Supplementary Material

A. Additional Methodological Details

Laplacian pyramid warping. When an image is warped from one view to another, different regions of the source image may undergo varying amounts of stretching or compression. This leads to frequency mismatch: highly stretched regions tend to lose high frequency details while other regions may retain them. We address this issue by performing Laplacian pyramid warping, as used in LookingGlass [2]. Given a source image, we construct its Laplacian pyramid, where each level isolates a specific frequency band. We then apply the geometric warp independently to every pyramid level. Since each level contains only a narrow range of frequencies, the warp produces smoother and more stable results with fewer large scale distortions. After all levels are warped, we collapse the warped pyramid to obtain the final aligned image. In our implementation, we use a five level Laplacian pyramid for all warping operations.

Time travel. Similar to LookingGlass [2], we also adopt the time travel strategy [7] to allow the model to blend different views better. The idea is to travel by l steps to a noisier timestep and then let the cleaner state $z_{0|t}$ guide the update, since using the cleaner prediction $z_{0|t}$ from the current timestep provides a more reliable estimate than $z_{0|t+l}$. At a chosen timestep t, we first sample a latent state at a noisier timestep $t + l$ using the transition distribution $q(z_{t+l} \mid z_t)$. We then restart the reverse process from $t + l$ and continue denoising to $t-1$. This operation effectively rewrites the recent sampling history, using a more reliable clean prediction at timestep t to guide subsequent updates. Repeating this procedure improves global consistency around challenging regions. In our implementation, we follow LookingGlass and apply time travel only between 20% and 80% of the sampling steps, with a travel length of one, and we repeat the procedure three times.

Foreground object relighting. To enhance realism, we relight the foreground object using a relighting (harmonization) model [5]. We provide a text description of the object and its desired illumination, and the model creates the soft shadows and caustics accordingly. We repeat 20 times of this process and choose the one that minimally changes the appearance w.r.t. PSNR in the transparent object region. For instance, for a glass sphere on the table of an artroom scene, we use a prompt such as "*Creating soft shadows and caustics under the glass sphere on the table.*" This relighting step improves the coherence between the inserted transparent object and the surrounding generated scene.

B. From Text to Object Parameters

For each scene category, we use the prompts p, p^-, p^{360}, and p^r to guide the generation process, as illustrated in Figure 1. The full prompt p describes the scene together with the transparent object, while the object free prompt p^- removes the object and instead specifies that the supporting surface is clean and empty. The panorama prompt p^{360} describes a high resolution equirectangular 360 degree panorama captured from the center of the transparent object, providing additional scene context beyond the perspective view. The relighting prompt p^r adds physically plausible shadows and caustics using the instruction "*Create soft shadows and caustics under* p^{obj}".

Given a text prompt p that describes a scene containing a transparent object, we derive the three components required by our method: a 3D model of the object, its refractive index, and its pose relative to the camera. These are obtained automatically using a sequence of lightweight language model queries together with a text to 3D generator.

Prompt decomposition. Starting from the full prompt p, we ask the large language model ChatGPT [6] to identify the span of text that refers to the transparent object. For the first prompt describing the living room in Figure 1, the LLM identifies the removable phrase "*with a big glass sphere on top*". From this phrase, we extract the core object description "*a glass sphere*", which we denote as p^{obj}. To construct the object free prompt p^-, we replace the entire identified phrase "*a big glass sphere on top*" with "*surface clean and empty*", while keeping all other scene details unchanged.

Refractive index extraction. We obtain the refractive index by querying the LLM directly with the object description p^{obj}. For common materials such as glass or water, the returned values are well defined and stable across queries. In the example above, asking "*What is the refractive index of a glass sphere*" typically returns a value of 1.5.

3D model generation. In the main paper, we use a text-to-3D generator TRELLIS [9] to produce high quality meshes that follow p^{obj}. The additional examples given in the supplement come from the RefRef dataset [10], which provide clean and watertight meshes. However, we have verified that TRELLIS can also generate suitable meshes for the associated prompts. An example is given in Figure 2.

Object pose estimation. To place the transparent object into the scene, we apply SAM 3 [1] to segment the horizontal surface referred to in the text prompt, and place the bottom center of the 3D model generated by TRELLIS [9] at the center of the surface.

	Full Scene Prompt p	Panorama Prompt p^{360}
Living Room	*A cozy living room with a polished wooden coffee table close to the camera,* ***with a big glass sphere on top,*** *surrounded by a beige sofa, a patterned rug, plants, bookshelves, framed wall art, and sunlight through sheer curtains.*	*A high resolution equirectangular 360 degree panorama captured on top of a polished wooden coffee table in a cozy living room.*
Dining Room	*A bright dining room with a wooden dining table in the center,* ***with a big glass sphere on top,*** *surrounded by upholstered chairs, a pendant lamp, fruit bowls, paintings, and daylight through tall windows.*	*A high resolution equirectangular 360 degree panorama captured on top of a wooden dining table in a bright dining room.*
Office	*A minimalist home office with a smooth wooden desk closer to the camera,* ***with a big glass sphere on top,*** *surrounded by a black office chair, bookshelves with plants, framed posters, a side table with a monitor.*	*A high resolution equirectangular 360 degree panorama captured on top of a smooth wooden desk in a minimalist home office.*
Kitchen	*A modern kitchen with a large marble island in the center,* ***with a big glass sphere on top,*** *surrounded by wooden cabinetry, bar stools, hanging lights, utensils, and reflections from stainless steel appliances under morning light.*	*A high resolution equirectangular 360 degree panorama captured on top of a marble island in a modern kitchen.*
Artroom	*An art classroom with a rectangular wooden worktable near the camera,* ***with a big glass sphere on top,*** *surrounded by easels, color splattered stools, sketches, jars of brushes, and warm daylight through wide windows.*	*A high resolution equirectangular 360 degree panorama captured on top of a wooden worktable in an artroom.*
Café	*A minimalist café interior with a square wooden table in the foreground,* ***with a big glass sphere on top,*** *surrounded by metal framed chairs, plants, hanging lights, a pastry counter, and sunlight on the tiled floor.*	*A high resolution equirectangular 360 degree panorama captured on top of a square wooden table in a minimalist café interior.*
Cave	*A rocky cave lit by a bright campfire, warm flickering light casting shadows,* ***a big glass sphere on the ground,*** *with scattered camping gear, tents, sleeping bags, backpacks, lanterns, and cooking pots, with smoke and embers in the air.*	*A high resolution equirectangular 360 degree panorama captured from the ground of a rocky cave, camping gear scattering around.*
Desert	*A high-noon desert scene with blinding sunlight and hard shadows, heat haze over sand and rocks,* ***a big glass sphere on the ground,*** *and camping gear in the foreground, tent, backpacks, and a small stove.*	*A high resolution equirectangular 360 degree panorama captured from the ground of a desert scene, with camping gear around.*
Karaoke	*A karaoke room with colorful lights, TV on the wall displaying music videos, a coffee table* ***with a big glass sphere on top,*** *in front of the TV, and a sofa around the coffee table.*	*A high resolution equirectangular 360 degree panorama captured from on top of a coffee table in a colorful karaoke room.*
Landscape	*A beautiful landscape with a river and mountains, viewed from a camera directly in front of a stone table and chairs in the foreground, filling the lower frame,* ***a big glass sphere on the table,*** *on the left of colorful food.*	*A high resolution equirectangular 360 degree panorama captured from on top of a stone table of a beautiful landscape.*

Figure 1. Prompts used for generating scenes. The full prompts p are shown on the left, where the transparent object phrases are highlighted in red bold text. The object free prompts p^- are obtained by replacing the transparent object description in each full prompt with wording that states the surface is clean and empty, while keeping all other scene details unchanged. The right column shows the panorama prompts p^{360}. The relighting prompt p^r is: *Create soft shadows and caustics under* p^{obj}.

C. Extended Experimental Results

C.1. Qualitative Results

In this section, we provide additional qualitative and quantitative results that complement the main paper. We present qualitative results for additional transparent objects, analyze the effect of varying refractive indices (e.g., water, plastic), provide the full per-scene quantitative comparison table, a quantitative ablation study table, and study the effect of different time travel repeat counts. In our implementation, generating a perspective–panorama pair requires approximately 126 seconds on an 80 GB NVIDIA A100 GPU (preprocessing time not included).

Complex transparent objects. We provide extended qualitative comparisons across a variety of transparent objects. Figures 3 and 4 show generated results for five additional objects. From these examples, we observe that our method continues to produce refractions that are visually correct and physically plausible, closely matching the Blender reference images. In contrast, the Flux inpainting model struggles considerably: while the main paper showed that it can sometimes produce a glass sphere, albeit with incorrect physics, it consistently fails once the object shape becomes more complex, even when provided with the exact foreground mask. The standard Flux model is more capable of generating glass-like objects, yet its outputs often ex-

Table 1. Full per-scene quantitative comparison of our method against FLUX-based inpainting model [5], FLUX-dev [5], FLUX.2-dev [4], Qwen-Image [8], and Stable Diffusion 3.5 (Large) [3]. The table is split into two sections to maintain readability in portrait orientation. Higher CLIP, ImageReward, and PSNR are better, while lower LPIPS is better. Our method consistently outperforms the baselines in terms of refraction fidelity (measured by PSNR and LPIPS), without sacrificing text alignment (CLIP).

	CLIP ↑						ImageReward ↑					
Scene	Inpaint	FLUX	FLUX2	Qwen	SD3.5	Ours	Inpaint	FLUX	FLUX2	Qwen	SD3.5	Ours
Artroom	**35.06**	31.81	33.44	34.46	34.37	34.12	0.14	0.09	**0.80**	0.10	0.52	0.30
Cafe	32.55	30.95	31.11	31.36	**33.72**	30.55	-1.42	-1.25	-0.85	**-0.78**	-1.00	-1.26
Cave	36.85	36.87	36.89	36.14	**37.77**	36.83	-0.28	-0.14	**0.24**	-0.07	-0.03	-0.24
Desert	34.93	35.26	35.02	35.22	35.44	**35.73**	0.67	1.09	**1.31**	1.01	0.97	0.79
Dining Room	**32.53**	28.15	30.76	30.03	32.19	32.04	-0.87	-0.40	**0.43**	-0.25	-0.04	-0.34
Karaoke	36.22	34.01	**37.49**	34.64	37.18	35.87	0.21	0.36	**1.18**	0.75	0.98	0.43
Kitchen	30.03	32.96	**33.09**	32.42	32.70	30.29	-0.77	0.30	**0.51**	0.39	-0.23	-0.71
Landscape	34.02	31.66	33.69	30.63	**35.69**	35.04	0.02	-0.99	-0.88	0.05	**0.52**	-0.18
Living Room	29.74	29.77	29.01	29.03	**32.74**	28.44	-1.18	-0.15	-0.07	-0.66	**0.10**	-0.73
Office	32.51	32.15	33.16	32.78	**33.82**	29.58	-1.22	-0.96	-0.88	**-0.47**	-0.68	-1.29
Average	33.44	32.36	33.37	32.67	**34.56**	32.85	-0.47	-0.20	**0.18**	0.01	0.08	-0.32

	PSNR ↑						LPIPS ↓					
Scene	Inpaint	FLUX	FLUX2	Qwen	SD3.5	Ours	Inpaint	FLUX	FLUX2	Qwen	SD3.5	Ours
Artroom	13.50	13.43	12.89	12.93	13.25	**16.67**	0.47	0.47	0.43	0.45	0.52	**0.27**
Cafe	11.75	11.81	11.21	11.69	11.46	**15.65**	0.44	0.45	0.43	0.45	0.53	**0.23**
Cave	13.76	14.05	13.58	13.66	13.67	**18.91**	0.50	0.50	0.49	0.50	0.56	**0.25**
Desert	15.26	16.36	14.31	15.04	14.05	**18.71**	0.35	0.35	0.41	0.38	0.43	**0.20**
Dining Room	14.20	14.09	14.02	14.26	14.43	**15.14**	0.44	0.43	0.44	0.43	0.49	**0.21**
Karaoke	11.01	10.93	10.21	10.79	10.29	**17.09**	0.53	0.52	0.56	0.54	0.57	**0.24**
Kitchen	11.79	11.72	11.51	11.82	11.63	**15.65**	0.51	0.55	0.55	0.53	0.58	**0.27**
Landscape	11.28	11.12	10.70	12.09	10.85	**15.05**	0.47	0.51	0.49	0.49	0.53	**0.29**
Living Room	12.83	12.07	12.17	12.26	12.00	**16.51**	0.48	0.50	0.53	0.49	0.55	**0.23**
Office	11.26	11.24	10.86	10.98	10.88	**15.71**	0.48	0.49	0.49	0.54	0.56	**0.23**
Average	12.66	12.68	12.15	12.55	12.25	**16.51**	0.47	0.48	0.48	0.48	0.53	**0.24**

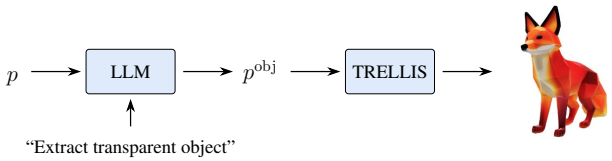

Figure 2. An example of extracting the object specific prompt p^{obj} using an LLM and generating a 3D mesh with TRELLIS. From the full scene prompt "*An art classroom with a rectangular wooden worktable near the camera, with a glass fox on top, surrounded by easels, color splattered stools, sketches, jars of brushes, and warm daylight through wide windows*" the LLM identifies "*a glass fox on top*" and distills the object description "*a glass fox*" as p^{obj}, which is then provided to TRELLIS to produce the 3D fox mesh.

hibit low transmittance or hollow interiors that do not correspond to the appearance of real transparent materials. These results collectively highlight the robustness of our method across diverse geometries and the difficulty existing generative models face in handling complex refractive behaviors.

Effect of refractive index. We analyze the effect of varying the refractive index for a glass sphere in Figure 5. This highlights the sensitivity of appearance to the refractive index and illustrates the importance of accurate material estimation when synthesizing transparent objects.

Time travel analysis. Finally, we analyze the effect of repeat counts in time travel. More repeats generally produce smoother results but suppress high frequency details, making the generated image appear overly uniform and less realistic. We first evaluate configurations where the repeat count is applied equally to both the perspective and panorama view ($R_{\text{main}} = R_{\text{pano}} \in \{1, 3, 5, 8\}$). As shown in Figure 6, increasing the repeat count does improve global blending and reduces artifacts like random spots. However, we observe that it becomes overly aggressive on the panorama branch. Since the panorama covers a much wider field of view with rich global structure, repeated smooth-

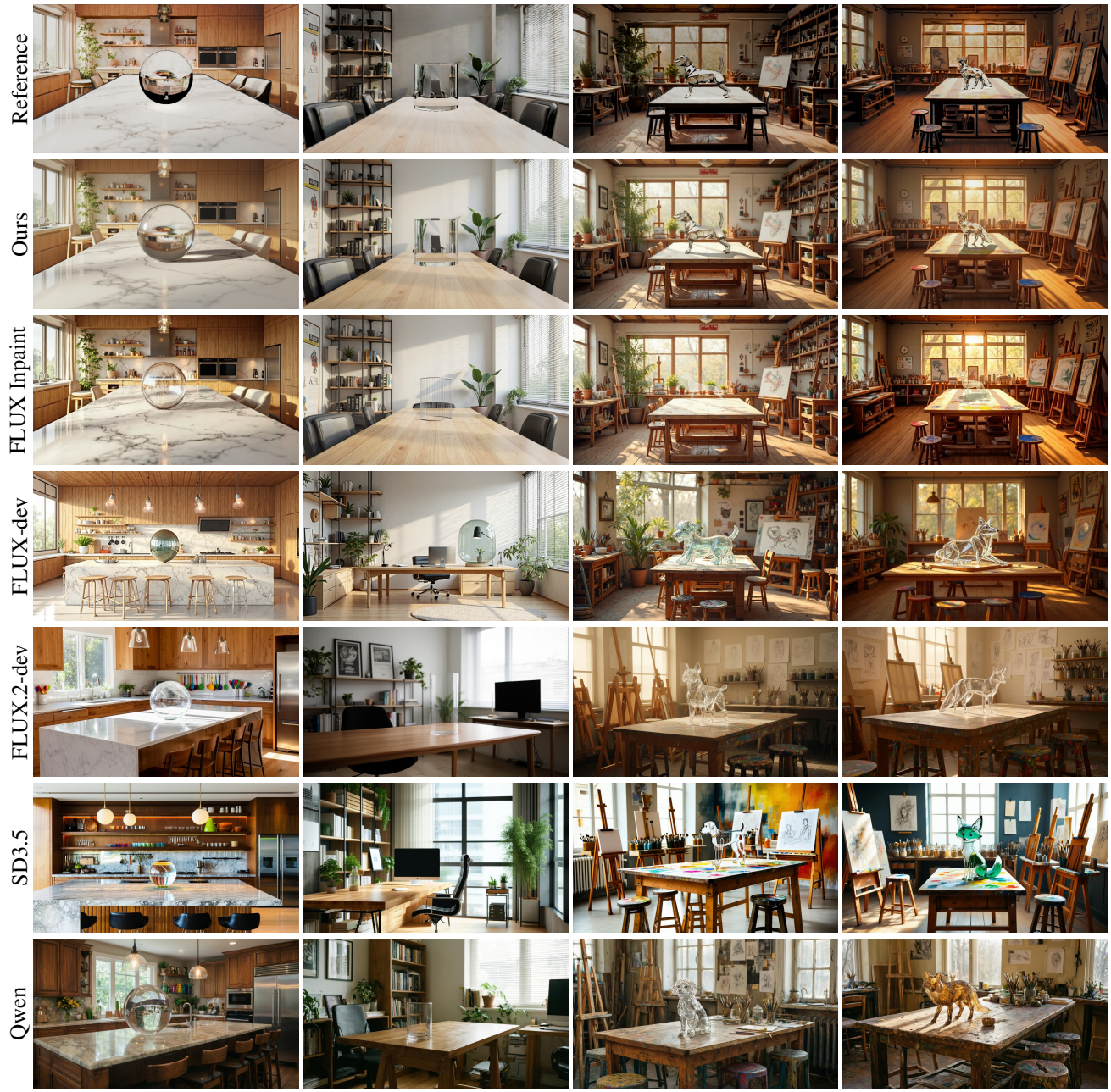

Figure 3. Additional qualitative results of our method against FLUX-dev [5], FLUX.2-dev [4], Qwen-Image [8], Stable Diffusion 3.5 (Large) [3], and FLUX-based inpainting model [5], on a diversity of shapes. These examples illustrate the robustness of our method across diverse object geometries. This figure is the first part of a two page layout.

ing removes important details and leads to blurry, unrealistic results. To address this issue, we apply time travel only to the perspective view while disabling it for the panorama ($R_{\text{main}} \in \{1, 3, 5, 8\}$ while $R_{\text{pano}} = 1$). This retains the benefits of smoothing and stabilizing the perspective branch while preserving the rich high frequency information present in the panorama. As illustrated in the second half of Figure 6, this configuration yields cleaner and more consistent results without sacrificing panoramic detail, offering a better balance between smoothness and realism.

C.2. Quantitative Evaluation

Full scene comparison. We provide the full per-scene quantitative comparison in Tab. 1, for every indoor scene prompt. Across CLIP, PSNR, and LPIPS, our method achieves consistently stronger results than both baselines,

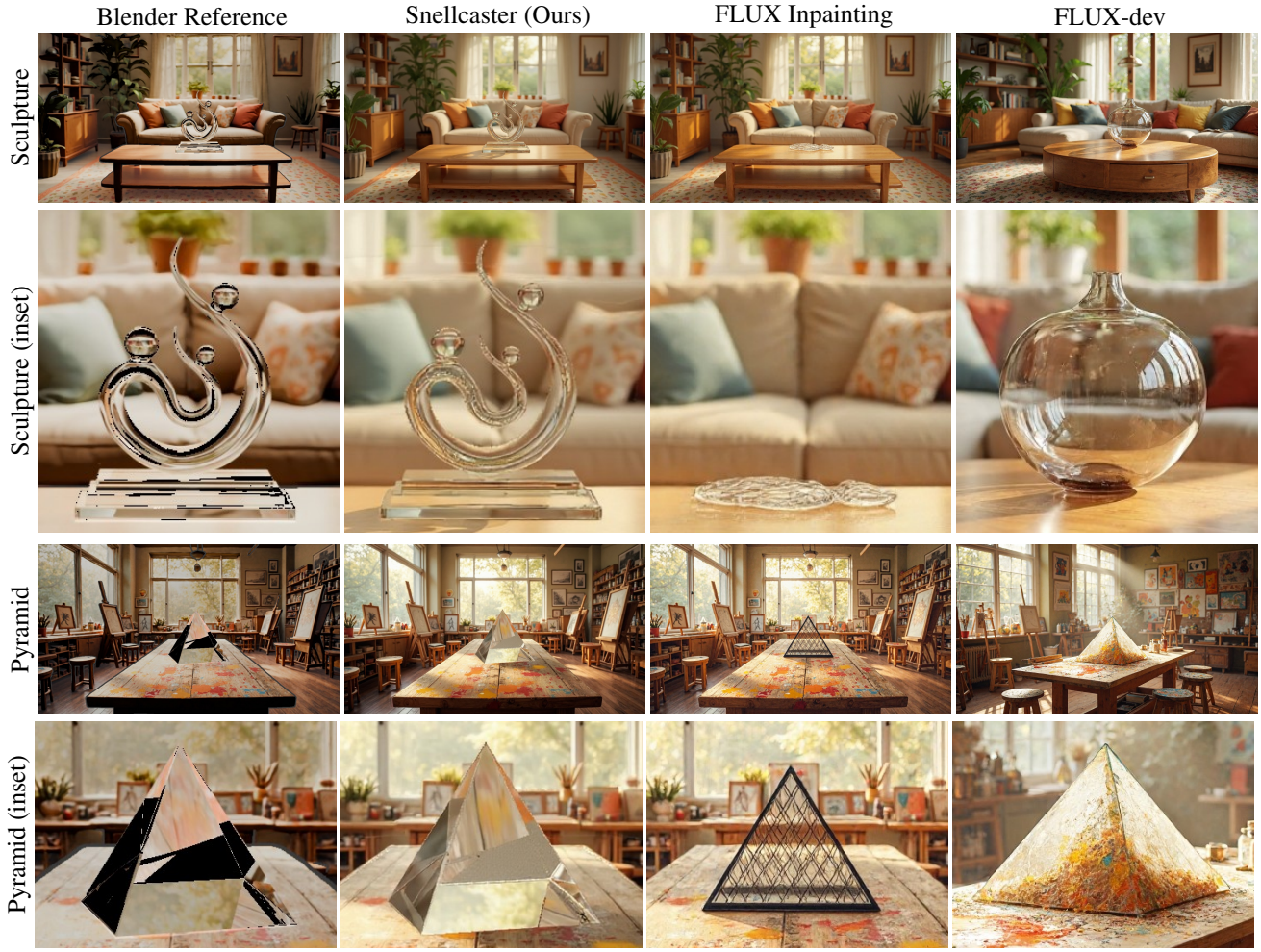

Figure 4. Additional qualitative results on complex transparent objects (continued). This second part supplements Figure 3 and is separated only for page layout. The results further demonstrate that our method maintains physically plausible refraction across different shapes, while existing FLUX variants fail to handle challenging geometries.

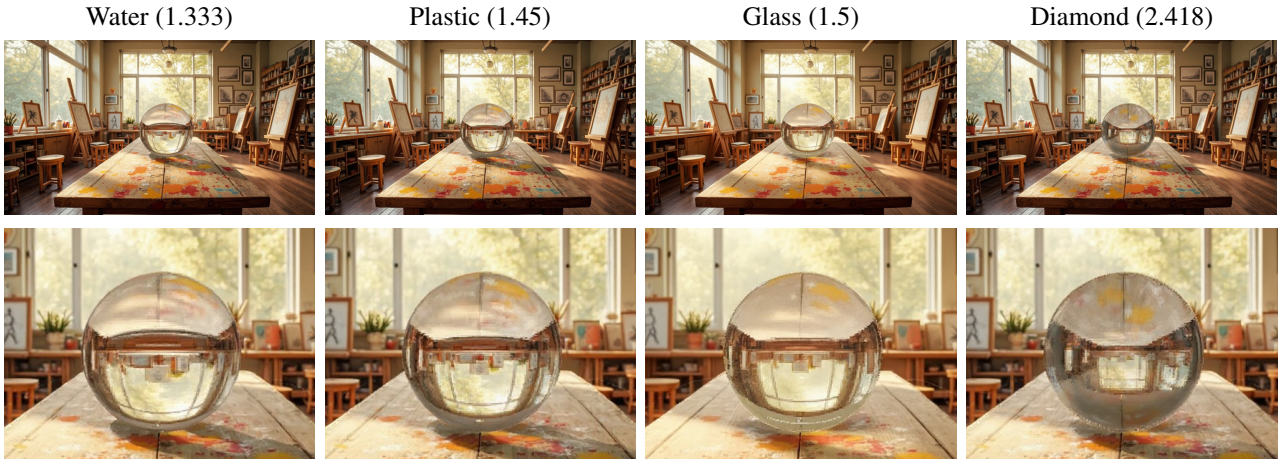

Figure 5. Appearance variation under different refractive indices for a glass sphere. We render the same scene while sweeping the refractive index and show the corresponding outputs from Snellcaster (ours). The results demonstrate the strong sensitivity of transparent object appearance to the refractive index and highlight the importance of accurate material estimation for physically consistent generation.

Figure 6. Qualitative comparison of time travel repeat counts. The top half of the figure shows configurations where the same repeat count is applied to both the perspective and panorama views ($R_{\text{main}} = R_{\text{pano}}$). The bottom half shows a setting where time travel is applied only to the perspective view while the panorama is left unmodified ($R_{\text{main}} \in \{1, 3, 5, 8\}, R_{\text{pano}} = 1$).

with particularly large gains in PSNR and LPIPS, which measure physical consistency with the Blender reference images. On these metrics, our approach performs significantly better than the baselines on every single scene. For ImageReward, the standard Flux model obtains a slightly higher score, although the difference is negligible. This is expected because ImageReward measures prompt alignment rather than physical plausibility, and our prompts mention many scene elements, so variations outside the object region can influence the score.

Additional ablation study. An additional ablation study is performed where we evaluate the effect of three components: detail-preserving averaging, Laplacian pyramid warping, and time travel. Qualitative results are shown in Figure 7. To better visualize the impact of each component, we show two zoomed-in regions of an artroom scene without applying foreground object relighting. This avoids differences caused by relighting itself, which can introduce slight variations even when using the same random seed, since the input images differ slightly. In the second row, we observe that removing either detail-preserving averaging or Laplacian pyramid warping leads to notably rough edges

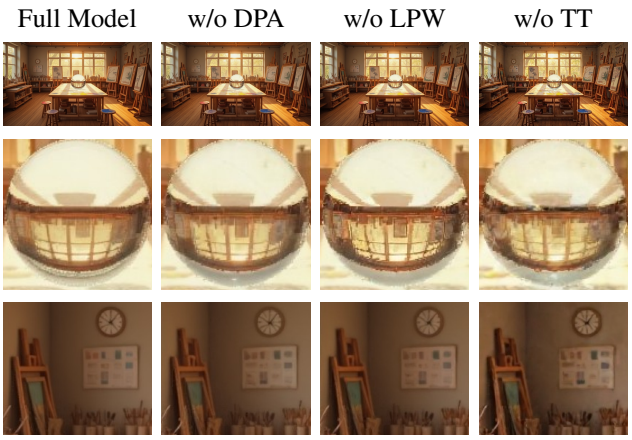

Figure 7. Additional ablation study of the proposed method. We compare the full model with variants that remove individual components: detail-preserving averaging, Laplacian pyramid warping, and time travel. The results are displayed without foreground object relighting, allowing the effects of each component removal to be observed more clearly. Removing detail-preserving averaging leads to the loss of sharp details, removing Laplacian pyramid warping introduces aliasing in regions with large stretching, and removing time travel makes the cross-view blending noticeably less natural with strong artifacts. The full model avoids these issues and yields sharper and more coherent results.

and artifacts. In particular, removing Laplacian pyramid warping introduces aliasing. Removing time travel, which is designed to improve blending consistency, produces obvious artifacts, including random spots, ghosting, and loss of detail. In contrast, the full model shows negligible aliasing within the sphere, and the refractions appear smooth and physically plausible. The third row highlights background regions outside the sphere, where the advantages of the full model are again evident. It produces smooth results while the other variants exhibit rough regions, noise, and visual artifacts. This shows that each component contributes meaningfully to the overall image quality and consistency.

Another quantitative ablation study is given in Tab. 2, showing how removing each component affects performance. We compare the full model without relighting with variants that remove individual components: detail-preserving averaging (DPA), Laplacian pyramid warping (LPW), and time travel (TT). First, a note of caution about the interpretation of the metrics. Those that evaluate the quality of the refracted region—the masked mean absolute error, the peak signal-to-noise ratio, and the LPIPS distance—are all in reference to the Blender-rendered image. However, this pseudo-ground truth is not harmonized with respect to lighting or reflection. That is, the light sources are in significantly different locations and have different colors, and the reflected background is entirely missing. This means that the refracted region in the "ground truth" has significant expected differences in color, luminance, and structure from the truth, and so these metrics are not a fully reliable guide to performance. Moreover, these metrics do not capture effects such as aliasing, which can be observed in the qualitative ablation study (Figure 7). Nonetheless, we observe small drops in quantitative performance when the different model components are removed, especially when time travel is ablated. Interestingly, and as previously observed, time travel smooths the images to some extent, which is helpful for the pixel-level metrics but harmful for the image-level ones (e.g., CLIP, ImageReward). On the whole, we direct the reader to the visual ablation study in Figure 7, where the actual effects of each component are more obvious.

Table 2. Additional quantitative ablation study of the sphere object across six scenes (artroom, café, living room, dining room, kitchen, and office). We compare the full model without relighting with variants that remove individual components: detail-preserving averaging (DPA), Laplacian pyramid warping (LPW), and time travel (TT). We report the masked mean absolute error (MAE), the peak signal-to-noise ratio, and the LPIPS distance, which all measure the fidelity of the refractive region to the Blender-rendered pseudo-ground truth, as well as the CLIP and ImageReward (ImgR) scores, which assess whether the overall image is reasonable. We note that the Blender image is not harmonized with respect to lighting or reflection, so the refracted region has significant expected differences in color, luminance, and structure, and so these metrics are not a fully reliable guide. We also note that these metrics cannot capture effects like aliasing, which can be observed in Figure 7.

Model Variant	MAE↓	PSNR↑	LPIPS↓	CLIP↑	ImgR↑
Ours	**0.0953**	<u>18.21</u>	<u>0.24</u>	<u>34.22</u>	<u>0.51</u>
w/o DPA	0.0955	18.15	<u>0.24</u>	34.18	0.46
w/o LPW	0.0957	**18.23**	**0.23**	34.08	0.46
w/o TT	0.1002	17.82	0.26	**34.56**	**0.58**

References

[1] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. 1

[2] Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, and Vinicius Azevedo. LookingGlass: Generative anamorphoses via laplacian pyramid warping. In *CVPR*, pages 24–33, 2025. 1

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling Rectified Flow Transformers for High-resolution Image Synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4

[4] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. 3, 4

[5] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 3, 4

[6] OpenAI. Chatgpt. https://openai.com, 2025. Version 5.1. 1

[7] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 1

[8] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3, 4

[9] 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 *CVPR*, pages 21469–21480, 2025. 1

[10] Yue Yin, Enze Tao, Weijian Deng, and Dylan Campbell. Refref: A synthetic dataset and benchmark for reconstructing refractive and reflective objects. *arXiv preprint arXiv:2505.05848*, 2025. 1