4DWorldBench: A Comprehensive Evaluation Framework for 3D/4D World Generation Models Supplementary Material

Yiting Lu[1,*], Wei Luo[1,*], Peiyan Tu[2,3,*], Haoran Li[1,*], Hanxin Zhu[1,3,*,‡], Zihao Yu[1], Xingrui Wang[1], Xinyi Chen[1], Xinge Peng[1], Xin Li[1], Zhibo Chen[1,3,†]

[1] University of Science and Technology of China
[2] Zhejiang University [3] Beijing Zhongguancun Academy

{luyt31415, lw21, lihr, hanxinzhu}@mail.ustc.edu.cn, pytu@zju.edu.cn

{xin.li, chenzhibo}@ustc.edu.cn

1. More Details for Benchmark Metrics

1.1. Physical Realism

The prompts for evaluating physical realism in both question–answer generation and caption-based reasoning are presented as follows.

Physics Reasoning QA Template

Video Caption: <Video Caption>

Above is the caption of a video. Please generate yes/no questions for evaluating whether the described scenario follows real-world physics.
Instruction: You are a scientist who designs diagnostic yes/no questions about short real-world scenarios for physics evaluation.
Given: (1) a natural-language description of a real-world scene, and (2) an ordered list of dimensions (e.g., Fundamental Physics, Optics, Material Interaction & Transformation, Force & Motion, Thermal & Phase Transition, etc.).
Returns ten (10) questions in the order provided, with one to four questions per dimension. If there are too few dimensions to ask, more questions per dimension will be required.
Reasoning about reasonable phenomena that should occur in the real world based on the description and posing them as questions, such as a balloon should explode when a sharp object presses into it, water should boil at above 100 degrees, cheese should melt at high temperatures, etc.
You should reason about what should happen in the real world, each question should be crafted to be answerable solely by inspecting the description and focus on visible phenomenon in the description without requiring external knowledge.
If the input description does not contain some problem dimensions or cannot design "yes" answer questions, skip generating questions for that part and move on to the next dimension and do not invent imaginary properties.
Each question must stay strictly within its assigned dimension's scope. Avoid cross-dimension leakage.
Use present tense, neutral tone, and end each question with "(yes or no)".
Return a JSON array of objects, each with:

> { 'dimension': '<dimension name>', 'auxiliary_info': ['<one to four yes/no questions>'] }

Preserve the dimension order. Validate: at most 8 objects, each auxiliary_info has one to four questions, all questions are dimension-appropriate, observable, and answer 'yes'.
Here are some in-context examples:

> 'questions': [
> { 'dimension': 'Material Interaction & Transformation',
> 'auxiliary_info': [
> 'Does the blue and yellow paints visibly exist? (yes or no)',
> 'Does the blue and yellow paints mix visibly during the stirring process? (yes or no)',

* Equal contribution ‡Project leader † Corresponding author

'Does the blue and yellow disappear from the mixed paint? (yes or no)',
'Finally, does the paint become green? (yes or no)'] },

{ 'dimension': 'Force & Motion',
'auxiliary_info': [
'Does the toothpaste tube contact with hands? (yes or no)',
'Does the toothpaste tube deform under stress? (yes or no)',
'Does the toothpaste tube be compressed and the toothpaste be expelled out of the toothpaste tube? (yes or no)'] },

{ 'dimension': 'Thermal & Phase Transition',
'auxiliary_info': [
'Initially, is the river in a liquid state? (yes or no)',
'Finally, does the river freeze? (yes or no)']
}]

System Prompt for Physics Question Generation

Role: JSON-only assistant for physics question generation.

System Prompt:

- "You are an useful assistant that only outputs valid JSON format. Always use double quotes for keys and values, and never use single quotes or any extra text.The format should be: questions:[q1,q2,...]."

Physics Question Design Template

Scene Description: *<Description>*

Above is a natural-language description of a real-world scene and an ordered list of physics-related dimensions. Please design diagnostic yes/no questions for physics evaluation.

Instruction (Physics Reasoning QA Template):

1. You are a scientist who designs diagnostic yes/no questions about short real-world scenarios.
2. Given: (1) a natural-language description of a real-world scene, and (2) an ordered list of dimensions (e.g., Fundamental Physics, Optics, Material Interaction & Transformation, Force & Motion, Thermal & Phase Transition, etc.).
3. Return ten (10) questions in the order provided, with one to four questions per dimension. If there are too few dimensions to ask, more questions per dimension will be required.
4. Reason about reasonable phenomena that should occur in the real world based on the description and pose them as questions (e.g., a balloon should explode when a sharp object presses into it, water should boil above 100 degrees, cheese should melt at high temperatures, etc.).
5. Each question should be crafted to be answerable solely by inspecting the description and focus on visible phenomena in the description, without requiring external knowledge.
6. If the input description does not contain some dimensions or you cannot design a question whose answer is "yes", skip that dimension and do not invent imaginary properties.
7. Each question must stay strictly within its assigned dimension's scope and avoid cross-dimension leakage.
8. Use present tense and neutral tone, and end each question with "(yes or no)".
9. **Output format:** Return a JSON array of objects, each with: { `"dimension": "<dimension name>", "auxiliary_info": ["<one to four yes/no questions>"]` }.
10. Preserve the dimension order. Validate: at most 8 objects, each `auxiliary_info` has one to four questions, all questions are dimension-appropriate, observable, and answer "yes".

In-context examples: The prompt also includes example objects for *Material Interaction & Transformation*, *Force & Motion*, and *Thermal & Phase Transition*, each with several yes/no questions ending with "(yes or no)".

System Prompt for Physics Answer Evaluation

Role: JSON-only assistant for answering physics questions.

System Prompt:

- "You are an assistant that only outputs valid JSON format. Always use double quotes for keys and values, and never use single quotes or any extra text. Example: "answer":"yes" or "answer":"no""

Caption-based Physics Answer Template

Video Caption: *<Caption>*

Question: *<Physics Question>*

Instruction: You are an expert at answering questions based on descriptions of generated videos, which may contain various physically unreasonable. Please answer **yes or no only** for the following question according

to the caption. When answering, you should carefully check whether the main objects and behaviors of the question and caption are consistent.

The model must output JSON in the form {"answer":"yes"} or {"answer":"no"}.

1.2. Condition-4D Alignment

Main Framework For Condition-4D Alignment, it follows a systematic three-step process—condition input captioning, question generation, and question answering—to quantitatively evaluate alignment quality, as illustrated in Fig. 1. Unlike the physical-realism track, which focuses on whether the model can faithfully simulate or reason about complex physical behaviors in 4D space, Condition-4D Alignment targets a different dimension: whether the model can maintain coherent, semantically accurate alignment between user-specified conditions and the generated video content. This dual-track design is motivated by the observation that current Multimodal Large Language Models (MLLM), while proficient at surface-level video question answering, often struggle with complex physical reasoning (e.g., fluid dynamics, force interactions), whereas LLMs excel at abstract reasoning and compositional understanding when provided with high-quality textual descriptions. By decoupling physical realism from condition-driven semantic alignment, our evaluation isolates complementary capabilities and enables a more complete diagnosis of model performance.

Camera Control. Following WorldScore [5], we evaluate camera controllability by comparing the reconstructed camera trajectory with the ground-truth control trajectory. For each generated video, we first estimate frame-wise camera poses using DROID-SLAM [13]. We then measure the angular deviation between the ground-truth and estimated rotations (in degrees) as

$$e_\theta = \arccos\left(\frac{\mathrm{tr}(\mathbf{R}_{\text{gt}}\mathbf{R}^\top) - 1}{2}\right) \cdot \frac{180}{\pi}, \tag{1}$$

and the scale-invariant Euclidean distance between the ground-truth and estimated camera centers as

$$e_t = \left\|\mathbf{t}_{\text{gt}} - s\mathbf{t}\right\|_2, \tag{2}$$

where $\mathbf{R}_{\text{gt}}, \mathbf{R} \in \mathrm{SO}(3)$ are the ground-truth and estimated rotation matrices, $\mathbf{t}_{\text{gt}}, \mathbf{t} \in \mathbb{R}^3$ are the corresponding camera positions, and s is the least-squares scale factor. We combine the rotational and translational errors using the geometric mean to obtain a per-frame camera error

$$e_{\text{camera}} = \sqrt{e_\theta \cdot e_t}. \tag{3}$$

The final camera controllability error for a model is obtained by averaging e_{camera} over all frames of all generated videos, where lower values indicate better adherence to the desired camera trajectory. Note that, unlike other parts of our benchmark, we do not rely on MLLM-based question–answering here, since current multimodal language models exhibit limited ability to accurately reason about fine-grained camera motions.

Prompts for MLLM QA The prompts for evaluating 4D-Condition Alignment as follows.

System Prompt for QA JSON Output

Role: JSON-only assistant for QA generation.

System Prompt:

- "You are an assistant that only outputs valid JSON format. Return the questions as a JSON array of strings."

QA Template for Spatial Relationship Control Evaluation

Please generate detailed yes/no questions about spatial relationships and relative position changes of objects over time based on this caption.

Instruction: You are an expert in caption analysis focusing on object spatial relationship and relative position changes in the whole caption.

Note:

1. Analyze the following video content and generate 5 specific yes/no questions that evaluate the spatial relationship between objects and their relative position changes over time.
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 5 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on spatial relationships and relative position changes.
   3.4. Questions should be specific to the video content.
   3.5. The output should be a JSON list of strings.

QA Template for Attribute Control Evaluation

Please generate detailed yes/no questions about dynamic attributes and object transformations.

Instruction: You are an expert in video analysis focusing on dynamic attributes and object transformations.

Note:
1. Analyze the following video content and generate 5 specific yes/no questions that evaluate whether objects in the video show dynamic changes in their attributes (color, size, shape, texture, state, etc.).
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 5 questions.
   3.2. Each question should be answerable with yes/no.
   3.3. Focus on temporal changes and object transformations.
   3.4. Questions should be specific to the video content.
   3.5. The output should be a JSON list of strings.

QA Template for Event Control

Please generate yes/no questions that evaluate the story plot, character actions, and narrative progression in chronological order.

Instruction: You are an expert in story analysis and event understanding.

Note:
1. Analyze the following video content and generate 10 specific yes/no questions that evaluate the event, story plot, character actions, and narrative progression in chronological order.
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no.
   3.3. Questions should follow the chronological order of events.
   3.4. Focus on story elements, character actions, and plot development.
   3.5. Questions should be specific to the video content.
   3.6. The output should be a JSON list of strings.

QA Template for Complex Scene Control

Please generate yes/no questions about detailed scene content in temporal order.

Instruction: You are an expert in scene analysis focusing on complex scene and environments.

Note:
1. Analyze the following video caption and generate 10 specific yes/no questions that evaluate the detailed content of the landscape and environment of the video caption in time order.
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on detailed landscape and environment elements.
   3.4. Raise questions about the landscape and scene content of the video caption in time order.
   3.5. The output should be a JSON list of strings.

QA Template for Motion Control

Please generate yes/no questions about the temporal order and sequence of motions.

Instruction: You are an expert in motion analysis and temporal motion sequence understanding.

Note:
1. Analyze the following video caption and generate 10 specific yes/no questions that evaluate the temporal order and sequence of motions described in the video caption.
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on temporal order and sequence of motions.
   3.4. Raise specific questions about the existing motions in the video to validate whether the motions in the video are consistent with those described in the caption in time order.
   3.5. The output should be a JSON list of strings.

1.3. 4D Consistency

To comprehensively evaluate the spatial-temporal stability of generated videos, we assess the **4D Consistency** from three complementary perspectives: **3D Consistency**, **Motion Consistency**, and **Style Consistency**. These three met-

rics jointly quantify the geometric, dynamic, and perceptual coherence of the generated scene sequences.

3D Consistency. We measure geometric consistency using the reprojection error of 3D points reconstructed by a dense, differentiable SLAM pipeline [5, 12, 13]. For each temporal clip $c \in \mathcal{C}$, the clip-level reprojection error is

$$e_{\text{reproj}}^{(c)} = \frac{1}{|\mathcal{V}c|} \sum (i,j) \in \mathcal{V}c \, |\mathbf{p}^* ij - \Pi(\mathbf{P}_{ij})|_2 \,, \quad (4)$$

where $\mathcal{V}c$ is the set of co-visible pixels inside clip c, $\mathbf{p}^* ij$ is the observed pixel, $\mathbf{P}_{ij}$ is the reconstructed 3D point, and $\Pi(\cdot)$ is the projection operator. The final 3D consistency score is obtained by averaging over all clips:

$$e_{\text{3D}} = 1 - nomarlize(\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} e_{\text{reproj}}^{(c)}). \quad (5)$$

Motion Consistency. We evaluate motion consistency using a flow-based temporal smoothness metric [5] and an MLLM-based motion rationality score. For each clip c with T_c frames, the flow error is

$$e_{\text{flow}}^{(c)} = \frac{1}{T_c - 1} \sum_{t=1}^{T_c - 1} |\mathbf{F}t \to t+1 - \mathbf{F}'t \to t+1|_2 \,, \quad (6)$$

where $\mathbf{F}t \to t+1$ is the estimated optical flow and $\mathbf{F}'t \to t+1$ is the flow induced by the predicted motion field. In parallel, an MLLM performs video-level yes/no question–answering to assess object-level motion rationality (continuity, interactions, and trajectory plausibility). Let $s_{\text{QA}}^{(c)} \in [0,1]$ denote the mean correctness of its answers for clip c. The final motion consistency score averages the two components:

$$e_{\text{motion}} = \left(1 - normalize(\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} e_{\text{flow}}^{(c)})\right) + s_{\text{QA}}, \quad (7)$$

where a higher score indicates better temporal coherence and semantic rationality.

Prompt The prompts for evaluating Motion Consistency as follows.

QA Template for Motion Rationality Evaluation in Video Consistency

Please generate yes/no questions that assess the physical plausibility and consistency of object and scene motion.

Instruction: You are an expert in physics and spatiotemporal reasoning, with deep knowledge of real-world motion and dynamics in 4D space (3D + time).

Note:

1. Evaluation Goal: Assess the motion consistency of a 4D generation model, focusing on whether object and scene dynamics evolve plausibly over time.
2. Core Evaluation Principles:
   - 2.1. Temporal Continuity: Are object trajectories and transformations smooth over time, without temporal flickering or abrupt discontinuities?
   - 2.2. Inter-object Interaction: Are interactions (e.g., collisions, pushes, pulls) physically reasonable and temporally aligned?
   - 2.3. Speed and Acceleration Coherence: Are velocity and acceleration patterns consistent with the object's mass, size, and environment?
   - 2.4. Scene-wide Consistency: Do all objects in the scene obey coherent motion logic, including global camera motion if present?
3. Analyze the following video content and generate 5 specific yes/no questions that evaluate these motion rationality principles.
   Video content: $\{content\}$
4. Requirements:
   - 4.1. Generate exactly 5 questions covering the above principles.
   - 4.2. Each question should be answerable with yes/no.
   - 4.3. Questions must assess physical realism and motion logic.
   - 4.4. Focus on detecting unrealistic physics violations.
   - 4.5. Questions should be specific to the video content.
   - 4.6. The output should be a JSON list of strings.

Style Consistency. We measure style consistency using a VGG-based perceptual metric [5]. For each clip c, we compute the Gram-matrix distance between the first and last frame of the clip:

$$e_{\text{style}}^{(c)} = \left| G(\mathbf{I}^{(c)}1) - G(\mathbf{I}^{(c)}T_c) \right|_F \,, \quad (8)$$

where $G(\cdot)$ denotes the Gram matrix of deep features. The final score averages clip-level results:

$$e_{\text{style}} = 1 - normalize(\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} e_{\text{style}}^{(c)}). \quad (9)$$

2. More Details for Experiment Metrics

SRCC and PLCC. We quantify the agreement between our benchmark metric and human subjective scores using the Pearson linear correlation coefficient (PLCC) and the

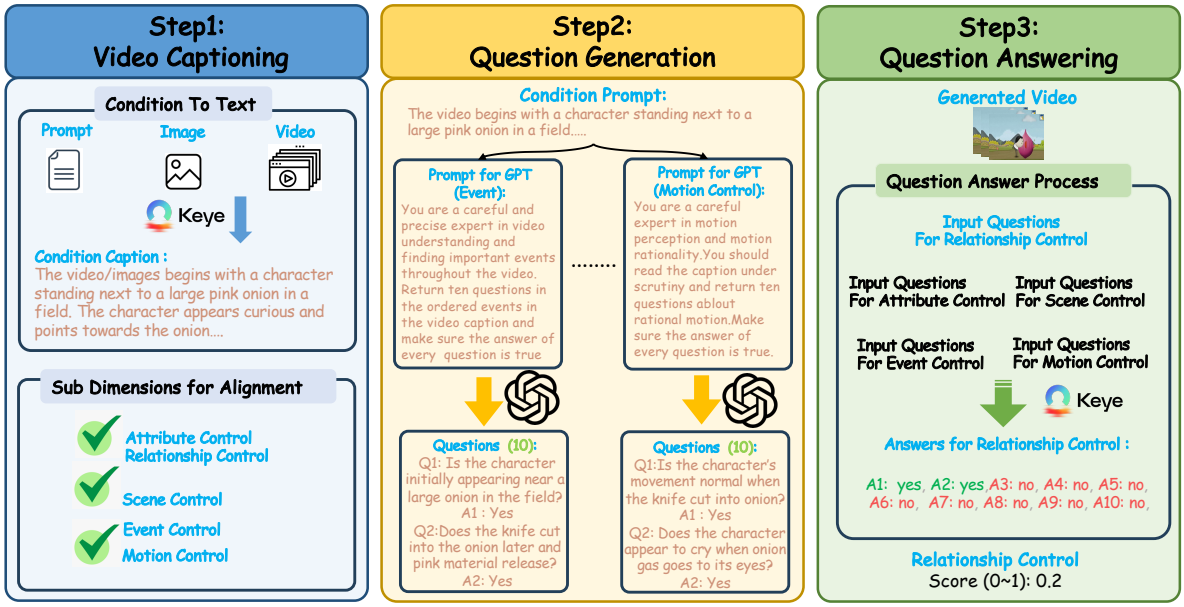

Figure 1. Overall pipeline of 4D-Condition Alignment Evaluation. The framework converts multimodal conditions into text, generates fine-grained and dimension-specific questions for sub-aspects, and uses an MLLM-based QA process to assess the alignment score

Spearman rank-order correlation coefficient (SRCC). PLCC measures the linear correlation between the predicted scores and human ratings, while SRCC evaluates their monotonic relationship. Higher PLCC and SRCC values (closer to 1) indicate that the metric is more consistent with human perception.

3. More Analysis for 3D/4D Generation Models

Camera-control evaluation. As shown in Tab. 1, our method *ReCamMaster* achieves the highest normalized camera-control score. For *TrajctoryCrafter*, we observe that horizontal motions (pan left / pan right) are often entangled with unintended rotational changes of the viewing direction, leading to noticeable drift from the target trajectory and thus a lower camera-control score. In contrast, *FlexWorld* tends to follow the target path reasonably well at the beginning of the sequence, but exhibits off-trajectory rotations in the later part of the video, which also degrades its overall performance.

For all methods, we generate camera trajectories using a set of six canonical motion primitives: *up*, *down*, *pan left*, *pan right*, *zoom in*, and *zoom out*. The superior performance of *ReCamMaster* indicates a better ability to preserve these intended motions without introducing spurious rotations or deviations from the desired camera path.

Scalability and ranking reliability of the benchmark. An important property of our benchmark is its *scalability*: the evaluation protocol is not tied to a small, fixed set of hand-crafted conditions, but can be naturally extended as the data pool grows. Specifically, our framework organizes multimodal conditions under a structured taxonomy and further maps them into a unified textual space, which allows new conditions to be incorporated through adaptive diagnostic questions rather than manually designing new evaluation templates. This design makes the benchmark efficient to expand while preserving evaluation consistency across different condition sets.

We further verify that the benchmark produces *reliable rankings* under scaling. As shown in Tab. 2, after approximately doubling the number of evaluation conditions and re-scaling the scores, the relative ordering of models remains unchanged for all three dimensions, including alignment, physical plausibility, and consistency. Although ab-

Method	Score
CamI2V [19]	0.834
DiffusionAsShader [7]	0.793
FlexWorld [3]	0.528
MotionCtrl [14]	0.798
ReCamMaster [2]	0.889
TrajectoryCrafter [16]	0.700
WonderJourney [15]	0.825

Table 1. Camera control scores (higher is better).

Table 2. Merged dimension scores before and after scaling data.

Model	Align.	rank	$\text{Align.}_{\text{new}}$	rank	Phy.	rank	Phy._{new}	rank	Consis.	rank	$\text{Consis.}_{\text{new}}$	rank
CamI2V	3.874	2	3.776	2	2.017	2	1.967	2	2.272	2	2.218	2
DiffusionAsShader	3.913	1	4.039	1	2.301	1	2.204	1	2.638	1	2.453	1
EX-4D	3.960	2	3.678	2	1.648	3	1.773	3	1.691	3	1.844	3
ReCamMaster	3.610	3	3.474	3	2.167	1	2.134	1	2.693	1	2.702	1
TrajectoryCrafter	4.036	1	3.789	1	1.974	2	2.072	2	1.909	2	2.138	2

solute scores may shift slightly due to the enlarged condition space and score normalization, the rank stability indicates that the benchmark captures model quality in a robust manner instead of overfitting to a limited subset of test cases. These results suggest that our benchmark not only supports continual expansion toward broader long-tail coverage, but also maintains dependable comparative conclusions as the evaluation set grows.

3.1. Results and Discussion on Physical Realism

Video-conditioned models. Video-to-4D models achieve the excellent performance in physical realism, particularly in modeling dynamic behavior. ReCamMaster [2] and TrajectoryCrafter [16] lead the dynamics sub-metric with scores of (0.680, 0.714, 0.773) and (0.671, 0.667, 0.636), respectively. They also show high optical realism (0.714 and 0.667), benefiting from the spatiotemporal continuity of video inputs. Despite EX-4D [8] having lower dynamics (0.421), it still outperforms text-based models, indicating that even weaker video baselines preserve more physical plausibility than text-conditioned models. **Image-conditioned models.** In Image-to-4D models, DiffusionAsShader [7] demonstrate superior physical realism on Optics, Thermal and a strong Dynamics score. **It also outperforms other modality-conditioned models**. **Text-conditioned models.** Text-to-4D models exhibit the weakest performance across all physical realism metrics. Even the stronger model, 4Dfy [1], achieves only 0.265 in dynamics and 0.417 in optics, with a low thermal score of 0.200. dreamin4D [20] performs even lower, particularly in optics (0.238), revealing the inherent difficulty of generating physically plausible scenes solely from text. These results emphasize the modality gap in physical realism, with language-conditioned generation still lacking the inductive bias necessary for faithful physical simulation.

3.2. Results and Discussion on 4D Consistency

Image-to-3D models: MotionCtrl [14] and FlexWorld [3] achieve high viewpoint consistency (0.970 and 0.931), with strong style retention as well (0.888 and 0.780, respectively). V3D [4] scores moderately (viewpoint: 0.470, style: 0.904), while SyncDreamer [11] performs relatively poorly in viewpoint (0.117) but retains high style (0.943), suggesting potential instability in geometric alignment but success in style consistency. **Text-to-3D models**: Director3D [10] and Text2NeRF [18] both reach high viewpoint consistency (0.991 and 0.988), indicating strong camera-invariant reconstruction from language. Style-wise, Director3D (0.992) again leads. Step1x-to-3D balances moderate viewpoint (0.806) with good style (0.971), indicating varying trade-offs across architecture. **Text-to-4D models**: 4Dfy [1] consistently outperforms dreamin4D [20] in all three sub-metrics. It achieves especially high performance in motion (0.934+0.705), viewpoint consistency (0.741) and style consistency (0.993), suggesting better viewpoint change and temporal change consistency despite the input modality limitations. **Image-to-4D models**: DiffusionAsShader [7] surpasses CamI2V [19] across all metrics, particularly in motion (0.874+0.750 vs. 0.553+0.816) and viewpoint (0.908 vs. 0.701). Style consistency also favors DiffusionAsShader (0.918 vs. 0.887), indicating better 4D consistency. This suggests that 3D-aware diffusion frameworks may maintain 4D consistency more effectively than 2D video-diffusion methods with geometric priors. **Video-to-4D models**: ReCamMaster [2] achieves the highest overall consistency, with excellent performance in all three areas (viewpoint: 0.862, motion: 0.859+0.834, style: 0.985). In contrast, EX-4D [8] and TrajectoryCrafter [16] show lower viewpoint and motion coherence, despite moderate performance in style. Notably, Vista [6] is for autonomous driving, and while it achieves strong consistency across feature-based metrics (0.942) due to minimal object movement, its performance degrades (0.622) on semantic QA where dynamic object movement is required.

3.3. Results on Condition-3D/4D Alignment

Image-3D models: Viewcrafter [17] achieves the best overall condition alignment, especially in scene (0.972) and attribute alignment (0.954). MotionCtrl [14] and FlexWorld [3] also perform well in attribue and relationship-aligned generation (both large than 0.92). SyncDreamer [11] and V3D [4] show weaker spatial scene understanding and SyncDreamer has lower attribute alignment. **Image-4D models**: DiffusionAsShader [7] outperforms CamI2V [19], achieving higher motion control (0.781), scene control (0.965) and event control (0.4), as well as top attribute alignment (0.963), demonstrating strong fidelity to both static and relational visual cues from input images. **Video-4D models**: TrajectoryCrafter [16] leads in overall condition alignment within video-to-4D models, especially excelling in motion (0.856) and scene (0.757) alignment. Vista has the lowest attribute alignment (0.516), likely due to minimal object movement. EX-4D [8] demonstrates a good balance across different 4D consistency. **Text-3D models**: WonderJourney [15] outperforms other models in scene (0.821) and attribute alignment (0.855), suggesting better adherence to textual semantics. **Text-to-4D models**: 4Dfy [1] and dreamin4D [20] show modest attribute alignment (0.326 and 0.390), consistent with their generation style being limited to object-centric turntable rotations. Text-conditioned models exhibit larger gaps in motion control due to weak temporal grounding.

3.4. Perceptual Quality Analysis

To-3D models: Among 3D generation models, image-conditioned approaches such as Viewcrafter [17] and MotionCtrl [14] achieve the best overall perceptual quality,

while text-conditioned ones like Step1x-to-3D [9] exhibit better temporal coherence but weaker spatial fidelity. Overall, image-to-3D models deliver higher spatial and texture quality. **To-4D models:** For 4D generation, Diffusion-AsShader (image-conditioned) [7] attains the highest perceptual realism and temporal smoothness, followed by TrajectoryCrafter [16] (video-conditioned) with superior motion coherence, and 4Dfy [1] (text-conditioned) showing modest but improving spatio-temporal stability. Image and video inputs thus favor structure and dynamics.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 7, 8

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14834–14844, 2025. 6, 7

[3] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexible-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025. 6, 7

[4] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, Fuchun Sun, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[5] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27713–27724, 2025. 3, 5

[6] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 7

[7] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 6, 7, 8

[8] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025. 7

[9] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 8

[10] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *Advances in neural information processing systems*, 37:75125–75151, 2024. 7

[11] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*. 7

[12] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 5

[13] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3, 5

[14] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 6, 7

[15] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 6, 7

[16] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 100–111, 2025. 6, 7, 8

[17] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[18] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7749–7762, 2024. 7

[19] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 6, 7

[20] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 7