ConsisVLA-4D: Advancing Spatiotemporal Consistency in Efficient 3D-Perception and 4D-Reasoning for Robotic Manipulation

Supplementary Material

A. Implementation Details

A.1. Model Details

CV-Aligner. CV-Aligner aligns multi-view visual observations with task instructions and integrates 3D information from VGGT [10] through a three-step process: feature modulation, token selection, and cross-view fusion.

- **Step 1 (FiLM Modulation).** Although each visual token $\mathbf{z}^{\text{sem},j}$ from SigLIP [11] inherits pretrained linguistic semantics, we further strengthen the alignment between observations and the task instruction $\mathbf{t}$ for robotic manipulation by modulating SigLIP with FiLM [8] scale and shift vectors $\gamma_i(\mathbf{t})$ and $\beta_i(\mathbf{t})$. Specifically, the instruction $\mathbf{t}$ is projected into the visual embedding space at every SigLIP transformer layer to produce layer-specific $\gamma_i(\mathbf{t})$ and $\beta_i(\mathbf{t})$, which are then applied as independent per-token multiplicative and additive modulations.
- **Step 2 (ES-Selection).** We first encode the task instruction $\mathbf{t}$ using $f_t^{\text{SigLIP}}(\cdot)$, then compute its cosine similarity with every visual token from each camera view. For each view, we retain the top $1/8$ of the original tokens ($256 \rightarrow 32$). ES-Selection is performed independently for each view, producing $\mathbf{z}^{\text{sem}}_{\{M,L,R\}}$.
- **Step 3 (Single-Fusion).** We fuse $\mathbf{z}^{\text{sem}}_{\{M,L,R\}}$ with the VGGT-derived $\mathbf{z}^{\text{3D}}_{\{M,L,R\}}$ via a cross-attention module. The fusion is instantiated as a 4-layer Transformer [9] configured with a 1152-dimensional hidden size, 16 attention heads, and a 2752-dimensional feed-forward block. The 1024-dimensional visual features are linearly projected into the Transformer's hidden space to serve as keys and values. This process produces $\mathbf{z}^{\text{obj-3D}}_{\{M,L,R\}}$, effectively injecting VGGT's 3D cues while preserving a fixed budget of 32 tokens per view.

CO-Fuser. Unlike VGGT and SigLIP, which fuse only their final token outputs, we leverage the structural similarity between VGGT and DINOv2 [7] to enable a deeper form of fusion. Both encoders are geometry-enhanced, non-textual semantic contrastive models, and VGGT's large-scale 3D pretraining is itself grounded in DINOv2. This architectural alignment makes it possible to perform dense, block-level fusion across the two encoder streams.

- **Step 1 (Group-Fusion).** For the l-th block, we compute

$$\mathbf{z}_l^{\text{geo-3D}} = (1-\alpha_l) \odot \mathbf{z}_l^{\text{geo}} + \alpha_l \odot \mathbf{z}_l^{\text{3D}},$$

where the frozen VGGT 3D features decay with layer depth following a cosine schedule, while the learned fusion features gradually gain weight through block-wise integration. In Eq. (15), $\alpha_0 = \psi = 0.2$, $\alpha_{\mathcal{L}'} = \psi \cdot \delta = 0.01$, and $\mathcal{L}' = 24$.

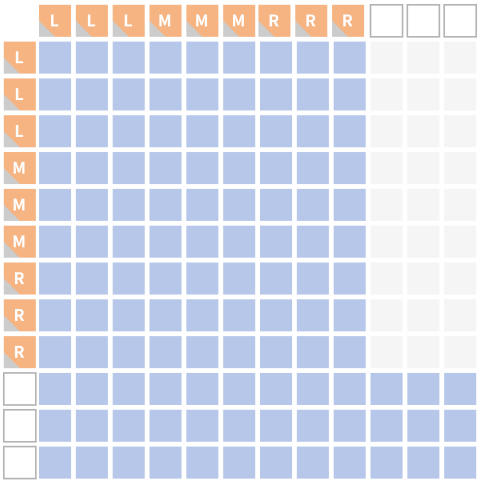

Figure 1. **Block-Wise Causal Self-Attention.** Multi-view observations $\mathcal{I} = \{M, L, R\}$ (Main, Left, Right). Compression ratio reaches 1/12. □ represents the Aggregation Token.

- **Step 2 (Aggregation Token Concatenation).** To mitigate visual ambiguities from single viewpoints by incorporating geometric relations across views, we initialize a set of aggregation tokens $\mathbf{z}_0^{\text{agg-3D}}$ and concatenate them to $\mathbf{z}_0^{\text{geo-3D}}$. These tokens are learned progressively through the blocks, with their number fixed at 64 (compressed to 1/8 for dual-view single-arm tasks and 1/12 for triple-view dual-arm tasks).
- **Step 3 (IG-Aggregation).** We design a block-wise causal self-attention mechanism, illustrated in Fig. 1, for IG-Aggregation. Causal attention is applied between $\mathbf{z}_l^{\text{geo-3D}}$ and $\mathbf{z}_l^{\text{agg-3D}}$, while bidirectional attention is used within each set. This ensures effective flow and aggregation of information from different views into $\mathbf{z}_l^{\text{agg-3D}}$. Finally, only the 64 aggregation tokens from the last block, $\mathbf{z}_{\mathcal{L}'}^{\text{agg-3D}}$, are retained.

CS-Thinker. Based on $\mathbf{z}^{\text{obj-3D}}_{\{M,L,R\}}$ from CV-Aligner and $\mathbf{z}_{\mathcal{L}'}^{\text{agg-3D}}$ from CO-Fuser, CS-Thinker learns implicit semantic and geometric knowledge for reasoning via Spatiotemporal Consistency Attention, a Training-Only dynamic decoder, and depth decoders. Specifically,

- **1)** three sets of initialized dynamic tokens ($3 \times 4 = 12$

Table 1. Comparison of the Performance between the α_l Settings in Eq. 15 and Linear Decay.

$\frac{d\alpha_l}{dl}$	**LIBERO SR ↑**	**Real-World SR ↑**
$-\frac{\psi\cdot(1-\delta)}{2}\cdot\frac{\pi}{\mathcal{L}'}\cdot\sin\left(\frac{l\pi}{\mathcal{L}'}\right)$	**98.1**	**78.3**
1.0	94.4 (-3.7)	73.3 (-5.0)
0.1	95.9 (-2.2)	75.0 (-3.3)

tokens) are used to decode single-view dynamic object representations;

- **2)** one set of initialized depth tokens ($1 \times 4 = 4$ tokens) is used to decode global depth across three views.

Accordingly, CS-Thinker employs 1 dynamic decoder and 3 independent depth decoders. Each decoder consists of 8 Transformer blocks with a hidden size of 1024, 16 attention heads per block, and a feed-forward ratio of 4.

α_l Settings. We analyze the design of the layer-wise adaptive weight α_l in the Group-Fusion module (Eq. 15), showing that the cosine decay mechanism is intentional. The derivative $\frac{d\alpha_l}{dl}$ approaches 0 as $l \to 0$ (shallow layers) and $l \to \mathcal{L}'$ (deep layers).

- **1)** At the initial stages of training, α_l remains close to a high value, ψ, with a near-zero slope, allowing the model to focus on absorbing prior geometric information and aligning it under strong constraints.
- **2)** In the later stages, α_l smoothly decreases toward its minimum value, $\psi \cdot \delta$, with the slope again approaching zero. This transition enables the model to steadily shift attention towards learned geometric features in the final layers, effectively exiting prior geometric constraints.
- **3)** The rate of change of cosine decay is concentrated in the middle layers ($l \approx \mathcal{L}'/2$), reflecting a common trend in feature learning: models tend to focus on integrating both low-level and high-level information. The cosine decay ensures that the intensity of geometric priors primarily shifts in this region, aiding in the acceleration and completion of complex feature abstraction.

In contrast, with linear decay, the rate of change of α_l with respect to l is constant. This results in a uniform removal of geometric priors across the entire network depth, leading to a sharp, discontinuous transition during optimization. Our α_l settings, with a variable rate, offer a performance advantage, as shown in Tab. 1.

A.2. Training Details

Single-Arm Task Training. We adopt OpenVLA [5] as the backbone, with an action chunk size of $K = 8$. Fine-tuning is performed using low-rank adaptation (LoRA [4]) with rank 32 and $\alpha = 64$. The model is trained for 80K steps with a batch size of 64 and an initial learning rate of 5×10^{-4}. Checkpoints are evaluated every 10K steps, and the best-performing checkpoint is reported.

Dual-Arm Task Training. For dual-arm tasks, the action chunk size is set to $K = 25$, and OpenVLA is fine-tuned using LoRA with rank 32 and $\alpha = 64$. The model is trained for 80K steps with a batch size of 32. The initial learning rate is 5×10^{-4} and decayed to 5×10^{-5} after 50K steps. Checkpoints are evaluated every 10K steps from step 80K onward, and the best-performing checkpoint is reported.

B. Experimental Details

B.1. Simulation Benchmark

LIBERO. LIBERO (LIfelong learning BEnchmark for RObot manipulation) [6] is a simulation-based evaluation platform designed to investigate lifelong learning and multi-task transfer in robotic manipulation. It consists of four structured task suites, with each suite containing 10 tasks, each evaluated over 50 trials (**10 tasks** $\times$ **50 trials**).

RoboTwin 2.0. We randomly selected 7 different types of dual-arm tasks based on ALOHA [2] in RoboTwin 2.0 [1] to test the model's robust dual-arm manipulation ability, conducting 100 trials for each task (**7 tasks** $\times$ **100 trials**):

- **Click Alarmclock:** Click the center of the top-side button on the alarm clock on the table.
- **Turn Switch:** Use the robotic arm to flip the switch.
- **Put Bottles Dustbin:** Use the arms to grab the bottles and place them in the dustbin to the left of the table.
- **Open Laptop:** Use one arm to open the laptop.
- **Press Stapler:** Use one arm to press the stapler.
- **Place Empty Cup:** Use one arm to place the empty cup on the coaster.
- **Blocks Ranking RGB:** Place the red, green, and blue blocks in the order of red, green, and blue from left to right, placing them in a row.

B.2. Real-World Setup

Experimental Platforms. We validate the real-world performance of ConsisVLA-4D on two advanced mobile manipulation platforms: the AgileX Cobot Magic [2] and the Galaxea R1 Lite [3].

- The AgileX Cobot Magic platform, developed by AgileX Robotics based on Stanford's ALOHA project, integrates a differential drive AGV chassis (Tracer), a four-arm collaborative system, and RGB-D sensors.
- The Galaxea R1 Lite platform, developed by Galaxea Dynamics, is a modular dual-arm mobile platform designed for data collection and embodied intelligence development. It features a 23-DOF configuration (6-DOF chassis, 3-DOF torso, 7-DOF single-arm with gripper), an omnidirectional chassis with three steering wheels, and a suite

of perception modules, including binocular cameras, a monocular depth camera, and LiDAR.

Task Description & Evaluation Protocol. In addition to the tasks reported in the main text (Task 1 through Task 4: Microwave Operation, Banana Peeling, Drawer Arrangement, and T-shirt Folding). Tasks 1 through 4 collected **60**, **60**, **60**, and **45** expert demonstrations, respectively. ConsisVLA-4D exhibited consistently strong performance across all tasks.

- **Task 1**: ***"Put the bread into a bowl and heat it in the microwave."***
- **Task 2**: ***"Peel the banana and place it on the plate."***
- **Task 3**: ***"Open the drawer, put the toy inside, and then close the drawer."***
- **Task 4**: ***"Fold the T-shirt."***

C. Supplementary Visualizations

C.1. Task Execution Visualizations

As shown in Fig. 2, we present visualizations of key execution-stage observations for Tasks 1-2 in real-world settings. Tasks 1 and 2 assess the model's long-range task execution capability in complex environments, its alignment with multi-stage semantic instructions, and its precise understanding of spatial and geometric relationships.

- In **Task 1** (Microwave Operation), the robot successfully executes a long sequence of actions involving object nesting and spatial constraints: it first places the bread accurately into the bowl, then smoothly inserts the bowl into the narrow interior of the microwave, and finally closes the microwave door.
- In **Task 2** (Banana Peeling), the model demonstrates exceptional dual-arm coordination. The robot uses one arm to stabilize the object while the other arm performs the delicate peeling operation, successfully placing the peeled banana onto a plate.

C.2. Additional CV-Aligner Visualization Results

As shown in Fig. 3, we present additional qualitative visualizations of CV-Aligner. Four specific instruction cases were selected, asking the robot to pick up alphabet soup, butter, cream cheese, and milk. Each row corresponds to a particular text instruction, with attention heatmaps unfolding over time steps in both the Main View and Wrist View. It is evident that CV-Aligner effectively filters out target objects that highly match the instruction semantics from redundant background objects, validating the effectiveness of the visual redundancy removal via the Top-K selection mechanism in Equations (11-12).

C.3. Additional CO-Fuser Visualization Results

As shown in Fig. 4, which differs notably from Fig. 3, the attention heatmaps of CO-Fuser illustrate how the model understands the global geometric layout of the scene and the spatial relationships between objects. Unlike focusing solely on a single object, CO-Fuser covers multiple spatial nodes relevant to the task. CO-Fuser initializes a set of Aggregation Tokens that account for only 1/12 to 1/8 of the original patch tokens, and through this set, implicitly captures the geometric relationships across all viewpoints, integrating discrete object features into coherent spatial structure information. This complements the functionality of CV-Aligner.

Comparing the heatmaps of the Main View and Wrist View, we observe that CO-Fuser constructs a more robust spatial representation by leveraging the complementary nature of multi-view information. The Main View provides global context on the relative positions of objects (e.g., distance between an object and a basket), while the Wrist View supplements with close-range geometric details. This cross-view geometric fusion effectively addresses spatial localization uncertainty in a single viewpoint.

References

[1] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 2

[2] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024. 2

[3] Galaxea. Galaxea r1 lite. *https://galaxea-dynamics.com/*, 2025. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[5] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2

[6] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. 2

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

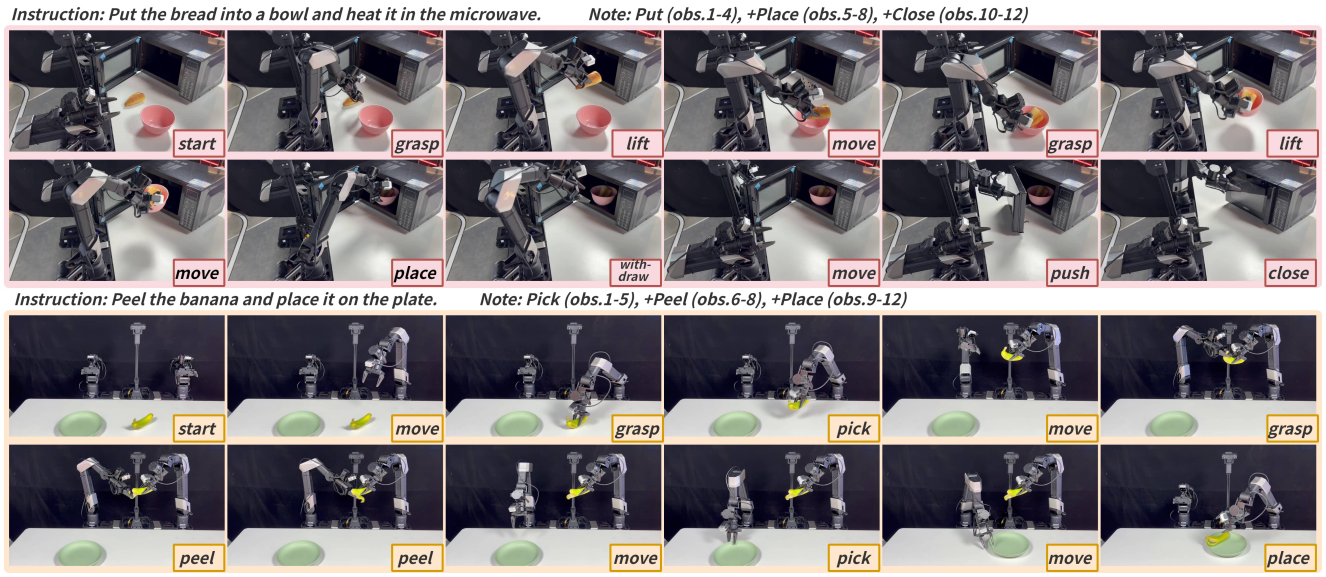

Figure 2. **Visualization of Task 1 and Task 2 Execution**, illustrating key execution-stage observations in full.

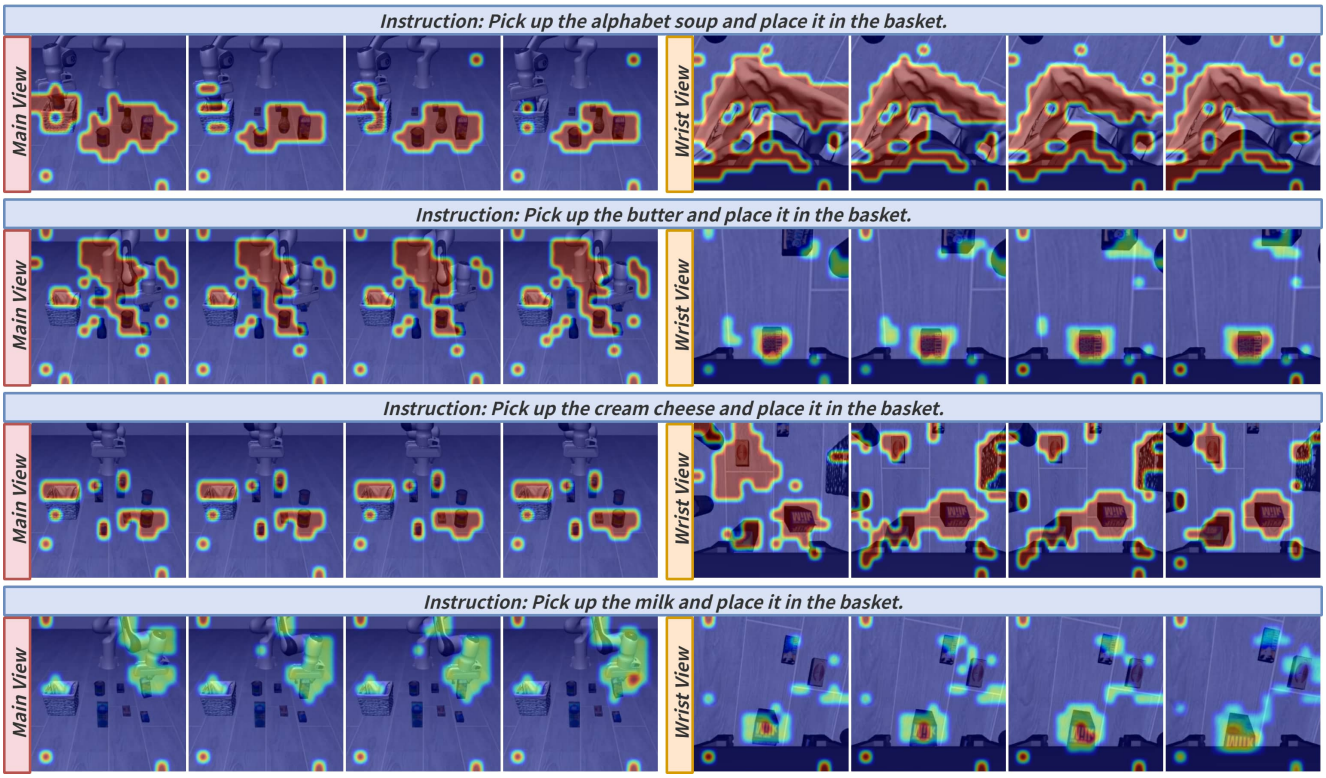

Figure 3. **Additional Qualitative Visualizations of CV-Aligner.** This figure illustrates the attention heatmaps generated by the CV-Aligner module in the Main View and Wrist View under different language instructions.

[8] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

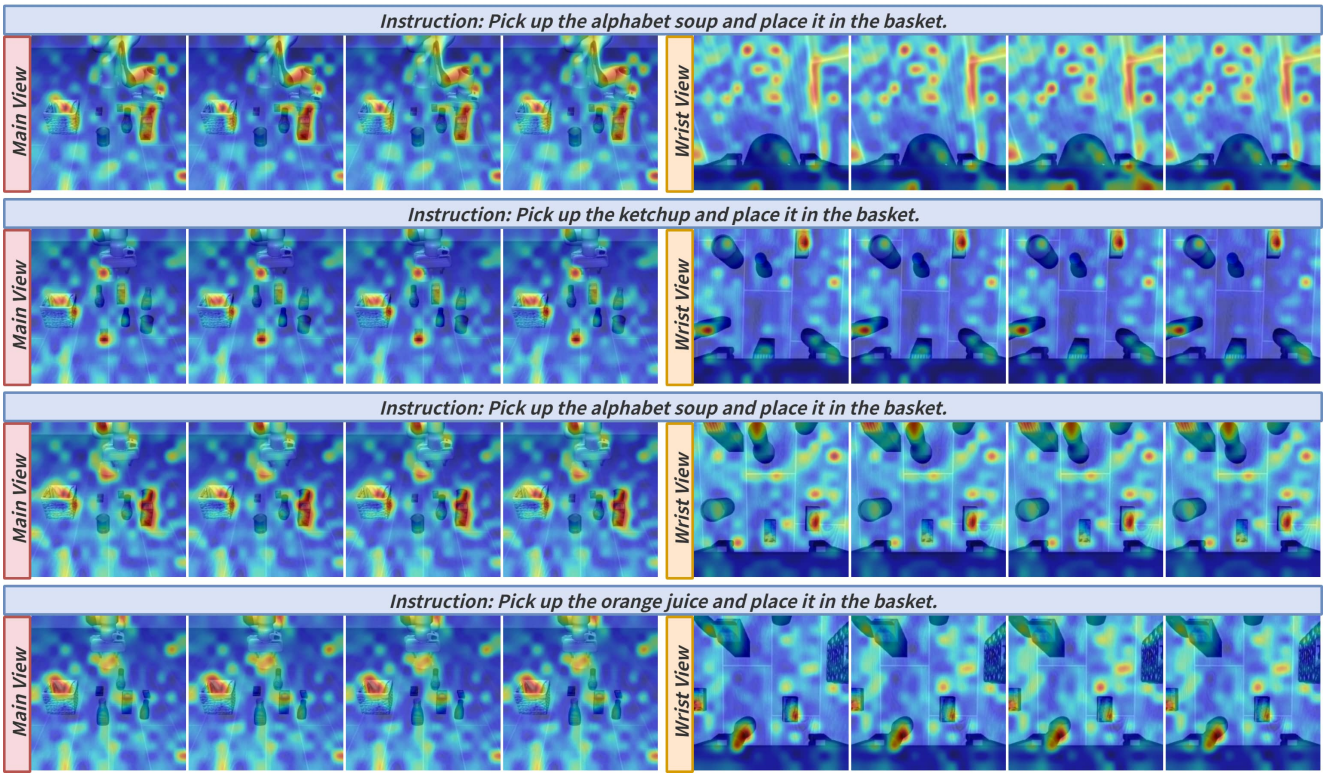

Figure 4. **Additional Qualitative Visualizations of CO-Fuser.** This figure illustrates the attention heatmap between the Aggregation Tokens extracted by the CO-Fuser module and the original visual patch tokens. Unlike the single-point focus of CV-Aligner, CO-Fuser presents a distributed attention pattern, complementing the focus of CV-Aligner on instruction-relevant objects.

[10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

[11] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1