warping, and inpaint, using Stable Diffusion v2, the small missing regions caused by occlusion. For depth-based warping, we first apply pre-trained depth estimation models [17] on perspective views of the generated panorama, and warp them to the corresponding frame of the video. Naive mesh-based warping following Sec. 3.1 may generate mesh faces between different objects, which is not ideal. Hence, we rely on point-based warping. To avoid grid-shaped sparsely distributed missing pixels (Fig. 14 left) and ensure the sharpness of the warped image, we apply a super-resolution-based approach similar to the strategy in Sec. 3.3. Specifically, we enlarge the resolution of the depth map from $512 * 512$ to $2048 * 2048$, and then warp the high-resolution depth map to each frame with a resolution of $512 * 512$ (Fig. 14 right).

To achieve super-resolution on the depth map, we first perform super-resolution on the RGB image, increasing its resolution to $2048 * 2048$. Since state-of-the-art depth estimation models are not effective on high-resolution images, instead of directly estimating the depth of the high-resolution image, we separate it into $13 * 13$ patches of resolution $512 * 512$ with overlappings between neighbouring patches, perform depth estimation on individual patches and align the depth map of each patch with the one of the low-resolution image to ensure a smooth depth transition over patches and a reasonable object geometry.

Figure 13. **Example of the biased Inception Score.** We show samples with high inception scores from the experiment in Fig. 6 (Left). Even though our method (right) generates scenes with a much better quality and multi-view consistency, the inception score can still be lower.

Figure 14. **Effect of super-resolution on depth-based warping.** Left: Naive point-based warping generates grid-shaped missing pixels distributed uniformly on the whole image. Right: Super-resolution effectively resolves this issue, making most parts of the warped image sharp and complete.

H. Limitation and Future Work

In terms of the limitation, L-MAGIC currently relies on the input prompt to encode the global scene layout information. Designing a fine-grained layout encoding mechanism that can ensure multi-view coherence at a more detailed level is an important and interesting future work.