

Action Anticipation at a Glimpse: To What Extent Can Multimodal Cues Replace Video?

Supplementary Material

The supplementary materials are organized as follows: Appendix A presents an extended ablation study, including the effect of the action recognizer on action history performance, an analysis of visual and textual features, and fusion strategies for visual and multimodal inputs in AAG. This section also includes qualitative results on depth frames per dataset. Additionally, we provide an ablation on the effect of modalities using video inputs. Finally, we present qualitative results on the performance of VLMs and VQA models for generating action history in Appendix B.

A. Extended Ablation

A.1. Effect of the Performance of the Single-Frame Action Recognition Model

As outlined in the main paper, the action history performs best when it is generated using action predictions provided by an action recognition method. In all experiments, a single-frame action recognition model with the same architecture as the proposed AAG model was employed, effectively adapting AAG for action recognition. As shown in Table A, the performance of this module is generally lower than video-based approaches, including those using temporal aggregation with AAG, as well as state-of-the-art (SOTA) methods. However, it is worth noting that single-frame AAG outperforms SOTA methods on the IKEA-ASM dataset.

Given that the accuracy of this module can influence the overall performance of the proposed AAG method, as demonstrated by baseline results with ground-truth action history, we conducted an ablation study to evaluate how the performance of the action recognition model impacts the quality of past action history and, consequently, action anticipation accuracy.

Table B shows the performance of AAG across the three benchmarks. In each case, we examined the impact of using (a) the single-frame action recognition method, and (b) the best performing video-based action recognition model reported in Table A, to populate the past action history.

By enhancing the model’s ability to recognize actions, we observe a significant improvement in the overall performance of the proposed single-frame action anticipation method. This improvement can be attributed to the higher-quality action recognition model, which generates more accurate and reliable action history, directly influencing the performance of the action anticipation task. When past actions are correctly predicted, they provide better context for

Methods	IKEA-ASM	Meccano	Assembly101
Video (RGB-only)	57.58 / -	45.16 / 73.75	-
Video (Multimodal)	64.25 / -	49.66 / 73.75	43.60 / -
AAG _{AR} (RGB + AH)	66.43 / 92.56	31.21 / 66.03	34.19 / 58.30
AAG _{AR} (RGB + Depth + AH)	66.84 / 92.29	31.14 / 66.52	32.77 / 56.19
AAG _{AR-vid} (RGB + AH)	70.54 / 93.17	32.38 / 66.06	34.73 / 60.20
AAG _{AR-vid} (RGB + Depth + AH)	71.35 / 94.96	33.48 / 69.32	34.87 / 60.77

Table A. Action recognition performance report (Top-1/5) of single-frame action recognition with the same structure as AAG. Baselines include: IKEA-ASM uses I3D [8] (Top-5 is not provided); Meccano uses SlowFast [14]; Assembly101 uses TSM [31] and is trained on all fixed cameras.

AH Source	Modalities	IKEA-ASM		Meccano		Assembly101
		Top-1	Top-5	Top-1	Top-5	Recall@5
Single-Frame	AH*	41.42	79.03	25.72	51.97	9.80
	RGB + AH	43.82	83.11	26.22	52.18	8.25
	RGB + Depth + AH	<u>44.66</u>	<u>82.87</u>	26.43	<u>54.95</u>	8.90
Best Video AR	AH*	44.50	82.31	27.78	59.17	15.93
	RGB + AH	47.46	84.51	27.46	52.05	11.23
	RGB + Depth + AH	48.02	85.03	28.27	58.46	11.35

Table B. Single-frame anticipation results under single-frame and video-based action histories. Underline indicates best single-frame. * AH generated from RGB, Depth, AH action recognizer, while the other AH is generated from the same inputs as those used in the anticipation-based AAG.

anticipating the next action, improving accuracy and reducing ambiguity in the current scene, which ultimately enables the model to make more accurate predictions. While the observed improvements are consistent across the three benchmarks, this effect is particularly pronounced on the Meccano dataset, in which AAG previously underperformed when action history was based on predictions. The lower performance of the action recognizer is partly attributed to the challenges of temporal aggregation, particularly in the presence of rapid view changes in head-mounted videos. However, we argue that the multimodal approach as in [39], incorporating both object and hand information in the ego-centric perspective, could provide valuable insights for future work on single-frame action anticipation.

Finally, it is worth noting that the modality contribution per dataset is consistent with previous results, demonstrating that the action history improves the robustness of the results, while not disturbing the conclusions drawn from previous analysis.

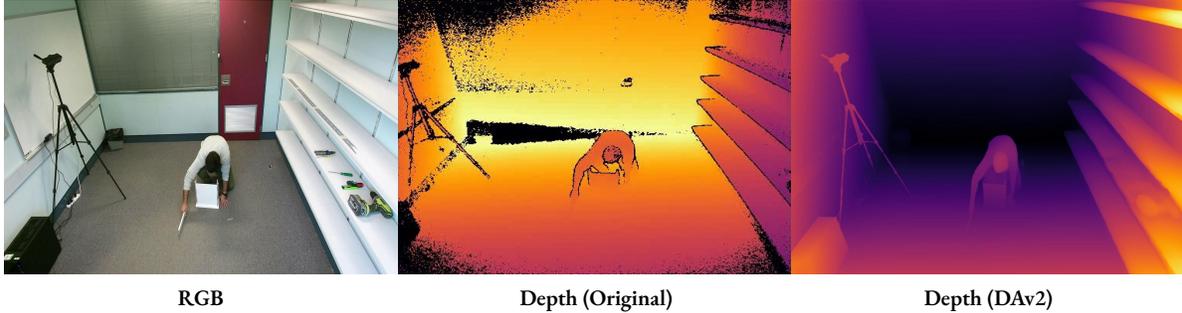


Figure A. Comparison of original depth frames and Depth Anything v2 estimates on the IKEA-ASM dataset. The higher noise in the original depth frames causes reduced performance when fused with RGB features.

A.2. Analysis on Visual Modalities

In the main paper, we show that depth information is more useful when extracted using Depth Anything V2 (DAv2) [55]. Similarly, we find that DINOv2 [36] combined with a cross-attention fusion mechanism is most effective for obtaining visual representations. This section ablates different strategies and encoders to validate these findings on the IKEA-ASM dataset.

Depth: is considered an auxiliary modality that contributes to a more robust visual representation when fused with RGB data. Choosing an appropriate fusion strategy is key to obtain optimal results. This statement is supported by the results in Table C. This table ablates different fusion strategies: concatenating and projecting the embeddings; sum of embeddings; soft-attention, defined as a softmax operation on depth embeddings followed by a dot product; self-attention using a transformer encoder; and cross-attention under the two possible settings, RGB or depth as query and the respective as key/value. Additionally, this table shows that depth information from ground truth, despite being more effective individually than the estimate, it shows diminished performance when fused with RGB data. As remarked in Section 3, we attribute this to the noisier nature of the depth ground-truth images, and shown in Figure A, where depth images compared to DAV2 estimates are noisier and may disrupt the fusion process, leading to suboptimal results.

Additionally, we can observe that our experiments revealed that among various fusion techniques for combining RGB and Depth features, cross-attention with $Q : \text{RGB}$ achieved the best performance. Our intuition is that this approach allows rich semantic and contextual cues from RGB data to guide the attention mechanism in selecting relevant information from the depth modality. Essentially, by leveraging RGB as queries, the model effectively prioritizes spatial and geometric features from the depth map that align with the observed scene, leading to more accurate and context-aware action anticipation.

Fusion Strategy	Depth Source	Top-1 / 5 Acc
Depth-Only		28.25 / 74.55
Concatenation		27.73 / 69.95
Sum		29.25 / 73.19
Soft-Attention	GT	27.69 / 74.79
Self-Attention		37.09 / 85.39
Cross-Att Q: Depth		26.65 / 71.91
Cross-Att Q: RGB		<u>37.29 / 85.23</u>
Depth-Only		26.33 / 68.95
Concatenation		35.53 / 82.43
Sum		33.85 / 78.07
Soft-Attention	DAv2	25.85 / 73.71
Self-Attention		38.34 / 85.23
Cross-Att Q: Depth		26.33 / 68.95
Cross-Att Q: RGB		38.82 / 86.19

Table C. Ablation on Visual Fusion Strategies and Depth Source.

Finally, we provide some qualitative results on depth frames across the selected benchmark datasets to support the statements in Section 5 regarding the contribution of depth across camera views. Figure B provides samples of depth frames from DAV2 compared to the original RGB image. While the third-person perspective on the IKEA-ASM dataset allows for clear distinction between background and foreground elements, the closer camera view in Meccano and Assembly101 limits the effectiveness of depth in these close-up perspectives. As a result, depth fails to offer valuable additional information and, in some cases, can even be misleading in these datasets, as shown in the experimental results. This further reinforces the varying utility of depth depending on the camera perspective, as discussed in Section 5.

Backbone Feature Extractor: The choice of an adequate feature extractors for RGB and depth features has a considerable impact on the performance of AAG. Table D ablates 3 different features extractors, (a) the 2D-CNN BN-Inception pre-trained on the TSN action recognition architecture [51], (b) ViT-L-14 [12] pre-trained on CLIP [38], and (c) DINOv2 [36]. From the three, the best encoding

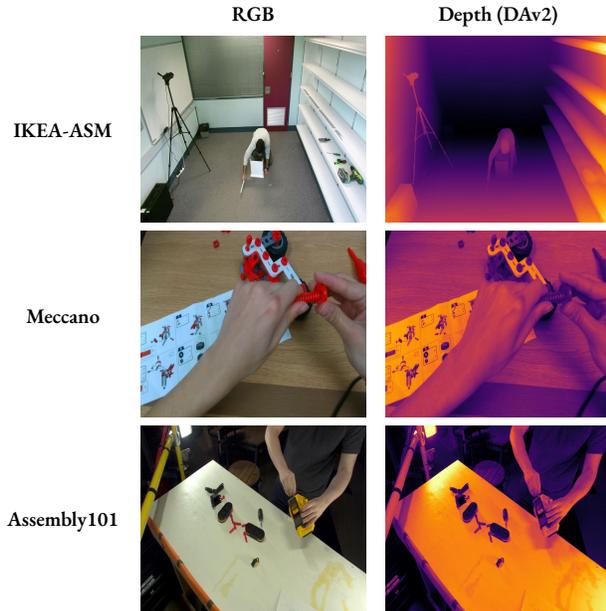


Figure B. Qualitative comparison of depth frames from DAV2 against the original RGB images across different benchmark datasets. The third-person perspective in IKEA-ASM clearly distinguishes background and foreground elements, enhancing depth’s utility. In contrast, the closer camera views in Meccano and Assembly101, where most objects and interactions occur within a nearly planar workspace, limit the effectiveness of depth.

performance on both RGB and depth modalities is given by DINOv2, whose self-supervised learning, enables us to derive richer and more generalized representations by removing the need for explicit labels during training.

Modality	Feature Extractor	Top-1 / 5 Acc
RGB	DINOv2	34.37 / 83.43
	ViT-L-14	26.33 / 69.07
	TSN BN Inception	26.33 / 68.95
Depth DAV2	DINOv2	26.33 / 68.95
	ViT	26.33 / 69.07

Table D. Ablation on Visual Feature Extractors.

A.3. Analysis on Multimodal Fusion

As for the visual fusion, we conduct an ablation on multimodal fusion strategies and text encoders under the best settings of AAG: RGB, depth and AH_{gt} , using a concatenation of the independent class name embeddings. Similarly, we use IKEA-ASM dataset. Table E depicts five different text encoders for AAG. Not only DistilBERT provides the best performance. We also evaluate CLIP’s text encoder on DINOv2 and CLIP’s ViT features. We find that the alignment between textual and visual embeddings for this task

yields the worst results, showcasing the ability of DINOv2 even when compared to visual and textual features in the same embedding space.

Text Encoder	Top-1 / 5 Acc
BERT [11]	60.42 / 90.20
RoBERTa [33]	56.78 / 88.68
DistilBERT [43]	61.83 / 89.64
CLIP Text Encoder [38]	39.66 / 84.83
CLIP Text + Visual Encoder [38]	27.89 / 68.95

Table E. Ablation on text encoder.

Table F ablates four different multimodal fusion strategies: concatenation of visual and textual features; sum of embeddings; a single self-attention transformer to fuse the three modalities, *i.e.*, not using cross-attention to fuse visual features first; and a self-attention transformer on visual and textual features, which is the one selected for AAG due to its superior performance.

Fusion Strategy	Top-1 / 5 Acc
Concat (Visual CA + Text)	56.06 / 87.23
Sum (Visual CA + Text)	51.22 / 86.43
Self-Attention (RGB + Depth + Text)	58.38 / 88.96
Self-Attention (Visual CA + Text)	61.83 / 89.64

Table F. Ablation on Visual-Text Fusion Strategies. CA = Cross-Attention.

A.4. Impact of Multimodal Cues Under Temporal Aggregation

We extend the analysis presented in Section 5 and particularly Table 1 to provide further insights on the effect of the selected modalities under temporal aggregation of video frames. First we compare the temporal aggregation mechanism from prior work [5, 52] with self-supervised video transformers. The temporal aggregation method consists of a transformer encoder with 3 layers and 8 attention heads per layer, fixed positional encoding to preserve temporal order, and a CLS token as a global representation. This strategy is selected as a simple baseline approach to temporal aggregation, having demonstrated superior performance compared to prior approaches, such as RNNs, in capturing temporal dependencies.

Method	Top-1	Top-5
Video Transformer Encoder	63.15	91.76
V-JEPA 2 [2]	58.58	90.96
TimeSformer [6]	57.22	90.24

Table G. Comparison of video feature extractors and temporal aggregation methods on AAG. Inputs: RGB, Depth, AH_{GT} .

Modalities	IKEA-ASM		Meccano		Assembly101	
	Top1 / 5 Acc	Recall@5	Top1 / 5 Acc	Recall@5	Top1 / 5	Recall@5
RGB	36.61 / 86.47	44.50	27.21 / 49.52	8.33	6.41 / 21.53	4.19
RGB, Depth	37.58 / 85.15	46.35	27.21 / 52.11	8.33	6.15 / 22.66	3.63
AH _{GT}	63.55 / 94.52	65.09	31.75 / 73.57	37.33	31.19 / 59.70	38.49
AH _{Pred} *	44.50 / 82.31	55.88	27.60 / 52.89	13.01	13.99 / 34.60	12.55
RGB, AH _{GT}	62.46 / 90.12	55.36	30.29 / 63.75	17.46	29.32 / 56.15	32.14
RGB, AH _{Pred}	41.06 / 82.07	45.94	26.43 / 55.52	13.25	14.06 / 31.77	10.08
RGB, Depth, AH _{GT}	63.15 / 91.76	60.97	32.88 / 67.90	25.95	30.27 / 56.51	33.05
RGB, Depth, AH _{Pred}	<u>46.02 / 83.43</u>	50.79	<u>27.24 / 54.52</u>	12.53	14.01 / 31.97	<u>11.24</u>

Table H. Analysis of the influence of modality selection on AAG performance with video aggregation performed on visual inputs. Underlined values indicate the best performance under realistic settings. * denotes that the action history was generated using the RGB, Depth, AH action recognizer. Otherwise, AH is generated from the same inputs as those used for anticipation.

By leveraging the same visual inputs as the single-frame method, we can evaluate the effect of temporal aggregation under the same training conditions. While the use of this training schedule may limit the performance of the video transformer, it is important to note that optimizing the video transformer is not the main goal of this work. Instead, the objective is to quantify its performance differences under identical conditions. On the other hand, self-supervised video transformers enable a one-to-one comparison between image-based and video-based self-supervised models, providing a clearer understanding of how temporal aggregation influences performance across modalities. Results in Table G show that self-supervised temporal aggregation fails to provide competitive performance, including both TimeSformer [6] and V-JEPA 2 [2]. We attribute this to the fact that aggregating self-supervised image features leverages robust individual frame representations, which are highly optimized for the task, while self-supervised video transformers may struggle to capture fine-grained temporal dependencies without fine-tuning on a specific task.

Table H presents an extended ablation of Table 1, incorporating video aggregation on the visual modalities. The results support our previous single-frame analysis. On visual inputs, we observe an improvement in IKEA-ASM when including depth, though this effect is diminished and only apparent in Top-5 accuracy for Meccano. As with the single-frame approach, depth is misleading when fused with RGB information in Assembly101. Action history, when populated from ground-truth data, continues to deliver the strongest performance on both IKEA-ASM and Assembly101. Similar to the single-frame approach, visual inputs yield misleading results compared to action history when it is generated from predictions of an equivalent action recognizer on Assembly101. For detailed accuracy results of each action recognizer, refer to Table A.

Given the results of state-of-the-art video-based methods on these datasets, we attribute the lower performance

observed here to the relatively simple transformer encoder used as a baseline. More refined temporal modeling, along with longer windows as seen in existing state-of-the-art approaches, is necessary to effectively capture interactions in datasets like Assembly101, which feature long-tail action distributions and high variability in the performed actions.

Notably, results on the Meccano dataset effectively leverage the three modalities. Despite the lower contribution of depth when combined with RGB alone, it still produces the strongest result in this setting. Additionally, the improvement in action recognition accuracy, along with the temporal aggregation of short-term observations, leads to the most effective results when all three modalities are fused, as highlighted in our single-frame analysis.

B. VLM Performance: Qualitative Evaluation

This section presents a qualitative analysis of vision-language models and extends the evaluation to the state-of-the-art visual question-answering (VQA) model, BLIP-2 [29]. Figure C provides qualitative comparisons of responses generated by Llama-3.2 Vision, GPT-4o, and BLIP-2 based on selected prompts.

BLIP-2 demonstrates a lack of reasoning beyond visual features, primarily focusing on human pose while disregarding contextual cues in the prompt. In contrast, VLMs exhibit a more human-like reasoning process, though they still present notable limitations, as acknowledged by the models themselves. For instance, Llama-3.2 (without context) explicitly states its difficulty in analyzing a single frame without additional information: “...However, I’m a large language model, I don’t have the capability to access the video or any additional context that may provide more information about the person’s previous actions. I can only make educated guesses based on the visual cues in the image”.

When contextual information is provided, VLM responses become more detailed and scenario-driven. Includ-

ing domain-specific knowledge, such as action class labels, leads to more concise and informative outputs, particularly in GPT-4o, which demonstrates stronger reasoning capabilities in structured environments.

The relevance of contextual information in the prompts is further demonstrated in Figure D. We include 2 prompts with and without dataset context. In this prompts we do not include a specific number of actions either, showcasing the importance of such information to get more structured responses. The main issues we may find in this descriptions is that due to the lack of context, Llama 3.2 focuses on a single item, a *dinosaur headpiece* and its responses is celebration-oriented (“...*suggesting they may be preparing for a party or a celebration...*”) rather than assembly-oriented. GPT 4o in contrast, while acknowledges the item is able to recognize the furniture pieces, though without specifying any particular previous task, which in this case, would be even hard for humans to do. By providing context, Llama 3.2 offers a more detailed description than GPT 4o, assuming that prior steps include preparation and inspection of the parts. This frame in particular corresponds to the beginning of a video sequence, so there are no prior actions annotated and the estimate provided by Llama is accurate enough. Notably, these prompts do not specify a fixed number of past actions, highlighting the role of structured input in obtaining more coherent responses compared to the ones in Figure C.

Overall, VLMs demonstrate strong general reasoning abilities but remain less effective for dataset-specific, fine-grained tasks. As a result, their generated descriptions can be misleading, whereas simple action labels provide more reliable and consistent guidance for AAG.



No context: What are the previous actions that lead to this frame?

BLIP-2: The person is doing a yoga pose.

Llama 3.2: Based on the image, it appears that the person has been assembling or disassembling some kind of furniture or equipment in the room. The scattered pieces of the desk suggest that they may have been taking it apart, and the person's posture and movements imply that they are in the process of reassembling it. However, I'm a large language model, I don't have the capability to access the video or any additional context that may provide more information about the person's previous actions. I can only make educated guesses based on the visual cues in the image.

GPT 4o: This frame captures a moment where a person is assembling furniture, likely a small table or shelving unit, in an office setting. The individual first identified the need for furniture assembly and gathered the necessary tools and components while clearing space in the office. They then unpacked the furniture parts, arranged the components systematically on the floor, and reviewed any provided instructions. In the initial assembly steps, the base panel was placed on the floor, and the person began attaching the vertical support pieces, with other parts like side panels nearby, ready for the next steps. At the current moment, the person is holding and aligning one of the vertical pieces with the base, possibly securing it in place by inserting screws or fitting pre-drilled holes, suggesting they are in the early-to-mid stages of the assembly, likely before securing all vertical supports and attaching the remaining panels.

Dataset context: What are the 5 previous actions that lead to this frame in the furniture building process?

BLIP-2: The person is doing a series of exercises to improve their balance and coordination.

Llama 3.2: The image depicts a person assembling a piece of furniture in a room with desks and chairs. The room is well-lit, and the person is surrounded by various components of the furniture, including a flat piece of wood, a long metal rod, and a smaller metal rod. The person is wearing a black t-shirt and red pants, and is kneeling on the floor, holding the flat piece of wood in place with one hand while using the other hand to attach the metal rods. To determine the previous actions that led up to this moment, we can analyze the scene and make some educated guesses: 1. Unpacking the components: The person likely started by unpacking the various components of the furniture from their boxes or bags. This would have involved carefully removing each item and placing it on the floor or a nearby surface. 2. Assembling the frame: With the components laid out, the person would have begun assembling the frame of the furniture. This would have involved attaching the metal rods to the flat piece of wood, using screws or other fasteners. 3. Attaching the legs: Once the frame was assembled, the person would have attached the legs to the bottom of the furniture. This would have involved screwing or bolting the legs to the frame. 4. Adding the tabletop: With the frame and legs in place, the person would have added the tabletop to the furniture. This would have involved attaching the tabletop to the frame using screws or other fasteners. 5. Final assembly: After adding the tabletop, the person would have performed any final assembly tasks, such as tightening screws or adjusting the furniture's position. These are the likely previous actions that led up to the moment depicted in the image.

GPT 4o: Based on the image, the person is assembling a piece of furniture, likely a small table or shelf. The five previous steps that led to this frame in the furniture-building process could be: unpacking the parts, where the person opened the packaging, took out all the necessary components, and laid them out on the floor for easy access; identifying the components, likely checking the instruction manual (if available) to understand which pieces go where, including screws, panels, and legs; placing the base panel, positioning the main wooden board flat on the floor as the foundation for the assembly; sorting the support beams, organizing the black support beams visible on the floor in preparation for attachment; and attaching the first beam, where the individual in the image appears to be securing the first leg or support beam to the base panel, possibly using screws or dowels.

Action Context: What are the 5 previous actions that lead to this frame in the furniture building process?

Possible actions: flip table, pick up back panel, pick up leg, push table top, spin leg, tighten leg, ...

Llama 3.2: Here are the previous actions that have led up to this moment: 1. Lay down side panel: The person has laid down the side panel of the furniture, ensuring it is properly aligned and secured. 2. Attach shelf to table: Before attaching the shelf to the table, the person would have ensured that the shelf is correctly positioned and aligned with the table's frame. 3. Flip table top: To prepare the table top for assembly, the person would have flipped it over to access the underside, where the necessary hardware and components are located. 4. Pick up side panel: The person has picked up the side panel, which was previously laid down, and is now holding it in place to ensure proper alignment and attachment. 5. Lay down front panel: After attaching the side panel, the person would have laid down the front panel, ensuring it is correctly aligned with the side panel and the rest of the furniture's frame. These actions have led up to the current moment, where the person is attaching the side panel to the table.

GPT 4o: Based on the image, here are five possible previous actions that may have led up to this moment: 1. Pick up table top. 2. Lay down table top. 3. Pick up legs. 4. Lay down legs. 5. Gather tools or hardware needed for assembly.

Figure C. Qualitative comparison of VLMs and VQA methods on 3 different prompts. Frame extracted from IKEA-ASM.



No context: What are the previous actions that lead to this frame?

Llama 3.2: The person in the image is wearing a dragon costume and standing in a living room, suggesting they may be preparing for a party or celebration. The presence of a table with a white tablecloth and a couch suggests that they may be hosting a gathering or event. The overall atmosphere of the image is playful and festive, with the person's costume and the decorations in the room contributing to a lively and joyful mood. It is likely that the person is preparing for a Halloween party or another type of celebration.

GPT 4o: The person in the image appears to be standing near a table with some materials laid out on it, possibly wooden boards or pieces. They are wearing a green costume headpiece that resembles a dragon or dinosaur, with spikes and horns. The setting looks like a living room or home office, with plants, a couch, and other furniture. It seems like the person might be preparing for a project, potentially building something with the materials on the table, though the exact task is unclear from the image.

Dataset context: What are the previous actions that lead to this frame in the furniture building process?

Llama 3.2: The person is wearing a dinosaur headpiece and appears to be looking at the white table with a piece of furniture on it. They are likely inspecting the assembly process to ensure everything is in order before proceeding. The person may be checking for any defects or issues with the furniture, such as scratches, dents, or misaligned parts. They may also be verifying that all the necessary parts are present and accounted for. Overall, the person is taking a careful and methodical approach to assembling the furniture.

GPT 4o: The person in the image appears to be in the process of assembling furniture. They are standing at a table with several wooden pieces laid out, which might be parts of a piece of furniture, possibly a shelf or a small table. They are wearing a costume, specifically a dinosaur or dragon-like headpiece, which adds a playful element to the scene. The setting is a cozy living room with a colorful rug, plants, and other home decor, suggesting this is a casual, at-home project.

Figure D. Qualitative results on the importance of contextual prompting. Frame extracted from IKEA-ASM. We use red color to highlight action-related errors, green for action-related valid answers, and orange for non-related visual descriptions.