

Supplementary Materials for *Martian World Models: Controllable Video Synthesis with Physically Accurate 3D Reconstructions*

A Additional Implementation Details

A.1 Automated Data Filtering Strategies

This subsection details the automated filtering pipeline applied to the raw Martian stereo image pairs obtained from the NASA Planetary Data System (PDS). The following paragraphs describe the specific filtering techniques employed:

Filtering Low-Quality Thumbnails and Grayscale Images. To eliminate uninformative or non-representative visual data, we begin by removing low-resolution thumbnails and grayscale images. This step relies on file-level heuristics including image resolution, file size, and RGB channel statistics. Images with dimensions significantly smaller than our expected minimum resolution or with anomalously small file sizes are classified as thumbnails and excluded. Additionally, we assess the variance across the RGB channels to detect grayscale images; those with minimal inter-channel variance are removed, as they lack the color information necessary for downstream multimodal analysis.

Removing Redundant Content via Perceptual Hashing. To prevent content-level redundancy caused by multiple captures of the same scene under different imaging conditions—such as varied white balance, contrast, or filters—we apply perceptual hashing. This technique generates compact hash codes that encode high-level structural similarity. By computing Hamming distances between these hashes, we identify visually near-duplicate images and discard those exceeding a similarity threshold. This step ensures the dataset maintains semantic diversity and avoids overrepresentation of particular scenes or textures.

Excluding Blurry and Low-Sharpness Images. Maintaining geometric fidelity is critical for tasks such as stereo reconstruction and surface normal estimation. To this end, we apply a sharpness filter using Laplacian variance, a well-established metric that quantifies edge contrast within an image. Frames with a variance below a pre-defined threshold are classified as blurry and automatically excluded. These typically result from motion blur or poor focus, and retaining them would degrade the quality of both photometric and geometric outputs in the synthesized dataset.

Filtering Out Visually Unusable Frames. We analyze the average color intensity histograms of each image to identify those dominated by irrelevant content such as large spacecraft segments, occlusions, or camera malfunctions. These images often display skewed or flat histogram profiles, indicating uniform color patches or unnatural saturation patterns. We flag and remove such frames to ensure that the final dataset primarily comprises clear, terrain-focused scenes with minimal visual obstruction. This enhances the quality and consistency of the data available for learning meaningful Martian representations.

A.2 Implementation of Grounded-SAM-assisted Semi-Automated Data Preprocessing

While the initial automated filtering pipeline, as described in Sec. A.1, addresses common image deficiencies, a subsequent semi-automated refinement stage is crucial for tackling more nuanced and complex visual challenges. These challenges include, but are not limited to, partial occlusions by rover hardware elements (e.g., wheels, antennas, or calibration targets), subtle lens-induced distortions not captured by generic filters.

This refinement phase incorporates a rigorous manual verification protocol for the image set that has passed the initial automated screening. The efficiency and accuracy of this manual review are substantially enhanced by leveraging the capabilities of Grounded-SAM, a sophisticated vision-language segmentation model. Grounded-SAM’s strength lies in its ability to perform open-vocabulary segmentation, identifying and delineating image regions based on arbitrary textual prompts. This is particularly advantageous for our application, as it allows for the flexible identification of diverse and potentially unforeseen rover components or artifacts without requiring model retraining or predefined class lists.

51 The operational workflow is as follows:

- 52 1. **Prompt Formulation:** Human domain experts, familiar with the rover’s morphology
53 and common imaging configurations, formulate targeted textual prompts. These prompts
54 typically reference known spacecraft components that have a high likelihood of intruding
55 into the image frame (e.g., "rover wheel", "robotic arm telemetry cable", "mast shadow on
56 terrain").
- 57 2. **Mask Generation:** Grounded-SAM processes each image in conjunction with these prompts
58 to generate segmentation masks. These masks highlight regions within the image that
59 correspond to the textual descriptions, effectively flagging areas suspected of containing
60 non-terrain elements or problematic features.
- 61 3. **Guided Manual Annotation:** The generated masks serve as precise visual guides for
62 human annotators. Instead of scrutinizing the entirety of each image for potential issues,
63 annotators can focus their attention on the regions highlighted by Grounded-SAM. This
64 significantly accelerates the review process and improves the consistency of identifying
65 obscured or compromised data.

66 Human annotators then perform the critical verification step. Based on the Grounded-SAM-proposed
67 masks and their own expert assessment, they make the final decision to:

- 68 • Confirm and accept the mask, leading to the flagging of the highlighted region for exclusion
69 from 3D reconstruction inputs.
- 70 • If the segmentation of Grounded-SAM is inaccurate or the image contains unrecognized
71 errors, manual annotation is carried out to obtain a clean image.
- 72 • Flag the entire image for exclusion if the problematic regions are too extensive or critical to
73 be simply masked out.

74 This process ensures that data compromised by non-Martian content or severe artifacts are meticu-
75 lously identified and appropriately handled.

76 The direct outcome of this Grounded-SAM assisted semi-automated refinement is a rigorously curated
77 collection of Martian stereo images. These images exhibit high visual integrity, characterized by
78 predominantly unobstructed Martian surfaces, more balanced illumination across the scene, and a
79 minimization of instrumental or environmental artifacts. Such a high-quality, clean dataset forms a
80 reliable and robust foundation essential for the subsequent stages of metric-aware 3D reconstruction
81 and, ultimately, for the training of generative simulation models like MarsGen.

82 A.3 Details of M3arsSynth construction

83 **The challenge of conversion from CAHVOR to pinhole model** This sections outlines the conver-
84 sion from a CAHVOR ($C_{CAHVOR}, A_{CAHVOR}, H_{CAHVOR}, V_{CAHVOR}, O_{CAHVOR}, R_{CAHVOR}$
85 vectors) camera model to a pinhole model, highlighting the critical parameters and potential imped-
86 iments. The conversion aims to derive pinhole model parameters (camera center $\mathbf{C}_{\text{pinhole}}$, rotation
87 matrix \mathbf{R} , intrinsic matrix \mathbf{K} , and radial distortion coefficients k_0, k_1, k_2) from the CAHVOR param-
88 eters.

- 89 • **Camera Center:** $\mathbf{C}_{\text{pinhole}} = \mathbf{C}_{CAHVOR}$.
- 90 • **Rotation Matrix \mathbf{R} :** Derived from normalized horizontal (\mathbf{H}_n) and vertical (\mathbf{V}_n) vectors,
91 and the optical axis (\mathbf{A}_{CAHVOR}).

$$\begin{aligned}\mathbf{H}_n &= (\mathbf{H}_{CAHVOR} - h_c \mathbf{A}_{CAHVOR}) / h_s \\ \mathbf{V}_n &= (\mathbf{V}_{CAHVOR} - v_c \mathbf{A}_{CAHVOR}) / v_s \\ \mathbf{R} &= \begin{pmatrix} \mathbf{H}_n^T \\ -\mathbf{V}_n^T \\ \mathbf{A}_{CAHVOR}^T \end{pmatrix}\end{aligned}$$

- **Intrinsic Matrix \mathbf{K} :** Determined by focal lengths ($f_u = h_s, f_v = v_s$) and principal point ($c_u = h_c, c_v = v_c$).

$$\mathbf{K} = \begin{pmatrix} h_s & 0 & h_c \\ 0 & v_s & v_c \\ 0 & 0 & 1 \end{pmatrix}$$

- **Radial Distortion k_0, k_1, k_2 :** Calculated from \mathbf{R}_{CAHVOR} , with k_1, k_2 also depending on h_s .

$$\begin{aligned} k_0 &= \mathbf{R}_{CAHVOR}[0] \\ k_1 &= \mathbf{R}_{CAHVOR}[1]/(\text{pixel_size} \times h_s)^2 \\ k_2 &= \mathbf{R}_{CAHVOR}[2]/(\text{pixel_size} \times h_s)^4 \end{aligned}$$

The conversion critically depends on four scalar parameters:

- h_s : Horizontal focal length scaling factor.
- v_s : Vertical focal length scaling factor.
- h_c : Horizontal principal point coordinate.
- v_c : Vertical principal point coordinate.

If these four parameters (h_s, v_s, h_c, v_c) are not available, the conversion to a pinhole model is not feasible. Therefore, these four scalar parameters are indispensable for a complete and accurate conversion from the CAHVOR to the pinhole model.

3D Gaussian Splatting for Photorealistic Scene Modeling We details key components and implementation specifics related to our 3D Gaussian Splatting (3DGS) model. The optimization of this model to accurately represent a 3D scene is driven by a combination of two primary loss functions: **Photometric Loss:** This loss ensures that the rendered images from the 3DGS representation closely match the input training images in terms of appearance. It penalizes differences in color and brightness, guiding the optimization towards visual fidelity.

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D\text{-SSIM}} \quad (1)$$

With the goal of guiding the model into plausible geometry, we introduced a geometric prior loss in the process of model optimization [], which is **Depth Regularization Loss**. This component utilizes the depth information output by the pre-trained large model and the depth obtained through the rasterization pipeline to supervise the geometric accuracy of the scene.

$$\mathcal{L}_{\text{depth}} = \|D_{\text{render}} - D_{\text{Metric Depth}}^*\|_1 \quad (2)$$

The bilateral grid is a key component in addressing photometric variations and appearance inconsistencies in Martian stereo imagery, which arise from factors like differing camera hardware, lighting conditions, or ISP pipeline transformations. This per-view learnable function transforms the rendered output of a 3DGS model to better match the target image’s appearance. Its implementation involves associating a 3D bilateral grid (a four-dimensional tensor $A \in \mathbb{R}^{W \times H \times D \times 12}$) with each training view, where W, H represent spatial locations, D represents pixel intensity values, and the final dimension stores parameters for a 3×4 affine color transformation matrix. For each rendered pixel, an affine transformation is retrieved via a differentiable slicing operation using trilinear interpolation based on its spatial location and a guidance intensity, allowing the grid’s parameters to be learned end-to-end. Integrated into the 3DGS optimization, the grid processes rendered images before the photometric loss calculation, encouraging it to bridge appearance gaps, and a Total Variation loss is applied as smoothness regularization to prevent overfitting and encourage the modeling of low-frequency changes. The grid’s resolution is typically much smaller than the input images (we have selected here is $16 \times 16 \times 8 \times 12$) to ensure computational efficiency and focus on low-frequency variations, with adaptive sizing possible based on scene characteristics. This joint training approach mitigates appearance inconsistencies, leading to more photorealistic and geometrically consistent 3D scene representations.

Camera Trajectory Synthesis and Motion Description A cornerstone of generating diverse and informative video sequences for the M3arsSynth dataset is the meticulous design and execution of virtual camera trajectories. We begin by defining a repertoire of canonical camera trajectory types.

These foundational trajectories, mathematically represented as a sequence of 6-Degrees-of-Freedom (6-DOF) poses $\mathcal{M}_{\text{traj}} = \{(R_t, T_t) \in \text{SE}(3) \mid t = 1, \dots, N\}$, where R_t signifies the camera’s rotation matrix and T_t denotes its translation vector at each discrete timestamp t , are engineered to encompass a wide spectrum of motion profiles. The spatial extent, or the overall scale and reach, of the predefined canonical trajectories is not fixed; instead, it is dynamically adjusted in response to the specific geometric characteristics of each individual 3D reconstructed Martian scene. This adaptation is primarily driven by the depth information derived from the reconstructed 3D model. Specifically, for regions within a scene that are identified as being in the near-field (characterized by relatively smaller depth values from the camera’s perspective), the corresponding segments of the canonical trajectories are programmatically contracted or scaled down. Conversely, for regions designated as far-field (characterized by significantly larger depth values, indicating distant terrain elements or horizons), the trajectory segments are expanded or scaled up. This depth-adaptive scaling strategy is paramount for ensuring that the synthesized video data effectively and consistently covers the scene’s content at appropriate levels of detail. Following the generation of these adaptively scaled trajectories, the precise 6-DOF pose parameters for each frame serve as the quantitative foundation from which natural language descriptions detailing the camera’s motion characteristics are subsequently derived, forming a key component of the textual modality within the M3arsSynth dataset.

Scene Content Captioning The M3arsSynth dataset incorporates rich textual descriptions to enable and enhance multimodal learning. While depth and normal maps are directly derived from the 3D reconstructed scenes, and camera motion characteristics are derived from the 6-DOF pose parameters of the trajectories, the acquisition of descriptive scene content captions involves a sophisticated process leveraging a Vision Language Model (VLM). We generate scene content captions by applying a VLM, referenced as ChatGPT-4o, to selected views from the synthesized video sequences. The generation process is guided by specific system prompts provided to the VLM. These prompts include the input image, the classification of the Martian terrain depicted (e.g., "Regolith/Rocky Terrain", "Dunes/Ripples (Sand/Dust)", as shown in Figure 4 of the paper), and a basic descriptive outline of the scene. By conditioning the VLM with this structured input, we obtain detailed and contextually relevant textual descriptions of the visual content. This methodology ensures that the textual modality is not only accurate but also aligned with the visual and geometric data, thereby creating a cohesive multimodal dataset suitable for training models like MarsGen for controllable video synthesis.

A.4 MarsGen Architecture and Training Specifics

Conditioning Mechanisms. The MarsGen model integrates multimodal information through distinct conditioning pathways. Textual prompts are initially concatenated with video tokens; this combined representation is then processed through a global attention mechanism to achieve feature fusion. Camera trajectory information is incorporated by first representing camera poses using Plücker embeddings, which are subsequently injected into the model via a ControlNet architecture. Finally, initial video frames are conditioned by concatenating them with the input noise distribution, which then undergoes a denoising process to guide the generation.

Fine-tuning Details. The fine-tuning of MarsGen was conducted with the following hyperparameters. The learning rate was set to 1×10^{-4} . We utilized the AdamW optimizer. A cosine learning rate scheduler was employed, incorporating a warm-up phase. The batch size was configured to 1 per GPU. Training was performed for 8,000 steps, with gradient accumulation implemented over 2 steps.

B Additional Experimental Results and Analysis

Qualitative Comparisons of Video Generation. The main paper presented quantitative comparisons of our generator against image-to-video and camera-controlled image-to-video models. This appendix provides additional quantitative results against other video generation models. Sora and ViewCrafter, for instance, evidently lack specialized modeling for dynamic Martian scenes, leading to uncontrollable video sequences inconsistent with the theme. This further validates the significance of our proposed dataset.

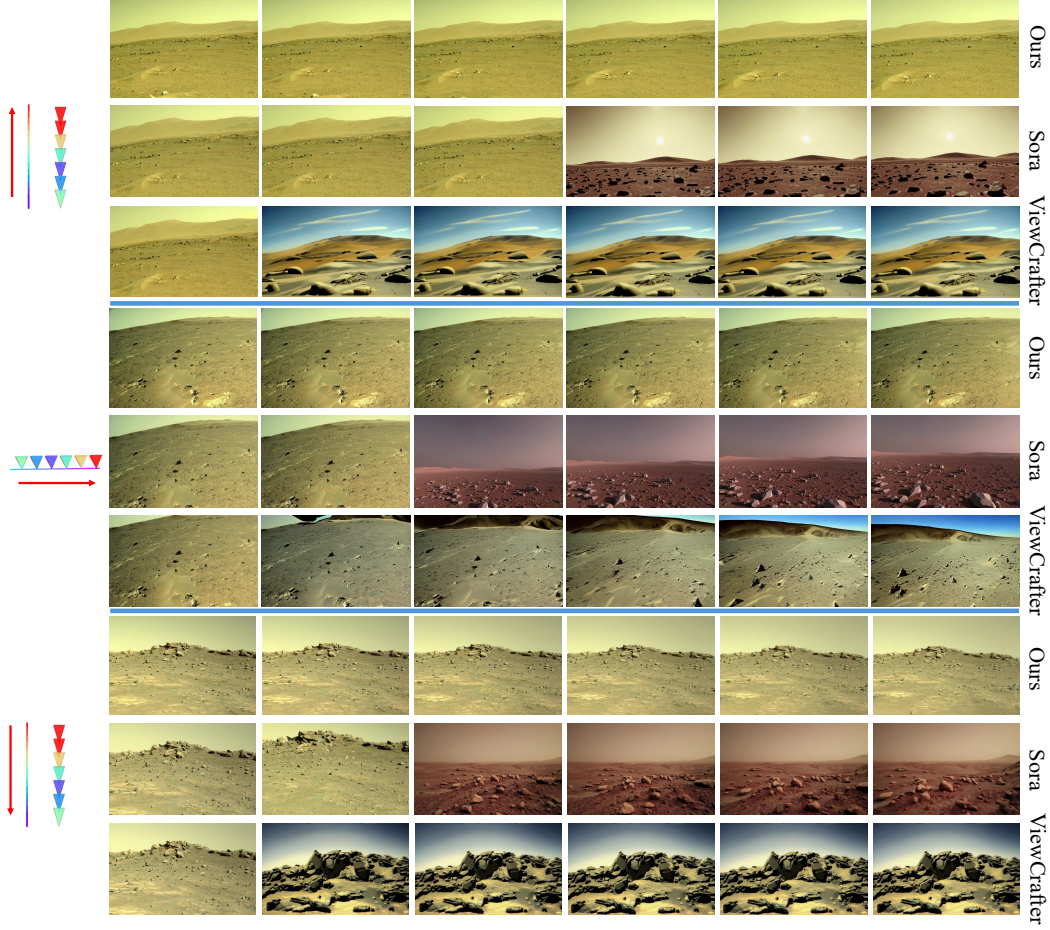


Figure A: **Qualitative comparison of video generation models on dynamic Martian scenes.** Each group of image sequences compares our model against Sora and ViewCrafter under a specific camera control condition. Our model demonstrates improved coherence with the intended camera control and greater thematic consistency for Martian landscapes, whereas Sora and ViewCrafter occasionally produce less controlled or thematically divergent outputs. The relevant comparison video files (in .mp4 format) are located in the corresponding subdirectories under the `comparison/` folder. For example, Sora’s examples are in the `comparison/sora/` directory,

181 For more comparisons of videos generated by our MarsGen model against the ground truth (GT),
 182 please refer to the `ours/` folder. Examples of other models’ failures in generating dynamic Martian
 183 scenes can be found in the `others/` folder.

184 **Qualitative Comparisons of 3D Reconstruction Pipelines.** Fig. B illustrates a qualitative comparison of 3D reconstruction outputs from different methodologies when applied to demanding Martian
 185 stereo image pairs. The top row of the figure presents results achieved by our M3arsSynth pipeline,
 186 which consistently demonstrates the ability to generate coherent and detailed 3D reconstructions
 187 of the Martian terrain. These outputs effectively capture the complex geometry and features of the
 188 landscape. These outputs effectively capture the complex geometry and features of the
 189 landscape.

190 In contrast, the second row of Fig. B displays reconstructions produced by the MAST3R pipeline.
 191 As indicated by the highlighted regions within the red boxes, MAST3R can encounter difficulties,
 192 leading to potential inaccuracies, loss of fine details, or artifacts. While MAST3R achieves 100% data
 193 utilization in quantitative assessments, it exhibits a significantly high reprojection error (46.98 px
 194 according to Table 2), which may stem from overfitting to unreliable depth priors.

195 Further highlighting the difficulties faced by existing techniques, traditional Structure-from-Motion
 196 (SfM) pipelines like COLMAP demonstrate extremely low utilization and robustness when pro-

197 cessing Martian data. For the specific visual examples shown in Fig. B, the COLMAP pipeline
 198 failed to generate a usable reconstruction in every instance, indicating its poor suitability for these
 199 challenging datasets. This qualitative observation is strongly supported by quantitative data presented
 200 in Table 2 of the main paper, which shows that COLMAP experiences failures on nearly 30% of
 201 preprocessed Martian image pairs. Such a high failure rate severely restricts its practical application
 202 for comprehensive 3D modeling of Martian environments.

203 In stark contrast, our proposed M3arsSynth pipeline achieves 100% data utilization and successfully
 204 reconstructs dense point clouds (averaging 250,000 points) while maintaining a competitive repro-
 205 jection accuracy (0.77 px, as detailed in Table 2). This robust performance, delivering both high
 206 data utilization and superior reconstruction quality, underscores the efficacy of our M3arsSynth data
 207 engine in producing the reliable and accurate 3D models that are essential for creating high-fidelity
 208 simulations and facilitating advanced video synthesis of Martian terrains.

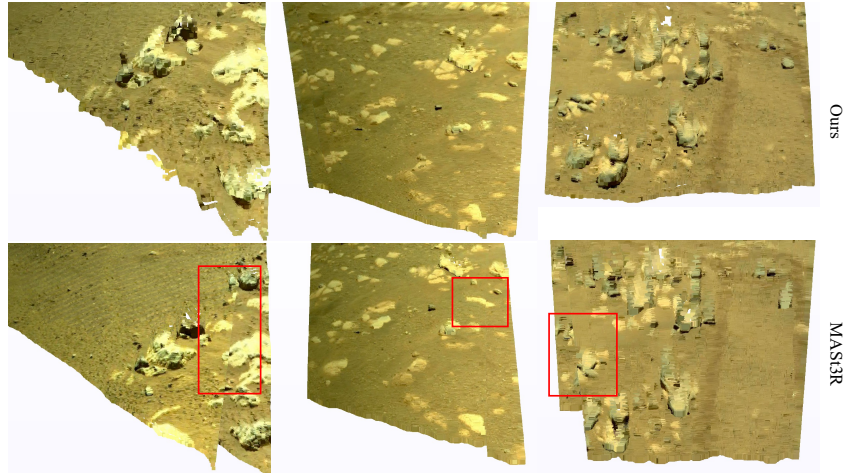


Figure B: **Qualitative comparison of 3D reconstruction pipelines on challenging Martian stereo imagery.** The top row displays reconstructions from our M3arsSynth pipeline, while the bottom row shows results from MASt3R, with red boxes highlighting areas of potential inaccuracy or detail loss. Notably, the COLMAP pipeline failed to produce reconstructions for all depicted examples, underscoring its limitations in robustness and data utilization on Martian datasets.