

A ADDITIONAL QUALITATIVE RESULTS

In this section, we present additional qualitative results from experiments conducted using our GSO-based evaluation datasets. Fig. A1 displays these additional qualitative outcomes. Textured meshes and their corresponding normal maps are generated to evaluate 3D shape quality and assess multi-view consistency.

As illustrated in fig. A1, our proposed method demonstrates better performance compared to the baselines with respect to shape reconstruction accuracy and exhibits significantly fewer artifacts in both texture and shape reconstruction outcomes.

B VIEW COVERAGE AND RECONSTRUCTION QUALITY

For feed-forward image-to-3D methods, the quality of the 3D reconstruction varies based on the number of input views and the extent of view coverage. In this section, we conducted experiments to assess how reconstruction quality changes under different view coverage conditions. We evaluate with 2-view, 3-view, 4-view, 6-view, and 8-view scenarios.

Table A1: Analysis regarding the reconstruction quality with the number of input views.

	CD↓	FS↑	PSNR↑	SSIM↑	LPIPS↓
2-view	0.156	0.456	20.411	0.896	0.146
3-view	0.080	0.645	23.708	0.912	0.117
4-view	0.065	0.706	25.746	0.923	0.101
6-view	0.063	0.717	26.373	0.926	0.098
8-view	0.063	0.722	26.433	0.926	0.098

As shown in table A1, we observed that reconstruction quality increases as the number of input views increases, regardless of whether the model is trained only on 4-view inputs. This result verifies the scalability of our model to more input views at inference time, which strengthens the practicality of the model.

C DETAILED INFORMATION OF TEXT PROMPTS

For the text-conditioned Stable Diffusion 2 (Rombach et al., 2022) Inpainting baseline, we generated text prompts in the format “A *photo of* {description}”, where the *description* consists of the object description for each item from the Google Scanned Objects dataset (Downs et al., 2022). Specifically, we used the first sentence of the model description provided in the GSO dataset as the object description for our GSO-based evaluation dataset. Examples of these text prompts are presented in table A2, where the object descriptions are highlighted in green.

D ANALYSIS OF INFERENCE TIME AND MEMORY

We assessed each model’s inference time and VRAM usage to evaluate the computational cost of our approach compared to the baselines. We recorded the time taken to synthesize four input view images and twelve novel view images, amounting to a total of sixteen images. Furthermore, we included the inpainting duration for those methods that use inpainting. For image-to-3D inference time, we record the time for only model inferences, except for post-processing such as view rendering or mesh extraction. Additionally, we calculated the average VRAM consumption based on a sampled subset from our memory usage evaluation dataset to assess memory efficiency.

As shown in table A3, inpainting-based methods require extra seconds to perform inpainting with the obstacle mask using Stable Diffusion. Inpainting-based baselines require considerably more VRAM for generation with Stable Diffusion, whereas our pipeline with masked multi-view Transformer demands only a slight increase in VRAM. Also, table A3 shows that our model is more efficient than the DiT-based generative model (Wu et al., 2025) in both inference time and VRAM usage.

Table A2: Examples of the text prompts for inpainting on our GSO-based evaluation dataset.

GSO Object Name	Text Prompt Condition for Inpainting
BUNNY_RATTLE	<i>A photo of a BUNNY RATTLE</i>
TZX_Runner	<i>A photo of a T-ZX Runner</i>
JS_WINGS_20_BLACK_FLAG	<i>A photo of a JS WINGS 2.0 BLACK FLAG</i>
JarroSil_Activated_Silicon_5exdZHIeLAp	<i>A photo of a JarroSil, Activated Silicon</i>
CONE_SORTING_kg5fbARBwts	<i>A photo of a CONE SORTING</i>
ZX700_mzGbdP3u6JB	<i>A photo of a ZX700</i>
Now_Designs_Bowl_Akita_Black	<i>A photo of a Now Designs Bowl, Akita, Black</i>
Spritz_Easter_Basket_Plastic_Teal	<i>A photo of a Spritz Easter Basket, Plastic, Teal</i>
FemDophilus	<i>A photo of a Fem-Dophilus</i>
ASICS_GELLinksmaster_WhiteCoffeeSand	<i>A photo of a ASICS GEL-Linksmaster - White/Coffee/Sand</i>
TERREX_FAST_X_GTX	<i>A photo of a TERREX FAST X GTX</i>

Table A3: Analysis on inference time and GPU memory usage for our method and baselines.

	Inpainting (s)	Image-to-3D (s)	Total (s)	Inpatining (GB)	Image-to-3D (GB)
Original LaRa (Chen et al., 2024a)	–	1.71	1.71	–	3.94
SD (Image Cond) (Rombach et al., 2022)	8.90	1.71	10.61	5.96	3.94
SD (Text Cond) (Rombach et al., 2022)	8.78	1.71	10.49	5.96	3.94
Grid Prior (Weber et al., 2024)	23.08	1.71	24.79	3.34	3.94
pix2gestalt (Ozguroglu et al., 2024)	29.92	1.71	31.63	11.53	3.94
Amodal3R (Wu et al., 2025)	–	9.303	9.303	–	13.75
Ours	–	2.30	2.30	–	5.47

E LIMITATIONS

Our method requires precise camera poses linked to sparse images, which may restrict its practical usability. Furthermore, our pipeline depends on an instance-level segmentation map for occlusion representation. We have found that SAM masks and COLMAP poses can be effectively applied to real-world data in certain contexts, particularly with object-centric data. To overcome this issue, we believe that developing an end-to-end framework that simultaneously estimates camera poses and object masks would be a promising direction for the future.

Additionally, since our model is based on an MAE-like structure, it can accurately generate the shape; however, it often produces a blurry texture for the occluded region. We believe that using texture-only generation methods or increasing the number of viewpoints can refine textures while still ensuring the practicality of our model.

F DECLARATION OF LLMs IN THE WRITING PROCESS

During the preparation of this work, the authors used Claude and Gemini to refine the writing. After using this tool, the authors rigorously reviewed and edited the content as needed, taking full responsibility for the content in the publication.



Figure A1: Additional qualitative results on our GSO-based evaluation dataset. We highlight the target object of each scene by intentionally brightening the non-target regions in the input images to enhance visibility. Please zoom in for more details.