Supplementary Material for: Compositional Context Fine-tuning Vision-Language Model for Complex Assembly Action Understanding from Videos

This supplementary document provides additional details that complement our main paper. It starts by introducing Dataset Details (Section 1), followed by Implementation Details (Section 2). Section 3 presents the error analysis, and Section 4 reports an ablation study that validates our layer partitioning strategy.

1. Dataset Details

This section provides comprehensive details of the two datasets introduced in this work: HA-ViD-VQA and IKEA-ASM-VQA. In Section 1.1, we present complete vocabulary lists including each action elements and holistic actions for both datasets. Section 1.2 then demonstrates sample distributions. In Section 1.3, we show the examples of VQA pairs for both datasets.

1.1. Complete Vocabulary Lists

HA-ViD-VQA Vocabulary In Table 1, 2, 3, 4, and 5, we provide the full lists of verbs, manipulated objects, target object, tools and holistic actions. For more information about the definition of the verbs and the details of the objects, please refer to the paper [1].

ID	Verb
1	insert
2	slide
3	place
4	rotate
5	screw
6	null

Table 1. HA-ViD-VQA Verb Vocabulary

IKEA-ASM-VQA Vocabulary In Table 6, 7, and 8, we provide the full lists of verbs, objects, and holistic actions in IKEA-ASM-VQA. For more details, please refer to the paper [2].

ID	Object	ID	Object
1	ball	14	small placer
2	assembly box	15	screw bolt
3	ball seat	16	hex screw
4	cylinder base	17	Phillips screw
5	cylinder cap	18	usb male
6	cylinder bracket	19	linear bearing
7	cylinder subassembly	20	worm gear
8	gear shaft	21	hand wheel
9	large gear	22	quarter-turn handle
10	small gear	23	hand dial
11	bar	24	nut
12	rod	25	null
13	large placer		

Table 2. HA-ViD-VQA Manipulated Object Vocabulary

ID	Object	ID	Object
1	ball	14	usb female
2	assembly box	15	screw hole C1
3	cylinder base	16	screw hole C2
4	ball seat	17	screw hole C3
5	cylinder bracket	18	screw hole C4
6	cylinder cap	19	worm gear
7	large gear	20	hole for the large gear
8	gear shaft	21	hole for the small gear
9	hole for the rod	22	hole for the worm gear
10	hole for the bar	23	screw bolt
11	hole for the bolt	24	nut
12	hole for the Phillips screw	25	null
13	stud on the assembly box		

Table 3. HA-ViD-VQA Target Object Vocabulary.

1.2. Dataset Statistics

Figures 1 and 2 respectively illustrate the distribution of action sample quantities for left and right hands in the HA-ViD-QVA dataset. The action samples are the same for three views in HA-ViD-QVA. The visualizations reveal a severe label imbalance problem for both hands, with *null*

ID	Tool
1	hex screwdriver
2	Phillips screwdriver
3	shaft wrench
4	nut wrench
5	null

Table 4. HA-ViD-VQA Tool Vocabulary.

ID	Actions	ID	Actions
1	insert the ball into the cylinder base	29	screw the nut onto the stud on the assembly box
2	insert the ball seat into the cylinder base	30	screw the nut onto the stud on the assembly box using the nut wrench
3	insert the cylinder cap into the cylinder bracket	31	screw the nut onto the screw bolt
4	insert the cylinder bracket into the cylinder base	32	screw the nut onto the screw bolt using the nut wrench
5	insert the large gear into the gear shaft	33	screw the screw bolt onto the nut
6	insert the small gear into the gear shaft	34	screw the hex screw into the screw hole C1
7	insert the bar into the hole for the bar	35	screw the hex screw into the screw hole C1 using the hex screwdriver
8	insert the rod into the hole for the rod	36	screw the hex screw into the screw hole C1 using a Phillips screwdriver
9	insert the large placer into the gear shaft	37	screw the hex screw into the screw hole C2
10	insert the small placer into the gear shaft	38	screw the hex screw into the screw hole C2 using the hex screwdriver
11	insert the screw bolt into the hole for the bolt	39	screw the hex screw into the screw hole C2 using the Phillips screwdriver
12	insert the hex screw into the screw hole C1	40	screw the hex screw into the screw hole C3
13	insert the hex screw into the screw hole C2	41	screw the hex screw into the screw hole C3 using the hex screwdriver
14	insert the hex screw into the screw hole C3	42	screw the hex screw into the screw hole C3 using the Phillips screwdriver
15	insert the hex screw into the screw hole C4	43	screw the hex screw into the screw hole C4
16	insert the hex screw into the cylinder bracket	44	screw the hex screw into the screw hole C4 using the hex screwdriver
17	insert the Phillips screw into the worm gear	45	screw the hex screw into the screw hole C4 using the Phillips screwdriver
18	insert the usb male into the usb female	46	screw the Phillips screw into the hole for worm gear
19	place the cylinder base onto the assembly box	47	screw the Phillips screw into the hole for worm gear using the Phillips screwdriver
20	place the cylinder bracket onto the assembly box	48	screw the Phillips screw into the hole for Phillips screw
21	place the worm gear onto the assembly box	49	screw the Phillips screw into the hole for Phillips screw using the Phillips screwdriver
22	screw the cylinder cap onto the cylinder base	50	slide the cylinder bracket
23	screw the gear shaft onto the hole for large gear	51	slide the linear bearing
24	screw the gear shaft onto the hole for large gear using the shaft wrench	52	rotate the worm gear
25	screw the gear shaft onto the hole for small gear	53	rotate the hand dial
26	screw the gear shaft onto the hole for small gear using the shaft wrench	54	rotate the quarter-turn handle
27	screw the nut onto the gear shaft	55	rotate the hand wheel
28	screw the nut onto the gear shaft using the nut wrench	56	null

Table 5. HA-ViD-VQA Action Vocabulary.

ID	Verb
1	align
2	attach
3	flip
4	insert
5	lay down
6	pick up
7	position
8	rotate
9	slide
10	spin
11	tighten
12	NA

Table 6. IKEA-ASM-VQA Verb Vocabulary.

(56) labels accounting for over one-third of all samples

ID	Object
1	leg
2	side panel
3	back panel
4	shelf
5	table
6	table top
7	pin
8	bottom panel
9	front panel
10	drawer
11	NA

Table 7. IKEA-ASM-VQA Object Vocabulary.

ID	Action
1	align leg
2	align side panel
3	attach back panel
4	attach shelf
5	attach side panel
6	flip table
7	flip table top
8	insert pin
9	lay down leg
10	lay down shelf
11	pick up back panel
12	pick up bottom panel
13	pick up front panel
14	pick up leg
15	pick up pin
16	pick up shelf
17	pick up side panel
18	pick up table top
19	position drawer
20	rotate table
21	slide bottom panel
22	spin leg
23	tighten leg
24	NA

Table 8. IKEA-ASM-VQA Action Vocabulary.

in each hand category. This distribution highlights both the challenging nature of the HA-ViD-QVA dataset and its authentic representation of real-world industrial scenarios. Figure 3 shows the label distribution in IKEA-ASM-VQA dataset. *NA* label accouts for 27.37% of all samples. We acknowledge that both datasets are relatively small in scale. This reflects our aim to assess the effectiveness of our proposed method under limited data conditions, along-

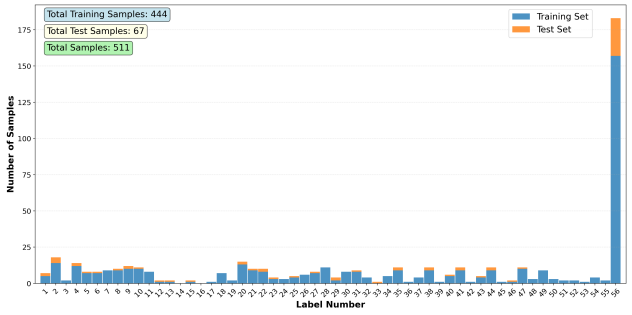

Figure 1. Distribution of Left-Hand Action Samples in the HA-ViD-VQA dataset.

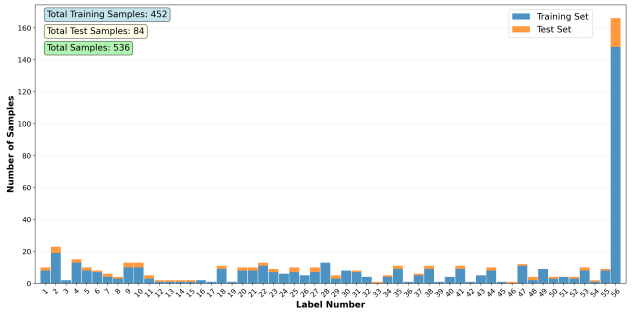

Figure 2. Distribution of Right-Hand Action Samples in the HA-ViD-VQA dataset.

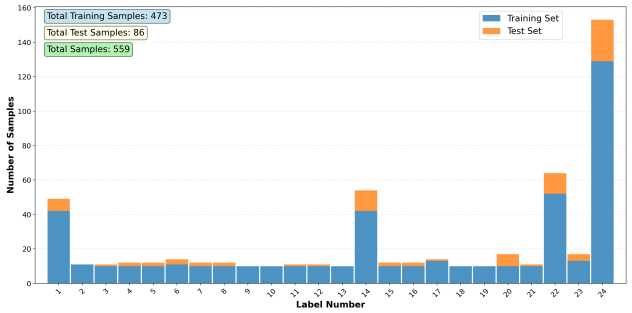

Figure 3. Distribution of Action Samples in the IKEA-ASM-VQA dataset.

side practical resource and time limitations. Critically, the dataset construction methodology we developed provides a scalable foundation for future expansion to larger, more diverse datasets.

1.3. VQA Template Examples

We adopt the structured VQA dataset format [1] provided by Qwen2.5-VL [3] for both datasets. Each JSON entry contains:

- Compositional questions targeting action elements (verbs, objects, tools).
- Compositional ground-truth answers.
- Video path.

Figure 4 illustrates a left-hand action sample from HA-ViD-VQA. The format remains consistent for right-hand actions, with questions adapted to target the right-hand verb. Figure 5 shows an example from IKEA-ASM-VQA. We create separate JSON files for each action element to enable LP-AT training.

```
[{
"messages": [
 {
"content": "<video>What is the action verb of the assembly action that the worker's left hand performed in the video?",
"role": "user"
 },
 {
"content": "The action verb is screw.",
"role": "assistant"
 }
 ],
"videos": ["video_path"]
},
]
```

(a)

```
[{
"messages": [
 {
"content": "<video>What are the objects being assembled during this assembly process?",
"role": "user"
 },
 {
"content": "The manipulated object is Phillips screw. The target object is hole for the Phillips screw.",
"role": "assistant"
 }
 ],
"videos": ["video_path"]
},
]
```

(b)

```
[{
"messages": [
 {
"content": "<video>What is the tool that the worker used in the video?",
"role": "user"
 },
 {
"content": "The tool is Phillips screwdriver.",
"role": "assistant"
 }
 ],
"videos": ["video_path"]
},
]
```

(c)

Figure 4. An Example of Compositional Question Answering Format for Left-hand Action Understanding in HA-ViD-VQA. (a) Action Verb QA Pair; (b) Objects QA Pair; (c) Tool QA Pair.

2. Implementation Details

While the main paper introduces the core implementation configurations for our method, this section provides additional implementation details necessary for reproducibility. We detail the computational requirements and resource usage for our method in Section 2.1, followed by baseline configuration specifics in Section 2.2.

2.1. Computational Requirements

Our experiments were conducted on a high-performance computing cluster with the following specifications:

2.1.1. Hardware Configuration

- **Training**: 4 × NVIDIA RTX A6000 GPUs (48GB VRAM each).

[1] https://github.com/QwenLM/Qwen2.5-VL/tree/main/qwen-vl-finetune

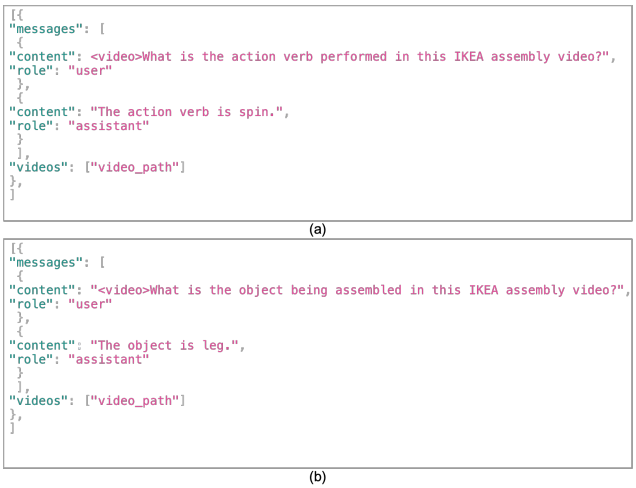

Figure 5. An Example of Compositional Question Answering Format in IKEA-ASM-VQA. (a) Action Verb QA Pair; (b) Objects QA Pair.

- **Inference**: Single NVIDIA RTX A6000 GPU.
- **CPU**: AMD Ryzen Threadripper PRO 3995WX 64-Cores.
- **RAM**: 512GB.

2.1.2. Model Specifications

- **Total Parameters**: 7B.
- **Frozen Components**: Visual encoders (675 million parameters) remain frozen during training.
- **Trainable Parameters**: 6.3B during fine-tuning (language decoders only).
- **Precision**: Mixed precision training (FP16) with gradient accumulation.

2.1.3. Resource Utilization

- **Peak Memory Usage**: 32 GB per GPU during training.
- **Batch Size**: 1 videos per GPU.
- **Inference Time**: 3.49 seconds per video (single GPU).

2.2. Baseline Implementation Details

We used the MMAction2 tool box [2] to benchmark TSM [4] and UniFormerV2 [5], while for VideoMAE V2 [6], we utilized the official implementation.

TSM Following the original paper's suggestions, we use the SGD optimizer with a dropout of 0.5. The learning rate was initialized as 0.0025 and decayed by a factor of 10 after epochs 20 and 40 frames were uniformly sampled from each clip. The TSM was pretrained on ImageNet [7], and we finetuned it on our 12 sub-datasets. We set the total training epochs to be 50 with a batch size of 16.

UniFormerV2 Following the default configuration from MMAction2, we use the AdamW optimizer. Epochs 1 to 5 have a linear learning rate with an initial learning rate of 0.00002 and start factor of 0.05. The subsequent epochs use a cosine annealing learning rate with the minimum learning rate ratio of 0.05. 8 frames were uniformly sampled from each clip. The Uniformerv2 was pre-trained on Kinetics-710 [5], a combined Kinetics-400/600/700 [8] benchmark. We set the total training epochs to be 50 with a batch size of 8.

VideoMAE V2 Following the default configuration from the official implementation [3], we employ the AdamW optimizer with an initial learning rate of 0.001, complemented by a 5-epoch linear warmup phase. The model uses a Vision Transformer base [9] architecture with 16×16 patches and 224×224 input resolution. We uniformly sample up to 16 frames from each video clip with a sampling rate of 4. The VideoMAE V2 model was pre-trained on Kinetics-710. Training is conducted for 50 epochs with a batch size of 4, incorporating regularization techniques including 0.3 drop path rate, 0.1 weight decay, and gradient clipping at 5.0.

3. Error Analysis

We conducted error analysis on the action recognition results of our method on HA-ViD-VQA and IKEA-ASM-VQA. This analysis reveals consistent patterns in the challenges of VLM-based assembly action recognition across different assembly contexts.

Tables 9 and 10 demonstrate the action element-wise striking patterns in the failures cases of our method across two datasets. For action verbs, the model frequently confuses verbs with similar motion trajectories, such as the rotate/screw and insert/place action pairs in HA-ViD-VQA, and slide/attach and tighten/spin action pairs in IKEA-ASM-VQA. This suggests VLMs still struggle to capture subtle differences in motion trajectories that distinguish these fine-grained assembly actions. Object recognition challenges reveal four distinct failure modes: small object size, visual similarity between objects, part-whole ambiguity, and limited spatial reasoning. In HA-ViD-VQA, small objects, e.g., the stud, holes, and the ball are particularly challenging (see Figure 6). Furthermore, visually similar components, such as small/large gears, ball/ball seat, and different types of holes, are often misidentified. IKEA-ASM-VQA exhibits part-whole confusion, where the entire table is misclassified as a leg. Finally, the target object recognition in HA-ViD-VQA shows severe spatial reasoning deficits, with screw hole positions (C1-C4) experiencing 76.73% error rate due to mutual confusion. These consistent patterns across datasets indicate that current VLMs

[2] https://github.com/open-mmlab/mmaction2

[3] https://github.com/OpenGVLab/VideoMAEv2

Action Element	Class	Error Rate/%	Common Misclassification
Verb	rotate	66.67	→ screw, insert
	place	61.44	→ insert
	insert	48.81	→ screw, place
MO	small gear	88.33	→ each other
	ball	66.67	→ ball seat
	ball seat	41.67	→ cylinder cap, ball
TO	stud	83.33	→ various targets
	screw hole C1-C4	76.73	→ other screw holes
	holes for [component]	76.76	→ other holes

Table 9. Most Problematic Classes by Action Element in HA-ViD-QVA. MO and TO denote manipulated object and target object. [Component] includes rod, bar, bolt, Phillips screw, large gear, small gear, and worm gear.

Action Element	Class	Error Rate/%	Common Misclassification
Verb	align	66.67	→ spin, attach
	tighten	61.44	→ spin
	slide	48.81	→ attach
Object	table	70.00	→ leg
	back panel	50.00	→ NA
	bottom panel	50.00	→ NA

Table 10. Most Problematic Classes by Action Element in IKEA-ASM-QVA.

lack the specialized spatial reasoning, motion understanding, and fine-grained visual discrimination required for complex assembly action recognition, highlighting the need for assembly-specific architectural enhancements or training strategies.

4. Layer Partition Experiment

Our partition strategy is motivated by two considerations: (1) the verb → object → tool progression intuitively follows the compositional order in our action decomposition framework, which also parallels the developmental trajectory in human motor learning: basic motor skills → object manipulation → tool use [10], and (2) the number of layers allocated to each action element corresponds to the estimated relative complexity of each recognition task. While this strategy is empirically rather than theoretically derived, we conducted an ablation study on a subset of the HA-ViD-VQA dataset (side-view, left-hand actions) to validate its efficacy. We selected HA-ViD-VQA for this study as it contains all four action elements (verb, manipulated object, target object, and tool).

Our layer partitioning strategy allocates 10-14-4 layers for verb-object-tool recognition respectively. To empirically validate this partitioning strategy, we compare three

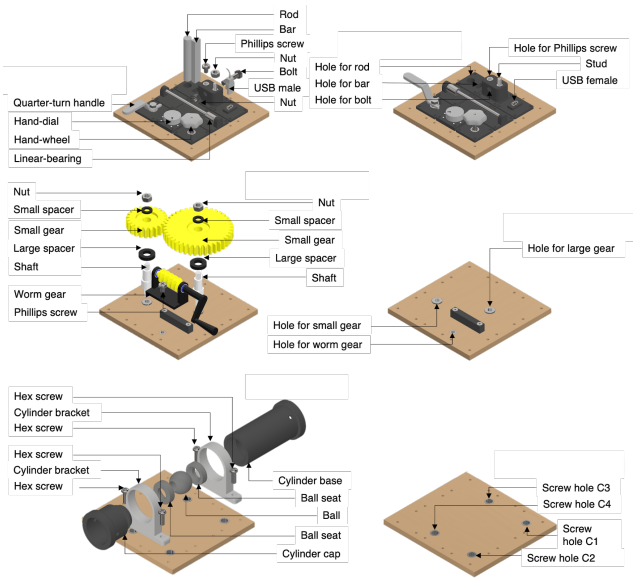

Figure 6. The objects in HA-ViD-VQA.

Layer partition verb-object-action	Verb v	MO o_m	TO o_t	Tool t	Action a
10-14-4 (ours)	61.19	**55.22**	**53.73**	88.06	**46.27**
5-19-4	37.31	52.24	43.28	**89.55**	40.30
15-9-4	**64.18**	49.25	43.28	80.60	44.78

Table 11. Ablation study of layer partitioning strategies on side-view left-hand action recognition. We report element-wise accuracy (%). MO: manipulated object, TO: target object.

allocation schemes:

- **Baseline (10-14-4)**: Balanced allocation based on task complexity.
- **Object-heavy (5-19-4)**: Less layers for verb recognition, more layers for object recognition.
- **Verb-heavy (15-9-4)**: More layers for verb recognition, less layers for object recognition.

The allocation for tool recognition was held constant at 4 layers, as this already yielded strong performance. All experiments used identical implementations and training protocols; table 11 reports the results with highest action recognition accuracy achieved by each for a fair comparison.

The results, presented in Table 11, demonstrate that our 10-14-4 allocation achieves the highest overall action recognition accuracy (46.27%) and maintains strong performance across all individual elements. While the verb-heavy configuration excels in verb recognition (64.18%), it significantly harms object recognition, leading to a lower final action accuracy (44.78%). The object-heavy configuration severely degrades verb recognition with only 37.31% accuracy, undermining overall performance. Besides, its lower performance of object recognition suggests that beyond a

certain threshold, additional layers provide no further benefit. This ablation study confirms that our empirically chosen partition strategy offers a superior balance, validating our initial design considerations.

References

[1] H. Zheng, R. Lee, and Y. Lu, "Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 67069–67081, Curran Associates, Inc., 2023. 1

[2] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, "The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 846–858, 2021. 1

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. 3

[4] J. Lin, C. Gan, K. Wang, and S. Han, "Tsm: Temporal shift module for efficient and scalable video understanding on edge devices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2760–2774, 2022. 4

[5] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "Uniformerv2: Unlocking the potential of image vits for video understanding," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1632–1643, 2023. 4

[6] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, 2023. 4

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 4

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. 4

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. 4

[10] A. W. Needham and E. L. Nelson, "How babies use their hands to learn about objects: Exploration, reach-to-grasp, manipulation, and tool use," vol. 14, no. 6, p. e1661. 5