Appendix

A. More Details about PAI-Bench-G

A.1. Data Source

We curated data by integrating established open-source benchmarks with web-sourced content. The open-source component comprises EgoExo4D [33], HowTo100M [57], Physics-IQ [58], WISA-80K [78], Agibot [2], Bridge-Datav2 [76], and Open X-Embodiment [59]. Complementing these, we also collected raw video data from YouTube and stock footage platforms, including Pexels, Pixabay, Free-video, MixKit, and Free-stock Video.

A.2. Details of Quality Score Calculation

For Quality Score, we adopt the evaluation protocol from VBench and VBench++ [38, 39, 99]. The specific components are defined as follows:

Subject Consistency. Evaluates the stability of the main subject's identity across frames. We extract frame-level features using DINO [15], which offers robust identity-sensitive representations [67]. For a video with T frames, the score is computed as:

$$S_{\text{subject}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}(\langle d_1, d_t\rangle + \langle d_{t-1}, d_t\rangle), \quad (1)$$

where d_i represents the unit-normalized DINO feature of the i-th frame, and $\langle\cdot,\cdot\rangle$ denotes the dot product. This metric averages the similarity of each frame with both the initial frame and its immediate predecessor.

Background Consistency. Assesses the temporal stability of the background scene, independent of foreground motion. We utilize the CLIP image encoder [63] to extract frame-level features. The score is calculated as:

$$S_{\text{background}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}(\langle c_1, c_t\rangle + \langle c_{t-1}, c_t\rangle), \quad (2)$$

where c_i denotes the unit-normalized CLIP feature of the i-th frame. Similar to subject consistency, this metric aggregates similarities relative to the first and preceding frames.

Motion Smoothness. Quantifies the physical plausibility and temporal coherence of motion. Following standard motion priors, we assume real-world motion is locally linear or quadratic. We subsample the generated video $[f_0, f_1, \ldots, f_{2n}]$ by dropping odd-indexed frames, reconstruct them using a frame interpolation model [49] to obtain $[\hat{f}_1, \hat{f}_3, \ldots]$, and compute the Mean Absolute Error (MAE) between the original and reconstructed frames:

$$S_{\text{smoothness}} = \frac{1}{T/2}\sum_{t=1}^{T/2}\|f_{2t-1} - \hat{f}_{2t-1}\|_1. \quad (3)$$

The final score is normalized to $[0, 1]$:

$$S_{\text{smoothness-norm}} = 1 - \frac{S_{\text{smoothness}}}{255}, \quad (4)$$

where higher values indicate smoother motion dynamics.

Aesthetic Quality. Measures visual appeal, encompassing composition, color harmony, and artistic quality. We utilize the LAION aesthetic predictor [45] to score each frame on a scale of $[0, 10]$. These scores are linearly normalized to $[0, 1]$ and averaged across all frames to derive the video-level metric.

Imaging Quality. Evaluates low-level image fidelity, specifically distortions such as overexposure, noise, and blur. We employ the MUSIQ predictor [42] (trained on SPAQ [25]) to obtain frame-level scores in $[0, 100]$. The final score is the average of these normalized values.

Overall Consistency. Measures the semantic and stylistic alignment between the generated video and the textual prompt. We employ the video-text consistency score from ViCLIP [82], which directly assesses the correspondence between the video content and the input description.

Image-to-Video Subject. Ensures the subject in the generated video remains faithful to the input reference image. Using DINO [15] features, we compute the similarity between the input image (s_{img}) and video frames (s_t):

$$S_{\text{i2v\_subject}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}\left(\langle s_{\text{img}}, s_t\rangle + \langle s_{t-1}, s_t\rangle\right). \quad (5)$$

The final score is averaged across all image-conditioned samples.

Image-to-Video Background. Verifies consistency between the generated background and the input image environment. This is critical for inputs emphasizing scene layout. We extract features using DreamSim [31], which is sensitive to layout variations, and compute:

$$S_{\text{i2v\_bg}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}\left(\langle b_{\text{img}}, b_t\rangle + \langle b_{t-1}, b_t\rangle\right), \quad (6)$$

where b_{img} and b_t are the unit-normalized DreamSim features of the reference image and the t-th frame, respectively.

A.3. Details of Domain Score Calculation

The Domain Score quantifies the adherence of generated videos to domain-specific physical and semantic constraints. We employ Qwen3-VL-235B-A22B-Instruct [9] as an automated evaluator, querying the model with a curated set of verification questions derived from the ground truth. The score is computed as the accuracy of the model's binary responses against the expected answers. For inference, we

Table 7. Generation parameters for the evaluated video models.

Model	Version	#Frames	FPS	Width	Height
CogVideoX [95]	-	49	16	720	480
CogVideoX1.5 [95]	-	81	16	1360	768
Cosmos-Predict2-2B [4]	-	93	16	1280	704
Cosmos-Predict2-14B [4]	-	93	16	1280	704
Cosmos-Predict2.5-2B [4]	`base/post-trained`	93	16	1280	704
DynamicCrafter [91]	`DynamiCrafter_1024`	16	8	1024	576
HunyuanVideo-I2V [43]	-	129	24	1184	768
LTX-Video-2B [35]	-	121	30	1216	704
LTX-Video-13B [35]	-	121	30	1216	704
MAGI-1-4.5B [73]	-	120	24	720	720
MAGI-1-24B [73]	-	120	24	1280	720
Veo3 [21]	`veo-3.0-generate-001`	192	24	1280	720
Wan2.1-I2V-14B [77]	`Wan2.1-I2V-14B-720P`	81	16	1280	720
Wan2.2-TI2V-5B [77]	-	121	24	1248	704
Wan2.2-I2V-A14B [77]	-	81	16	1280	720

uniformly sample frames at 2 fps and utilize greedy decoding to ensure deterministic evaluation. The specific prompt template used for evaluation is illustrated in Figure 8.

PAI-Bench-G Domain Evaluator System Prompt

You are a helpful AI assistant that answers questions about videos. Answer with just YES or NO. I'll show you a video with several frames. Please look carefully at all frames to understand what's happening in the video, then answer the question about the video with either YES or NO.

Input
{Sampled Frames 0}
{Sampled Frames 1}
...
{Sampled Frames N}

Question:
Question: *{Question}*

Please answer with YES or NO and explain your reasoning.

Figure 8. **Prompt template utilized for automated domain evaluator.** The model receives uniformly sampled frames and a specific constraint question to verify physical compliance.

A.4. Details about Experiments

Model Configurations. To ensure reproducibility, we detail the inference specifications for all evaluated VGMs in Table 7. We report the specific checkpoint versions alongside their corresponding spatial resolutions ($W \times H$) and temporal settings (frame count and FPS) used during the generation process.

Human Evaluation Protocol. To validate our automated metrics against human preferences, we conducted a pairwise comparison study using a web-based interface, as shown in Fig. 20. Annotators were presented with a text prompt, a reference image, and two generated videos (labeled A and B). Evaluation focused on two distinct dimensions: (1) *Video Quality*, which assesses visual coherence, motion smoothness, and clarity (correlating with our Quality Score); and (2) *Physical Plausibility*, which evaluates adherence to physical laws and domain-specific realism (correlating with our Domain Score). For each dimension, participants selected one of four outcomes: *A Better*, *B Better*, *Both Good*, or *Both Bad*.

ELO Rating Formulation. We quantified model rankings using an ELO rating system initialized at 1000 with a K of 32. Pairwise comparisons are converted into numerical scores: a definitive preference is assigned 1.0 to the chosen model and 0.0 to the other, while ties expressed as *Both Good* or *Both Bad* are assigned 0.5 to each. We maintained independent ELO trackers for video quality and physical plausibility. For the Overall score, we computed a separate ELO rating by averaging the quality and plausibility preference scores for each comparison and using these averaged preference values as the inputs to the ELO update process. We present the detailed ELO rating results in Table 8.

A.5. Qualitative Visualization

We present qualitative visualizations of generated samples to illustrate model performance across the diverse domains of PAI-Bench-G. Representative examples corresponding to autonomous vehicles, robotics, industrial settings, common sense reasoning, human activity, and physical dynamics are provided in Figures 11, 12, 13, 14, 15, and 16, respectively.

[t] [1]

Given: Modalities $\mathcal{M} = \{\text{Blur}, \text{Edge}, \text{Depth}, \text{Seg}\}$ Extraction operators $\mathcal{E} = \{E_m\}$: E_{Blur}: blur kernel E_{Edge}: Canny edge detector E_{Depth}: VideoDepthAnything E_{Seg}: GroundingDINO+SAM2 Fidelity operators $\mathcal{F} = \{F_m\}$: F_{Blur}: SSIM (Blur SSIM) F_{Edge}: F1 score (Edge F1) F_{Depth}: si-RMSE (Depth si-RMSE)

Table 8. Human-study (ELO) vs. automated evaluation on the PAI-Bench-G benchmark.

Models	ELO Scores			Evaluation Scores		
	Overall	Domain	Quality	Overall	Domain	Quality
Source Videos	1225.0	1278.1	1176.0	82.6	87.1	78.0
Veo3 [21]	1112.2	1122.8	1102.9	82.1	86.7	77.6
DynamicCrafter [91]	734.9	749.9	718.9	69.7	65.6	73.7
HunyuanVideo-I2V [43]	964.3	946.6	981.1	77.4	76.8	78.0
CogVideoX1.5 [95]	907.9	896.4	918.1	78.3	80.1	76.6
MAGI-1-24B [73]	988.9	969.3	1007.7	78.5	80.5	76.5
Wan2.1-I2V-14B [77]	976.0	961.4	988.7	80.5	83.8	77.3
Cosmos-Predict2.5-2B [4]	1008.1	968.5	1047.8	81.4	84.9	78.0
Wan2.2-I2V-A14B [77]	1082.6	1106.9	1058.7	81.6	85.6	77.5

F_{Seg}: mIoU w/ IoU matching (Mask mIoU) Conditional video X, generated video $\hat{X}$; number of prompts K

Part I: Condition Alignment each modality $m \in \mathcal{M}$ $X_m \leftarrow E_m(X)$; $\hat{X}_m \leftarrow E_m(\hat{X})$ $s_m \leftarrow F_m(X_m, \hat{X}_m)$ **Alignment Score** $\leftarrow \{s_{\text{Blur}}, s_{\text{Edge}}, s_{\text{Depth}}, s_{\text{Seg}}\}$

Part II: Generation Diversity (LPIPS) each condition c Generate K videos $\{\hat{X}^{(1)}, \ldots, \hat{X}^{(K)}\}$ all $i < j$ $d_{ij} \leftarrow \text{LPIPS}(\hat{X}^{(i)}, \hat{X}^{(j)})$ $\text{Div}_c \leftarrow \frac{2}{K(K-1)} \sum_{i<j} d_{ij}$ **Diversity-LPIPS** $\leftarrow \frac{1}{|\mathcal{C}|} \sum_c \text{Div}_c$

Part III: Generation Quality Quality Score $\leftarrow \text{DOVER}(\hat{X})$

Alignment Score, **Diversity-LPIPS**, **Quality Score**

B. PAI-Bench-C Score Calculation Details

To comprehensively evaluate conditional VGMs, we propose a metric suite assessing three critical dimensions: 1) control fidelity (faithfulness to input signals), 2) visual quality, and 3) generative diversity under identical conditions. A unified formulation of the extraction and fidelity operators, alongside the complete scoring pipeline is in Algorithm A.5.

B.1. Control Fidelity

We quantify the alignment between generated videos (V_{gen}) and reference control signals (V_{ref}) by projecting both into shared representation spaces, specifically varying levels of abstraction including low-frequency structure, edges, depth geometry, and semantic segmentation.

Vis Alignment: We apply the same blurring operation to both V_{gen} and V_{ref}. We then compute the Structural Similarity Index Measure (SSIM) [84] between the blurred representations, averaging scores across the dataset.

Edge Consistency: We evaluate boundary alignment by extracting binary edge maps from V_{gen} and V_{ref} using the Canny edge detector. The alignment is measured via the classification F1 score [75] on the pixel-wise binary maps.

Geometric Fidelity: To measure 3D geometric consistency, we extract depth maps using DepthAnythingV2 [94]. We calculate the scale-invariant Root Mean Squared Error (si-RMSE) [23] between depth estimations of V_{gen} and V_{ref}.

Semantic Alignment: We assess semantic layout consistency using a segmentation pipeline powered by GroundingDINO [51] and SAM2 [65]. To mitigate redundancy from open-set detection, instance masks are aggregated by caption phrase to form class-level masks. We establish correspondence between V_{gen} and V_{ref} masks via an Intersection over Union (IoU) matching algorithm. Matches with an IoU < 0.1 are filtered out, and the final score is reported as the mean IoU (mIoU) over valid correspondences.

B.2. Visual Quality

We assess the aesthetic and technical fidelity of generated videos using DOVER-technical [87, 88, 90], a reference-free video quality assessment metric. We compute the mean score across the entire dataset. Higher values correspond to superior visual clarity and stability.

B.3. Generation Diversity

To quantify the diversity of model's outputs under identical conditioning, we employ the Learned Perceptual Image Patch Similarity (LPIPS) metric [98]. For a fixed condition, we generate a set of N videos corresponding to N distinct text prompts. We calculate the pairwise LPIPS distance between all $\frac{N(N-1)}{2}$ unique pairs within this set. The final *Diversity-LPIPS* score is derived by averaging these pairwise distances across all samples in the dataset. Higher values indicate greater diversity and reduced mode collapse.

B.4. Qualitative Visualization

We present example test case and its corresponding model generated results from each of the autonomous vehicel, robotics, and human on PAI-Bench-C, as shown in Figures 17, 18, and 19 respectively.

C. More Details about PAI-Bench-U

C.1. Detailed Data Curation Process

This section details the curation process for the embodied reasoning benchmark. To ensure rigorous evaluation, we

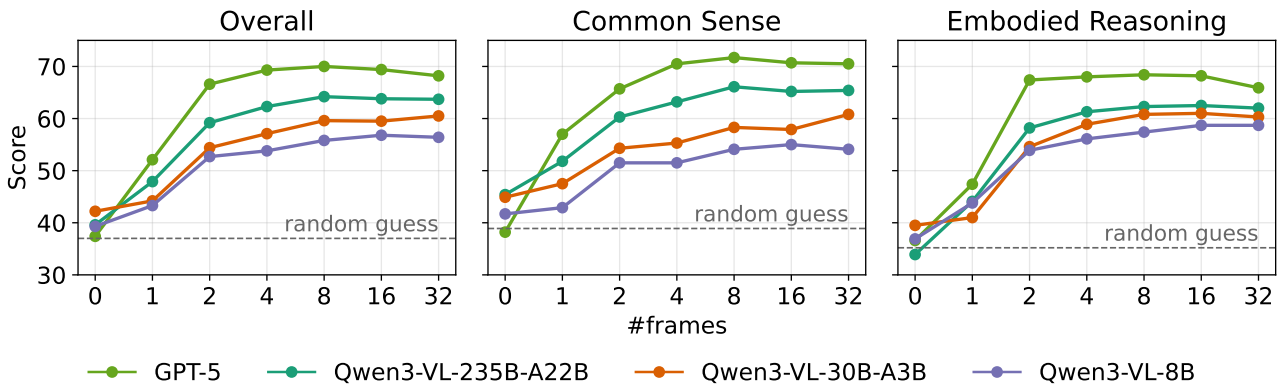


Figure 9. Overall, Common Sense, and Embodied Reasoning scores versus evenly spaced num_frames for Qwen3-VL-235B-A22B, Qwen3-VL-30B-A3B, Qwen3-VL-8B, and GPT5. Dashed horizontal lines denote random-guess baselines.

adhere to two core design principles: 1) **Unified Question Templates:** We standardize question formulation to ensure reasoning is grounded in visual input rather than exploited through textual biases. 2) **Unified Action Granularity:** To resolve ambiguity in next-action prediction (where multiple granularities, *e.g.*, "grab can" vs. "move hand left", may be valid), we adopt the hierarchical taxonomy proposed by [13]. We categorize behaviors into atomic-level *actions*, coarse-grained *subtasks*, and dataset-specific *goals*. The curation details for specific data sources are as follows:

RoboVQA. We curated 101 clips from the RoboVQA [68] validation split. We formulated binary multiple-choice questions (Yes/No) to evaluate two capabilities: *task-completion verification* (determining if an instruction was successfully executed) and *affordance* (assessing task feasibility given the visual context).

RoboFail. We manually annotated 100 examples from RoboFail [54] to assess reasoning under failure conditions. These samples evaluate action affordance and task completion verification in challenging scenarios characterized by: (1) complex temporal dynamics requiring observant perception, (2) physical constraints impeding execution, and (3) nuanced reasoning requirements beyond simple perception mismatches.

BridgeData V2. We processed the BridgeData V2 [76] validation split to extract 100 clips for multiple-choice QA. Conditioned on the task instruction and the visual history, the model is queried to predict the most plausible immediate next action.

AgiBot. We derived 100 multiple-choice samples from AgiBot [2]. For each clip, we provide the high-level task information and ask the model to identify the correct next subtask. Distractors are randomly sampled from the subtask sequence within the clip's full trajectory.

HoloAssist. We constructed 100 QA pairs from HoloAssist [81]. Providing the coarse-grained action annotation as the overall goal, we query the model for the most likely next subtask. Distractors are randomly sampled from other fine-grained action annotations associated with that coarse-grained goal.

AV (Proprietary). We curated 100 videos from a proprietary dataset to construct multiple-choice QA pairs. These videos exhibit diverse lateral and longitudinal behaviors with rich agent interactions. The questions are designed to: (1) predict the likely next immediate action of the ego vehicle, (2) verify the completion of a previously executed action, and (3) assess the affordance of specific actions within the given scenario.

C.2. Detailed Experiment Setup

System Prompt. The system prompt used for all MLLM inferences on PAI-Bench-U is provided in Figure 10.

PAI-Bench-U System Prompt

You are a helpful assistant. Answer in the format:
<think>reasoning</think>
<answer>answer</answer>.

Input
{Sampled Frames 0}
{Sampled Frames 1}
...
{Sampled Frames N}

Question: *{Question}*

Figure 10. **System prompt** used on PAI-Bench-U.

Experimental Model Settings Here, we describe the configurations used during inference. For all proprietary models, we adopt their default inference settings. For GPT-5, we refer to medium reasoning effort as enabling the think-

ing mode and minimal reasoning effort as running without the thinking mode. The corresponding model versions and checkpoints are summarized in Table 9. For all open-source models, we used vLLM as the inference backend and adopted the default inference parameters provided by each model.

Table 9. List of proprietary models evaluated in our experiments.

Model	Vendor	Version
GPT-4o	OpenAI	gpt-4o-2024-08-06
GPT-5	OpenAI	gpt-5-2025-08-07
Claude 3.5 Sonnet	Anthropic	claude-3-5-sonnet-20241022-v2

C.3. Discussion on Input Frame Count

To analyze how the number of sampled video frames affects model performance in PAI-Bench-U, we systematically evaluate GPT-5 [61], Qwen3-VL-235B-A22B [9], Qwen3-VL-30B-A3B [9], and Qwen3-VL-8B [9] with 0, 1, 2, 4, 8, 16, and 32 uniformly sampled frames.

Impact of Temporal Sampling Rate. As illustrated in Fig. 9, increasing the input frame count (< 8) initially yields consistent performance gains, as the model acquires the effective visual information necessary for question answering. However, performance saturates and remains stable beyond 8 frames. We attribute this plateau to two primary factors. First, we hypothesize a reasoning bottleneck: while the visual information provided by 8 frames is likely sufficient for the task, the models' failure to achieve further gains highlights intrinsic limitations in their spatial-temporal reasoning capabilities, rather than a deficiency in visual data. Second, excessive sampling may introduce temporal ambiguity. High frame rates significantly reduce the visual variance between adjacent frames, which can paradoxically hinder motion perception. For instance, when determining the movement direction of a robotic arm, sparse sampling (*e.g.*, start and end frames) presents distinct visual states that make the trajectory obvious. Conversely, dense sampling results in minute inter-frame differences. Current models often lack the sensitivity to process these fine-grained temporal shifts, leading them to misinterpret the high similarity between consecutive frames as a static scene, thereby degrading judgment.

C.4. User Study Details

To establish a human performance baseline for PAI-Bench-U, we conducted a user study where participants evaluated the benchmark questions based on the provided video clips. The aggregated participant accuracy serves as a reference for model evaluation. The annotation interface used for this study is depicted in Fig. 21.

D. PAI-Bench Leaderboard Visualization

To illustrate the performance of different models more clearly across the various dimensions of PAI-Bench, this section visualizes the PAI-Bench leaderboard using radar plots. As shown in Fig. 22, Veo3 and GPT-5 demonstrate strong performance across all domains. To enable a fair comparison across heterogeneous metrics, we apply appropriate normalization when constructing the radar charts.

E. Discussion and Future Work

In this section, we identify several open challenges and outline promising directions for future research:

Enhancing Metric Robustness with Advanced Encoders. The precision of overall consistency metrics is tied to the capabilities of the underlying video–text foundational models. Currently, models like VCLIP [85] exhibit constraints when processing lengthy and highly detailed textual prompts. Future work could leverage next-generation encoders to improve sensitivity to complex instructions, thereby enabling more granular consistency evaluations.

Addressing Conservative Generation Strategies. Our user study highlights a nuanced phenomenon in video generation: a trade-off between motion dynamism and generation safety. We observe that some models tend to adopt a conservative strategy that prioritizing static fidelity (*e.g.*, holding a racket) over complex, high-risk dynamics (*e.g.*, rapidly swinging it) to avoid artifacts. While this has a limited impact on quantitative rankings, developing methods that encourage risk-taking in motion generation without compromising visual quality remains a vital direction.

Necessity and Imperfection of MLLM Judges. While we utilize SOTA MLLMs like GPT-5 [61] to assess high-level semantics, automated evaluation remains an evolving field. Even the most advanced models possess inherent boundaries as visual judges, particularly when interpreting temporal dynamics in video. However, given the lack of effective alternatives for scalable semantic assessment, MLLM-based evaluation represents the current best practice.

The video begins with a view from inside a vehicle, likely captured by a dashboard camera, showing a wide, open road ahead under an overcast sky. The road is marked with white directional arrows indicating left or straight-right turns. Two lines of orange traffic cones are placed along the center of the road, suggesting some form of construction or roadwork. Within the area enclosed by the traffic cones, an arrow board displays flashing arrows pointing left and right to guide traffic around the detour. On either side of the road, there are grassy areas with small trees and shrubs. In the background, modern buildings, possibly part of an urban or suburban area, are visible, with a pedestrian bridge crossing above the road. The overall atmosphere is calm and quiet, with no other vehicles or pedestrians in sight.As the video progresses, the car changes lanes to the left and continues along the road. The camera angle shifts slightly to show the car moving further down the road, passing more traffic cones, and approaching the pedestrian bridge, which remains visible above the road.

(a) Input condition signals

(b) Source video

(c) Veo3

(d) CogVideoX1.5

(e) Cosmos-Predict2.5-2B

(f) DynamicCrafter

(g) HunyuanVideo-I2V

(h) MAGI-1-24B

(i) Wan2.2-I2V-A14B

Figure 11. **Example of autonomous vehicles domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

The video opens with a view of a testing environment, characterized by a large wooden table at the center. On this table, two robot arms are positioned at opposite ends, with the left arm closer to the camera and the right arm further away. Between the hands lies a dark wooden shelf with a red spherical object on its top rack, likely serving as a platform or obstacle. In the background, various pieces of equipment, including a tripod, a chair, are visible. [TRUNCATED] As the video progresses, the right robotic hand extends outward, moving from its initial position towards the red spherical object on the shelf. The hand then picks up the object and places it on the lowest rack of the shelf, completing a smooth, deliberate manipulation. The left robotic hand remains stationary throughout the sequence. No new objects appear in the video; all existing elements maintain their positions except for the movement of the right robotic hand. The scene concludes with the right robotic hand returning to its initial position, while the left hand continues to rest on the table. The overall environment remains unchanged, with the focus remaining on the interaction between the robotic hands and the wooden block, highlighting precise control during the demonstration.

(a) Input condition signals

(b) Source video

(c) Veo3

(d) CogVideoX1.5

(e) Cosmos-Predict2.5-2B

(f) DynamicCrafter

(g) HunyuanVideo-I2V

(h) MAGI-1-24B

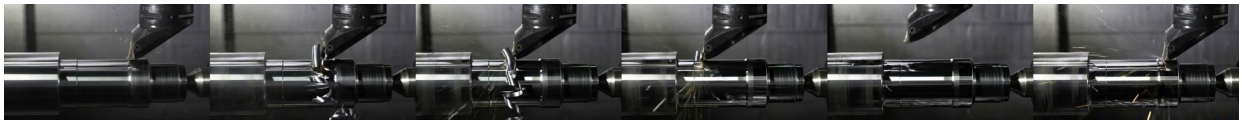
(i) Wan2.2-I2V-A14B

Figure 12. **Example of robotics domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

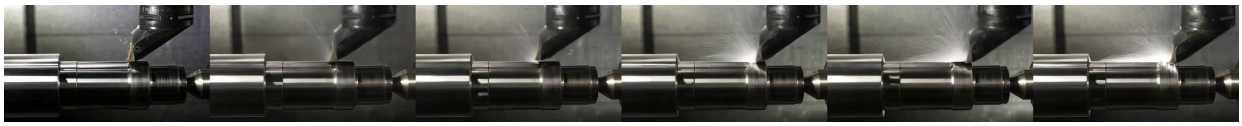
A close-up of a precision metalworking process in a controlled industrial setting. The first frame captures a cylindrical metal workpiece securely mounted on a lathe, rotating smoothly as a cutting machine, held by a black, angular fixture, approaches from above. The cutting machine, marked with numerical identifiers (5513 020-10), engages with the workpiece, shaving off thin metal shavings that are visibly ejected into the air, creating a fine mist around the machining area. [TRUNCATED] As the video progresses, the cutting machine continues its linear motion along the length of the workpiece, maintaining a steady pace. The tool's engagement with the material results in consistent metal shaving, producing a continuous stream of shavings that are dispersed into the surrounding space. The workpiece remains stationary relative to the camera's perspective, ensuring a clear view of the cutting metal process. The environment suggests a well-lit workshop, emphasizing the precision and efficiency of the operation. By the final frame, the cutting machine has almost completed its pass along the workpiece, leaving behind a smooth, polished surface. The metal shavings continue to be ejected, and the overall scene maintains a focused and industrious atmosphere, underscoring the meticulous nature of the metalworking process.

(a) Input condition signals

(b) Source video

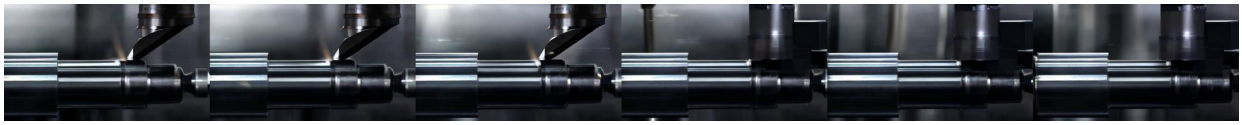


(c) Veo3

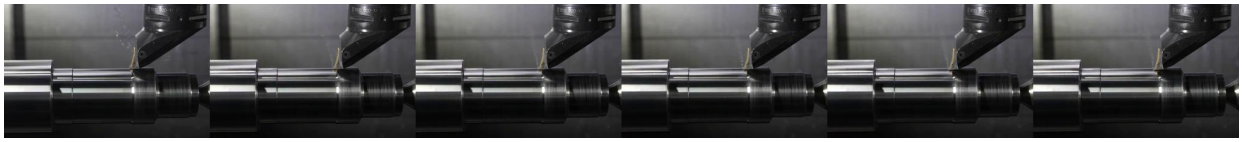


(d) CogVideoX1.5

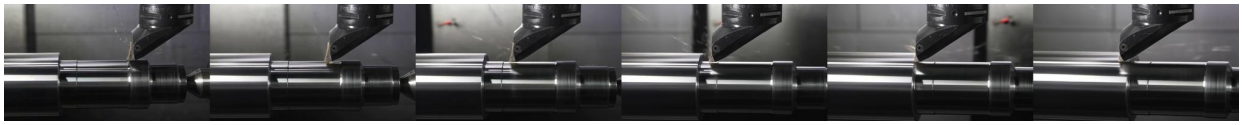
(e) Cosmos-Predict2.5-2B



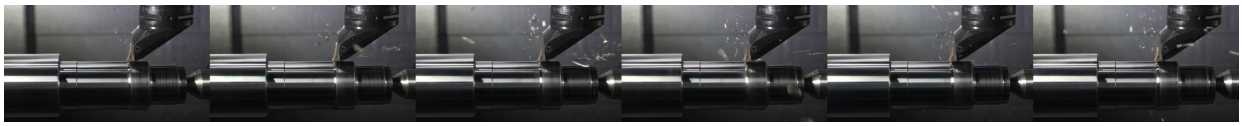
(f) DynamicCrafter



(g) HunyuanVideo-I2V



(h) MAGI-1-24B



(i) Wan2.2-I2V-A14B

Figure 13. **Example of industry domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

A close-up shot captures a wet, dark surface, likely a puddle or damp ground, under an overcast sky. The surface is covered with small ripples and bubbles, indicating recent rain. Several droplets of water form small, round bubbles that float on the surface, while others create concentric circles as they spread outwards. Tiny bits of debris, including leaves and twigs, are scattered across the surface, adding texture. As the video progresses, the rain continues to fall, creating more ripples and bubbles. Each raindrop impacts the water, causing a series of concentric circles that expand outward from the point of impact. The bubbles grow slightly larger as they float on the surface, reflecting the light and adding a shimmering effect. The scattered debris remains stationary, but the ripples and bubbles continuously shift and evolve, capturing the dynamic nature of the rainfall. The scene remains focused on the wet, reflective surface, with the ongoing rain creating a continuous pattern of ripples and bubbles.

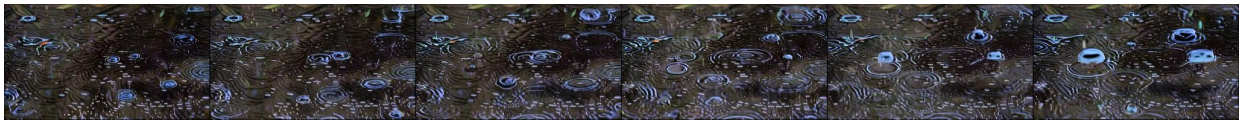
(a) Input condition signals

(b) Source video

(c) Veo3

(d) CogVideoX1.5

(e) Cosmos-Predict2.5-2B



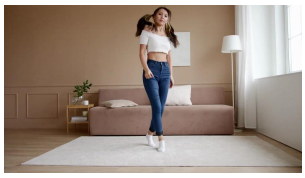
(f) DynamicCrafter

(g) HunyuanVideo-I2V

(h) MAGI-1-24B

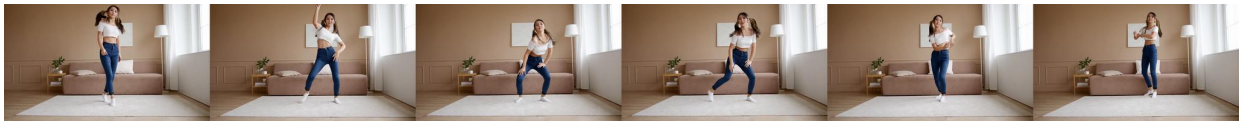
(i) Wan2.2-I2V-A14B

Figure 14. **Example of common sense domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

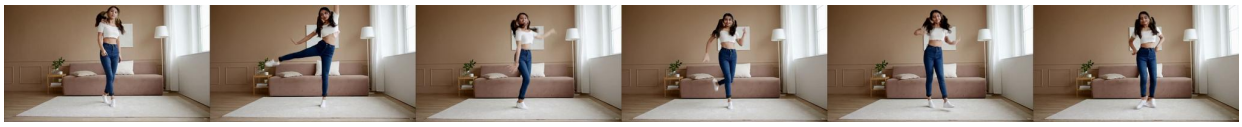


A modern, minimalist living room serves as the backdrop for a dynamic dance performance. A woman, positioned centrally on a light beige rug, executes a series of graceful dance moves. She wears a white cropped top, high-waisted blue jeans, and white sneakers, with her long hair tied up into two low pig tails that flow freely as she dances. [TRUNCATED] The video begins with the woman in a poised stance, her right leg crossed over her left, with both her arms to her sides. As she starts her dance routine, she lifts her right leg high into the air, balancing on her left foot, and extends her right arm gracefully above her head. She shifts her weight back down, landing smoothly and continuing with fluid, rhythmic movements that include hip sways and arm gestures. Her hair flows dynamically with her movements, adding a sense of energy and liveliness to the scene. By the final frame, the woman is caught mid-step in her dance routine, standing upright with her arms slightly bent in mid-air at the elbows. Her hair continues to flow, indicating the dynamic nature of her movements throughout the sequence. [TRUNCATED]

(a) Input condition signals



(b) Source video

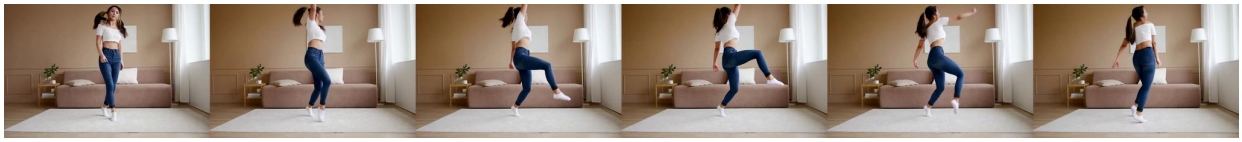


(c) Veo3

(d) CogVideoX1.5

(e) Cosmos-Predict2.5-2B

(f) DynamicCrafter



(g) HunyuanVideo-I2V

(h) MAGI-1-24B

(i) Wan2.2-I2V-A14B

Figure 15. **Example of human domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

A close-up shot captures a small, round object mounted on a metal pole in a sandy area. The background features a building with a red stripe across its facade and some trees to the right side. Tire tracks mark the ground, indicating recent vehicle activity. The object remains stationary until it suddenly ignites, producing a burst of bright light and smoke. The explosion is intense, sending sparks flying outward and creating a large cloud of white smoke that billows into the air. The smoke gradually disperses, revealing scattered debris around the pole. The scene remains static after the explosion, focusing on the remnants of the blast and the surrounding environment.

(a) Input condition signals

(b) Source video

(c) Veo3

(d) CogVideoX1.5

(e) Cosmos-Predict2.5-2B

(f) DynamicCrafter

(g) HunyuanVideo-I2V

(h) MAGI-1-24B

(i) Wan2.2-I2V-A14B

Figure 16. **Example of physics domain and model generations from PAI-Bench-G.** *Best viewed with zoom.*

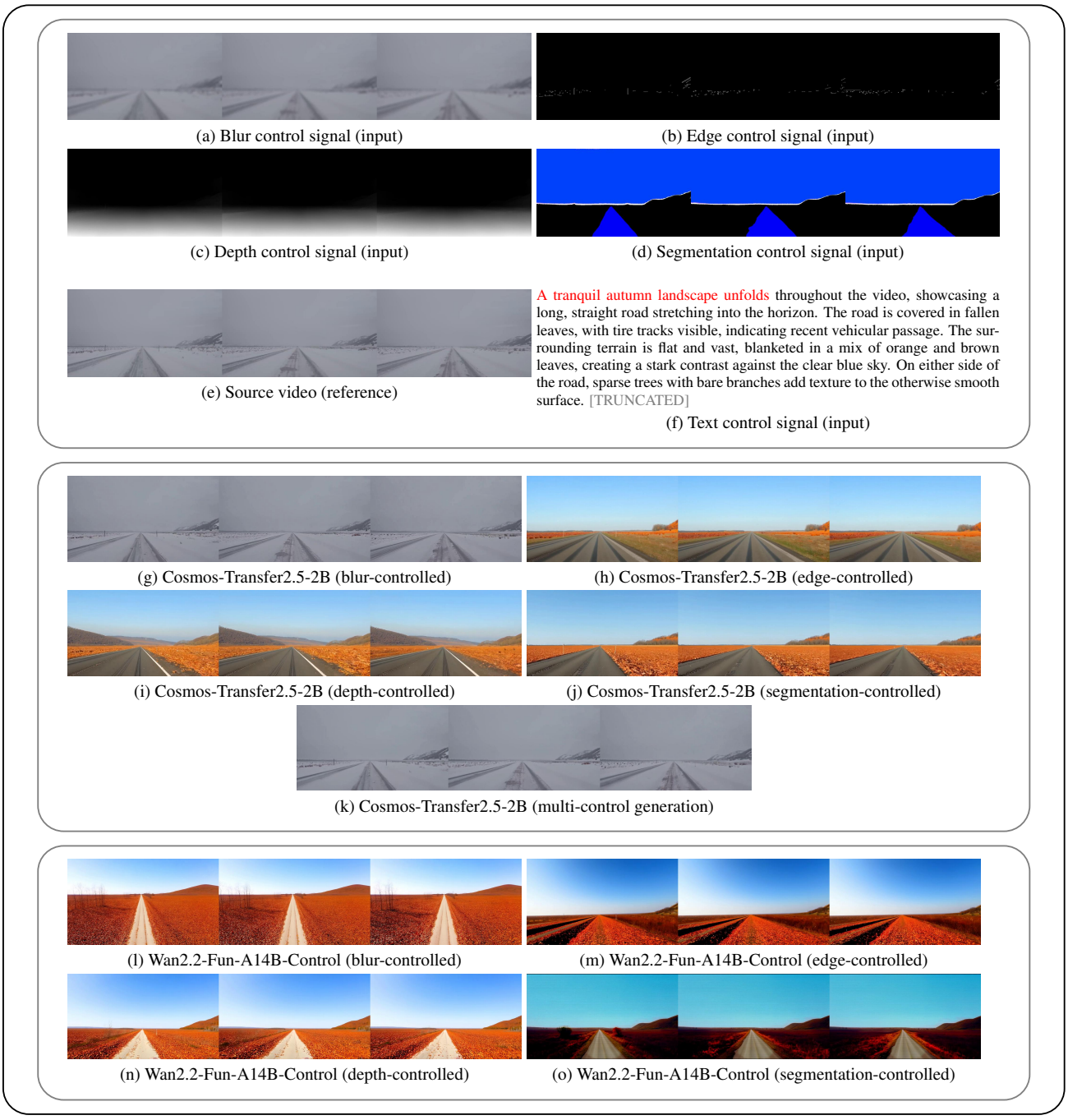
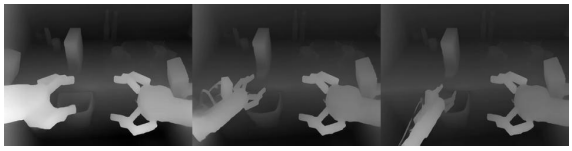


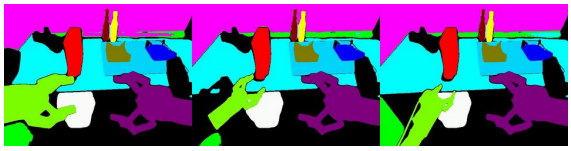
Figure 17. **Example of autonomous vehicle domain control signals and model generations from PAI-Bench-C.** *Best viewed with zoom.*

(a) Blur control signal (input)

(b) Edge control signal (input)



(c) Depth control signal (input)



(d) Segmentation control signal (input)

(e) Source video (reference)

In this video, robotic arms operate in a contemporary kitchen featuring a wooden countertop and a glass-top gas stove with two burners. Three ceramic jars are aligned on the counter. A black appliance, likely a water filter, is placed to the left. The wall has a textured stone appearance, adding to the kitchen's stylish design.

(f) Text control signal (input)

(g) Cosmos-Transfer2.5-2B (blur-controlled)

(h) Cosmos-Transfer2.5-2B (edge-controlled)

(i) Cosmos-Transfer2.5-2B (depth-controlled)

(j) Cosmos-Transfer2.5-2B (segmentation-controlled)

(k) Cosmos-Transfer2.5-2B (multi-control generation)

(l) Wan2.2-Fun-A14B-Control (blur-controlled)

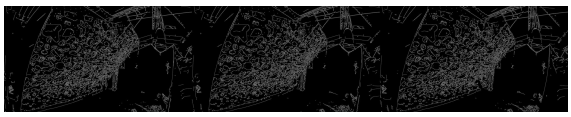
(m) Wan2.2-Fun-A14B-Control (edge-controlled)

(n) Wan2.2-Fun-A14B-Control (depth-controled)

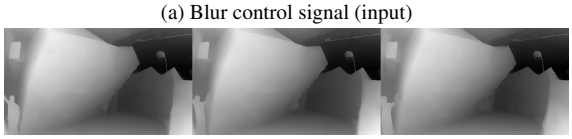
(o) Wan2.2-Fun-A14B-Control (segmentation-controlled)

Figure 18. **Example of robotics domain control signals and model generations from PAI-Bench-C.** *Best viewed with zoom.*

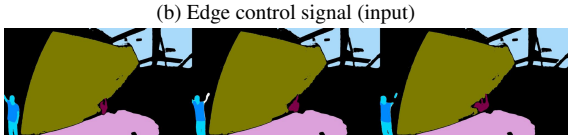
(a) Blur control signal (input)



(b) Edge control signal (input)



(c) Depth control signal (input)



(d) Segmentation control signal (input)

(e) Source video (reference)

The video is set in an indoor climbing gym, featuring a large, overhanging bouldering wall covered with a variety of neon-colored climbing holds. The holds are arranged in a complex pattern, with colors including electric blue, neon pink, lime green, bright yellow, and vivid orange, creating a visually dynamic backdrop. The wall is angled steeply, challenging the climber with its overhang. [TRUNCATED]

(f) Text control signal (input)

(g) Cosmos-Transfer2.5-2B (blur-controlled)

(h) Cosmos-Transfer2.5-2B (edge-controlled)

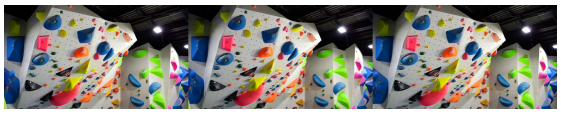
(i) Cosmos-Transfer2.5-2B (depth-controlled)

(j) Cosmos-Transfer2.5-2B (segmentation-controlled)

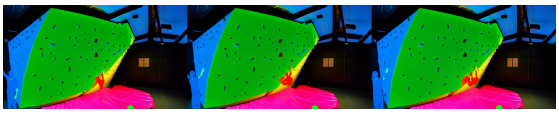
(k) Cosmos-Transfer2.5-2B (multi-control generation)

(l) Wan2.2-Fun-A14B-Control (blur-controlled)

(m) Wan2.2-Fun-A14B-Control (edge-controlled)



(n) Wan2.2-Fun-A14B-Control (depth-controlled)



(o) Wan2.2-Fun-A14B-Control (segmentation-controlled)

Figure 19. **Example of human domain control signals and model generations from PAI-Bench-C.** *Best viewed with zoom.*

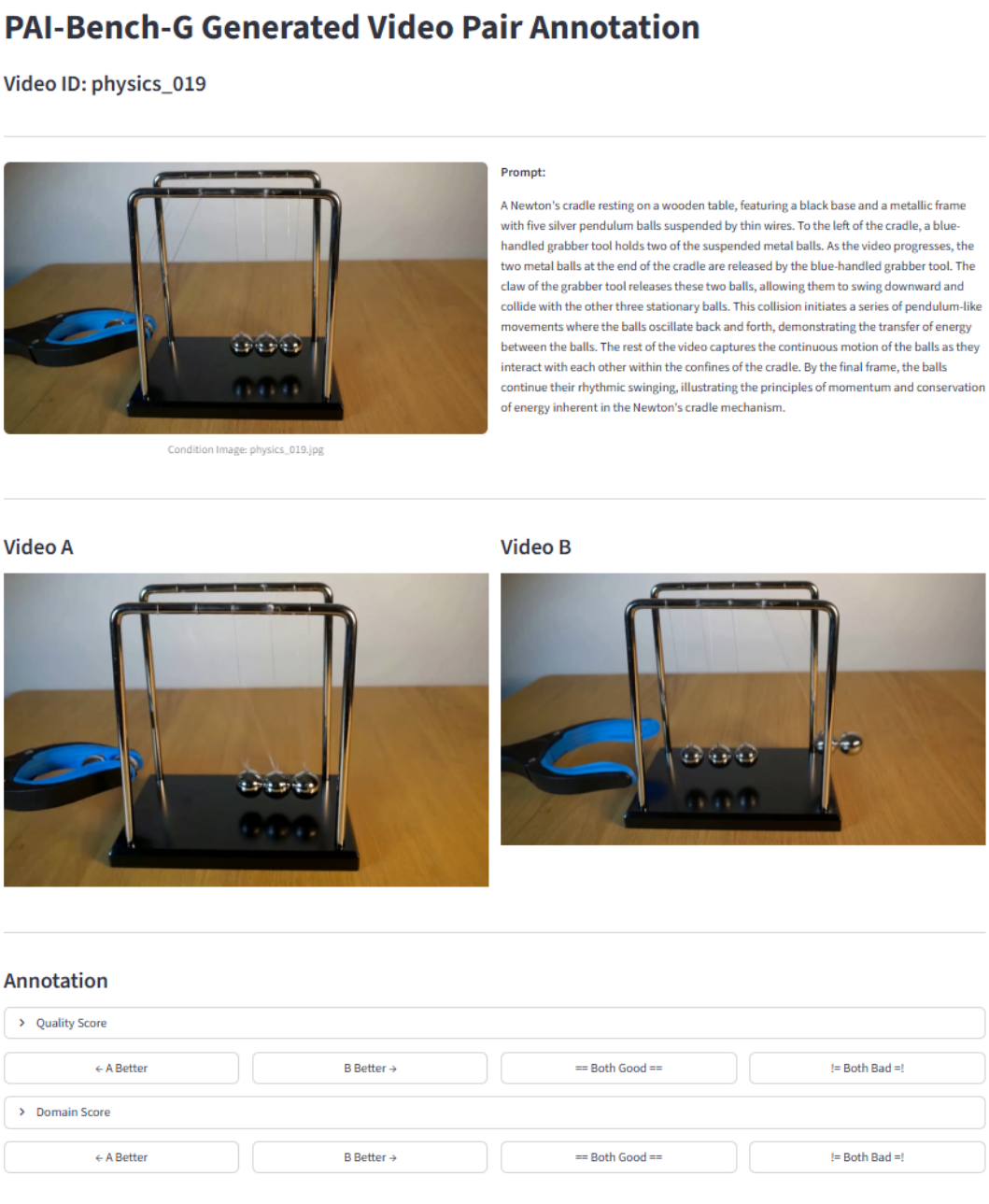


Figure 20. User study interface for preferences on PAI-Bench-G.

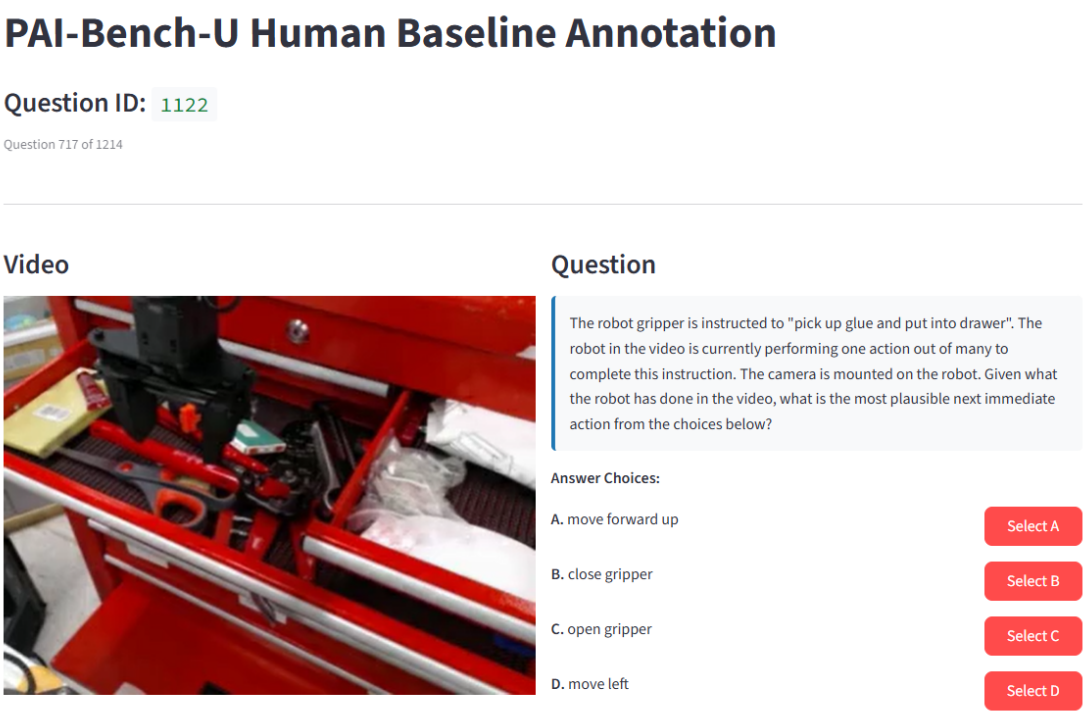


Figure 21. User study interface for establishing human baselines on PAI-Bench-U.

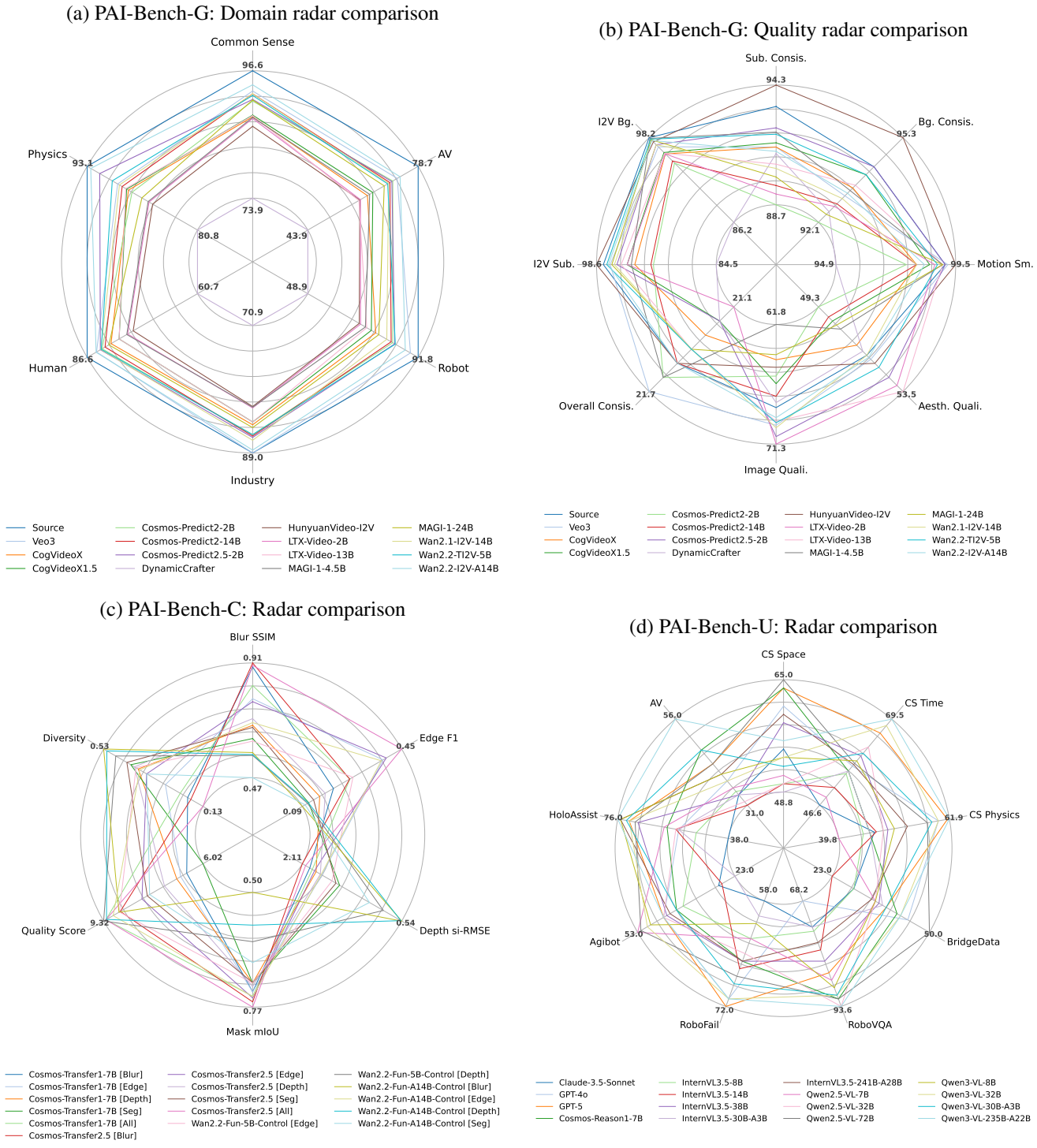


Figure 22. **Cross-suite radar comparisons in PAI-Bench.** Each subfigure highlights different capability dimensions across evaluation suites: