IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

BFA: Best-Feature-Aware Fusion for Multi-View Fine-grained Manipulation

Zihan Lan[1*] Weixin Mao[2*†] Haosheng Li[3*] Le Wang[4] Tiancai Wang[1] Haoqiang Fan[1] Osamu Yoshie[2]

***Abstract*—In real-world scenarios, multi-view cameras are typically employed for fine-grained manipulation tasks. Existing approaches (e.g., ACT [1]) tend to treat multi-view features equally and directly concatenate them for policy learning. However, it will introduce redundant visual information and bring higher computational costs, leading to ineffective manipulation. Fine-grained manipulation tasks typically consist of multiple stages, where the best view may vary across different phases. This paper proposes a plug-and-play Best-Feature-Aware (BFA) fusion strategy for multi-view manipulation tasks, which is adaptable to various policies. Building upon the visual backbone of the policy network, we design a lightweight subnetwork to effectively predict the importance score of each view. Based on the predicted importance scores, the reweighted multi-view features are subsequently fused and fed into the end-to-end policy network for seamless integration. Notably, our method demonstrates outstanding performance in fine-grained manipulations. The experimental results show that our approach outperforms multiple baselines by 22-46% success rate on different tasks. Our work provides new insights and inspiration for tackling key challenges in fine-grained manipulations.**

***Index Terms*—Computer vision, Deep learning, Robotics, Robot control, Learning systems**

I. Introduction

Imitation learning [2], [3] for robot manipulation [1], [4]–[6] enables robots to learn and replicate operations from human demonstration data. While existing approaches achieve promising results in general manipulation tasks, fine-grained manipulation remains particularly challenging due to its demand for high precision. Addressing these challenges requires comprehensive scene understanding, which demands multi-view observations to capture both the global context and detailed local interactions.

Notably, the importance of these views varies significantly across different stages of the manipulation process. As shown in the Figure. 1 (a), When a robotic arm initially approaches an object, the top view capturing the entire scene becomes crucial, providing essential information about global spatial relationships, scene layout, and target object positioning - while the wrist cameras may not even have the target objects in view. The emphasis shifts markedly during fine-grained manipulation like precise grasping or insertion, where the head or top camera view becomes invaluable by capturing detailed local interactions between the end-effector and target object, enabling precise alignment and depth perception. Therefore, it is important to switch view dynamically during the manipulation process, which can help the model mainly focus on the best camera view at each manipulation stage to better capture crucial spatial and contextual information and enhance operational precision. However, existing methods [1], [4]–[7] typically adopt oversimplified strategies that treat all views as equally important. They either simply concatenate features from different views or directly stack multiple images, neglecting the dynamic significance of each view. This uniform treatment overlooks the evolving importance of different views during manipulation and may bring a significant amount of distracting, unnecessary information, finally leading to reduce manipulation effectiveness and precision.

In this paper, we propose a novel learning-based Best-Feature-Aware (BFA) fusion strategy to address this often overlooked challenge. Our framework dynamically predicts the importance scores of multiple viewpoints by assessing the current interaction state between the robotic arm and objects. Specifically, a lightweight Score Network is introduced to evaluate the significance of each view. Based on the predicted importance scores which can be viewed as signal-to-noise ratios (SNR) of each view, we reweight and fuse the multi-view features, ensuring that the most useful information is effectively integrated to enhance policy performance as shown in the Figure. 1 (b). This plug-and-play component enhances the interaction with the environment through adaptive visual perception.

Furthermore, we designed an automated annotation framework using Vision-Language Models (VLM) that produces the multi-view ground-truth of importance score. Our system analyzes linguistic and visual proprioception information to categorize the current **state** of each robotic hand into one of four states: “holding”, “approaching”, “operating” and “returning” . Through carefully designed task-specific rules for combining these states as shown in the Tab. I, we decompose the entire manipulation process into distinct **stages**, with each stage focusing different views. We use the annotations from the VLM annotation system to train the above Score Network. Our method is evaluated on bimanual manipulation platform ALOHA [1] in real world as shown in Figure. 1(a). The effectiveness of our BFA strategy is further validated on two typical state-of-the-art imitation learning methods, RDT [4] and ACT [1]. Remarkably, our method demonstrates outstanding performance across various complex fine-grained manipulation tasks, such as “unzipping bag” and “opening box”. It achieves a success rate improvement of **22-46**% with existing methods. Moreover, the proposed BFA strategy can reduce the overall computation burden thanks to dynamic view selection. BFA can be regarded as a dynamic network from the

[1] MEGVII Technology, Beijing, China
[2] Waseda University, Tokyo, Japan maowx2017@fuji.waseda.jp
[3] Institute of Software, Chinese Academy of Sciences, Beijing, China
[4] Beihang University, Beijing, China
* indicates contributed equally to this work. † indicates the corresponding author.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

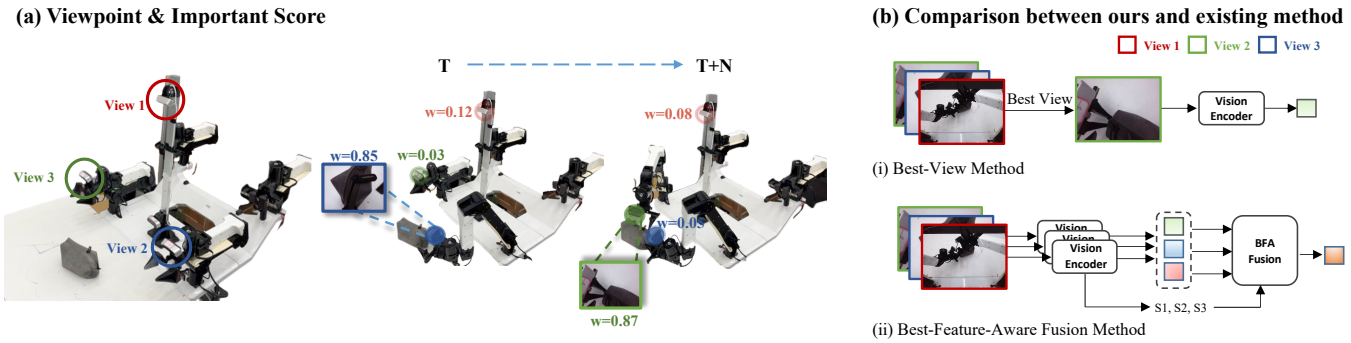


Fig. 1. (a) Our BFA method evaluates the importance score of each view during manipulation, as indicated by w. (b) Unlike the common best-view-based methods, our approach does not simply select the best view. Instead, it performs feature fusion based on the predicted importance scores $S1, S2, S3$ to derive the most suitable observation features for the manipulation task.

perspective of visual perception, improving the effectiveness of policy approaches. We hope the BFA mechanism can provide some new insights for the robotic manipulation community.

II. RELATED WORK

A. View Planning in Manipulation

View planning in robotics has been widely introduced in [8], [9], which seeks to determine the maximum information gain viewpoint and ensure the sequence of sensors. Among the various domains of view planning, one key area of interest is manipulation , which has been explored through various approaches to optimize task performance. Arruda et al. [10] proposed a geometry-based method that prioritizes object visibility and graspability, improving both the quality of reconstruction and the success of grasp. Alternatively, Jun Lv et al. [11] introduced the differentiates between the manipulation arm and viewpoint arm, magnifying the operation area through viewpoint following to enhance grasping stability. Other methods [12]–[14] selected the view with the greatest information gain during operation to address issues such as occlusion. In recent years, learning-based approaches have been increasingly used to optimize view planning. Some approaches [15]–[17] optimizes viewpoints planning process via reward functions during manipulation. Additionally, Multi-View Picking [18] applied a self-supervised state representation methods to focus on the target by changing views, enabling the completion of complex manipulation tasks.

B. Fine-Grained Robotic Manipulation

Current methods often employ imitation learning strategies to complete fine-grained manipulation tasks. By leveraging expert demonstrations, imitation learning enables the agent to efficiently acquire complex skills. Some methods [1], [5], [6], proposed an imitation learning framework based on transformer architecture [19], leveraging multi-view information and joint data as demonstration inputs to predict future action sequences. Additionally, Some works [7], [20], [21] integrate the diffusion process into imitation learning. Moreover, Some works [22], [23] have introduced a multi-task approach within these two paradigms, aiming to use a single model for handling multiple tasks. However, all these methods integrate multi-view information by directly concatenating all visual representations, without considering the unequal information provided by different viewpoints.

III. METHODOLOGY

In this section, we will first describe the overall architecture of BFA applied in existing policy networks (see Sec. III-A). Then we will provide the detailed implementation of BFA, as shown in Sec. III-B. After that, the VLM annotation system is presented to generate the ground-truth of importance scores, which is shown in Sec. III-C. In Sec. III-D, we provide an in-depth analysis of the mechanism behind the effectiveness of BFA.

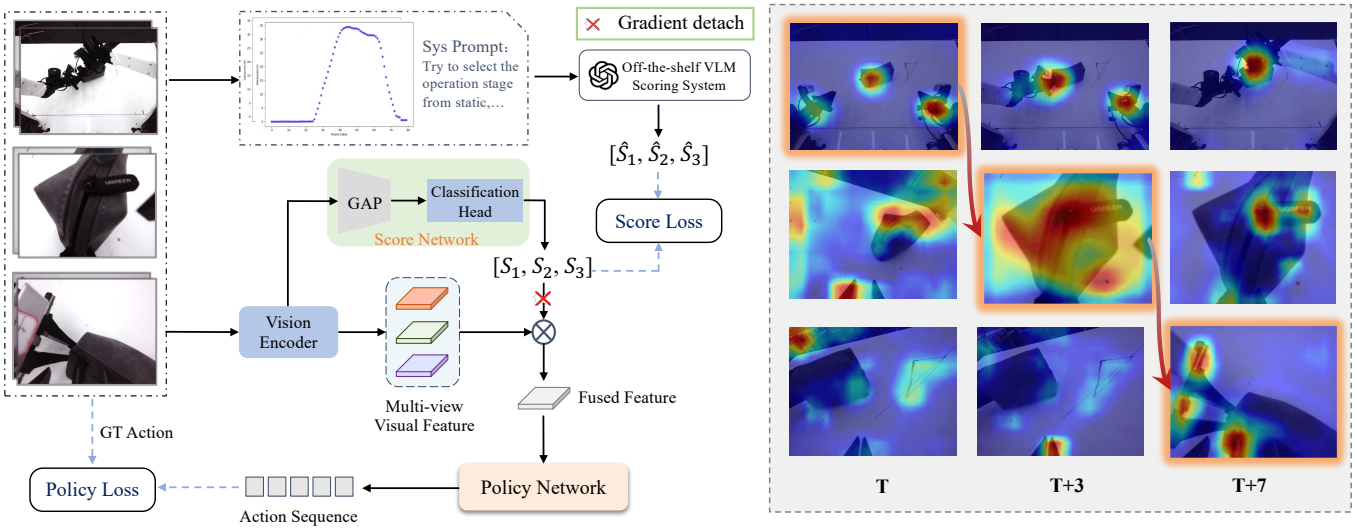
A. Overall Architecture

Our proposed best-feature-aware (BFA) fusion method is a general strategy which can be viewed as a plug-and-play module used in different end-to-end imitation learning methods. As shown in Fig. 2 (a), given multi-view RGB images from top-view and wrist cameras, the vision encoder (e.g., ResNet-18 [24], SigLIP [25]) extracts the multi-view visual features respectively. BFA takes the multi-view features as the inputs and uses Score Network (e.g., Multi-Layer Perceptions) to generate the importance score for each view. The generated importance scores are used to reweight and fuse the multi-view features, producing one fused feature which also is the input of subsequent policy network. The policy network (e.g., ACT [1], RDT [4]) predicts the action sequences for the deployment of real arm. During training, the Score Network is supervised by the importance score generating from the vision-language model (e.g., GPT-4o [26]).

B. Best Feature Aware

Previous studies [1], [4], [27]–[29] typically treat images from different views equally. We observe that each view contributes differently at various stages of the task. Accordingly, we argue that views should not be treated uniformly; instead, the most informative ones should play a dominant role during their most relevant stages. While primary features typically carry critical information, secondary features often provide complementary cues that further improve performance. Discarding them may limit the model's performance, so we

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Overall Architecture

(b) Attention Transfer Visualization

Fig. 2. The overall pipeline of best feature aware fusion strategy applied in the end-to-end policy network. Top-view and wrist-camera images are first fed into visual backbones to extract per-view features. Those features go into a lightweight scoring network, which assigns an importance weight to each view. The importance scores are further used to reweight and fuse the multi-view features. The fused features are finally served as the input of the policy network to generate the action sequence for real-arm deployment. Moreover, an off-the-shelf VLM annotation system is used to annotate the importance scores of each view offline. During training, the whole network is jointly optimized by the score loss and the policy loss.

propose retaining selective secondary features via a Best-Feature-Aware (BFA) fusion strategy.

Usually, existing imitation learning policies tend to employ the vision backbone (e.g., ResNet-18, SigLIP [25]) to extract the visual features. Let $f_i \in \mathbb{R}$ represent the feature vector extracted from the i-th view, where $i \in 1, 2, \ldots, N$ and N is the total number of views. The vision backbone processes each view respectively and generates a set of multi-view features:

$$\mathcal{F} = \{f_1, f_2, \ldots, f_N\}.$$

To achieve prioritized viewpoint selection, we reuse the multi-view visual features and design a plug-and-play lightweight network F_{cls} to predict the importance scores for each view. For both ACT [1] and RDT [4], the multi-view features $\{f_1, \ldots, f_N\}$ are compressed to low-dimensional representations via global average pooling (GAP). Then we fed the extracted feature to the Class Head F_{cls} to predict the importance score, $s_1, \ldots, s_N$ for all views.

$$s_i = F_{cls}(GAP(f_i)) \tag{1}$$

In our practice, we employ a three-layer linear network as Class Head to predict the importance scores. To better integrate multi-view feature information, not only the visual feature of the most important view is used as the observation feature of policy network. Instead, we utilize the predicted importance scores to reweight the features of corresponding views and then perform the element-wise addition operation to fuse the reweighted multi-view features.

$$\hat{f} = \sum_{i=1}^{N} f_i \times s_i \tag{2}$$

Here, $\hat{f}$ is the fused visual representation for fine-grained manipulation. The fused feature $\hat{f}$ is then fed into subsequent policy network for action sequence prediction. This element-wise addition allows for a more comprehensive fusion, avoiding the risk of the total information loss of unimportant views. Additionally, the continuous adaptation of feature weights ensures a smoother and more stable integration.

The policy loss L_p is used to optimize the parameters of policy networks as well as the visual backbone. Correspondingly, we add the score loss L_s as a auxiliary task.

$$L_{\mathrm{s}} = BCE(s, \hat{s}) \tag{3}$$

where BCE is the binary cross-entropy loss and $\hat{s}$ is the ground-truth of the importance score annotated by the VLM, which is described in Sec. III-C. Therefore, the overall loss function can be formulated as:

$$L = \lambda_1 L_{\mathrm{s}} + \lambda_2 L_p \tag{4}$$

where λ_1, λ_2 represents the weights for two loss functions. During training, the gradients of policy loss are propagated through all components except Score Network as shown in the Figure. 2, while the score loss gradients are simultaneously propagated to the vision encoder.

C. VLM Scoring System

To generate the ground-truth of importance score, we develop a VLM scoring system. To achieve human-level annotation quality in our system, we combine the rule-based methods with a Vision-Language Model (VLM) [26]. Using proprioceptive information from the robotic arms, we divide the entire episode into different stages For each frame, we generate two scatter plots that represent the change in distance from the grippers of the left and right robotic arms to their respective resting positions over time. These scatter plots are then input into the VLM [26], which is queried to

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

determine the current state of each robotic arm (e.g. “holding”, “approaching”, “operating” and “ ”returning”), as shown in Fig. 4.

To identify the state of both the left and right robotic arms, we combine the phases of both arms and apply hard rules which is shown in the Tab. I to calculate the view score of our system for the different viewpoints at that moment. For instance, if the left arm is in a “holding” state while the right arm is in an “operating” state, more attention should be given to the right-hand view. In this scenario, the importance score would be [0, 0, 1], allowing the system to focus mainly on the right-hand view. Furthermore, to optimize resource usage, we assume that during transitions — when the robotic arms move from the resting position towards an object or return to the resting position — the focus should shift to the top camera view. In this case, the importance score would be [0, 1, 0].

TABLE I
STATE AND STAGE RULES OF SOME TASKS

Task Name	Left Arm State	Right Arm State	Important Score
Unzip bag	Approaching	Holding	Top [0,1,0]
	Operating	Holding	Left [1,0,0]
	Operating	Approaching	Top [0,1,0]
	Operating	Operating	Right [0,0,1]
	Returning	Returning	Top [0,1,0]
Fold towel	Approaching	Approaching	Top [0,1,0]
	Operating	Operating	Left&Right [1,0,1]
	Returning	Returning	Top [0,1,0]
Open box	Approaching	/	Top [0,1,0]
	Operating	/	Left [1,0,0]
	Returning	/	Top [0,1,0]

Additionally, we implement a frame-skipping annotation strategy in which we annotate every five frames. If the annotations before and after the 5-frame window are consistent, we apply the same annotation to all frames within that window. However, if there is an inconsistency, we perform frame-by-frame annotation for the 5 frames. Combining these two optimization methods significantly reduces the computation time required for annotation.

D. Mechanism Analysis

To explore the mechanism of Best-Feature-Aware (BFA), we further provide an in-depth mechanism analysis on the strategy to explain why it works for fine-grained manipulation tasks. Considering that the introduced score network is essentially a classification network, we conduct the attention visualization on three views using Grad-CAM [30]. The core of CAM is that its attention region tend to focus on the target object with the highest classification score.

As shown in Fig. 2 (b), the visualization of three views shows that the attention of score network is shifted along the temporal axis. For the T timestep when the gripper accesses the manipulated objects, the attention focuses mainly on the gripper and objects from the top view. For the $T+3$ timestep when the left arm operates on the bag, the attention shift from the top view to the left wrist view. Finally, the attention focuses on the gripper of right arm when right arm unzips the bag. The predicted importance score also indicates the best feature transfer (red line). Furthermore, Fig. 3 shows that our method exhibits stronger responses in Regions-of-Interest (RoI) [31], such as the gripper and target objects, whereas the baseline method displays more dispersed attention. The BFA strategy achieves best-view selection and transfer by predicting the importance score.

IV. EXPERIMENTS

A. Robot Setup and Tasks details

In real world, we build five dexterous manipulation tasks with *ALOHA* which is an open-source Cobot Magic platform. As shown in Fig. 5 , among the 5 challenging tasks, two require dual-arm coordination (***Fold Towel***, ***slide Bag***) while three involve single-arm manipulation (***Play Chess***, ***Open Box***, ***Close Box***). This platform includes four AgileX robotic arms and four Orbbec DaBai RGB cameras mounted on the Tracer chassis. The camera mounted on the top reveals the global view, while each arm is equipped with a wrist camera, as described in the Fig. 1(a). All cameras capture images at 30Hz frequency. At each timestep, the robotic system captured frames from the cameras, each delivering 640×480 pixels RGB images.

B. Implementation Details

For all five real-world tasks, we collected demonstrations with each episode requires 300 to 650 timesteps to perform a complex task, given the control frequency of 25 Hz. We record 50 to 500 episodes for each task. We recorded the average success rate on the fine-grained manipulation tasks. For each task, we conducted ten trials using original ACT policy and RDT policy, respectively.

Both networks were trained from scratch with random initialization, without pre-training or fine-tuning. The ACT model integrated with BFA method contains approximately 106M parameters and is trained independently for each task. Training completes in approximately 2 hours on a single NVIDIA RTX 4090 GPU, achieving an inference time of 0.015 seconds on an RTX 4060 GPU (8GB). The RDT model with BFA method, comprising around 170M parameters, is trained from scratch per task on NVIDIA A100 GPUs. The training process takes 3 hours using 8 A100 GPUs (80G) in parallel, or 1 hour on a single A100 GPU, with an inference time of approximately 0.73 seconds on an RTX 4060 GPU.

We implement ACT with the same action chunking size as to ACT-BFA, which is 24. RDT and RDT-BFA’s chunk size are 64.

C. Experiment Results

As shown in Tab. II, We integrated our modules into two state-of-the-art imitation learning methods ACT [1] and RDT [4] and conducted comparative experiments. ACT-BFA and RDT-BFA consistently outperform the baseline models in success rates at every stage across the five tasks. Overall, both models show significantly higher success rates, with ACT-BFA achieving a 46% improvement and RDT-BFA showing a 22% increase compared to their respective baseline models.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

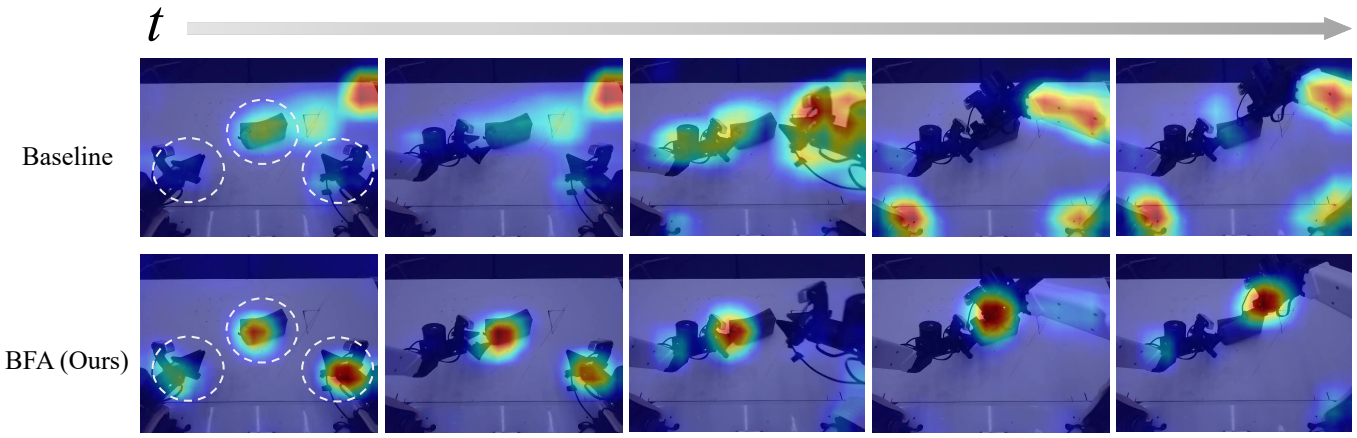


Fig. 3. Grad-Cam Heatmap Comparision between baseline and BFA. Our BFA affords more attention on end-effector and target object edge across the wall-time, while baseline distracts in the manipulation process.

TABLE II
SUCCESS RATE COMPARISON OF DIFFERENT METHODS ACROSS MULTIPLE FINE-GRAINED MANIPULATION TASKS

Method	Average Suc.	Unzip bag			Play Chess		Open box		Close box		Fold towel	
		Grab	Pinch	Slide	Pick	Position	Align	Lift	Align	Flip	Grab	Fold
ACT	32%	**100%**	60%	30%	40%	20%	50%	30%	20%	30%	50%	50%
ACT-BFA (Ours)	**78%**	**100%**	**90%**	**70%**	**100%**	**80%**	**100%**	**90%**	**70%**	**80%**	**80%**	**70%**
RDT	20%	50%	30%	20%	10%	0%	50%	40%	20%	20%	30%	20%
RDT-BFA (Ours)	**42%**	**80%**	**80%**	**30%**	**50%**	**0%**	**70%**	**70%**	**70%**	**40%**	**80%**	**70%**

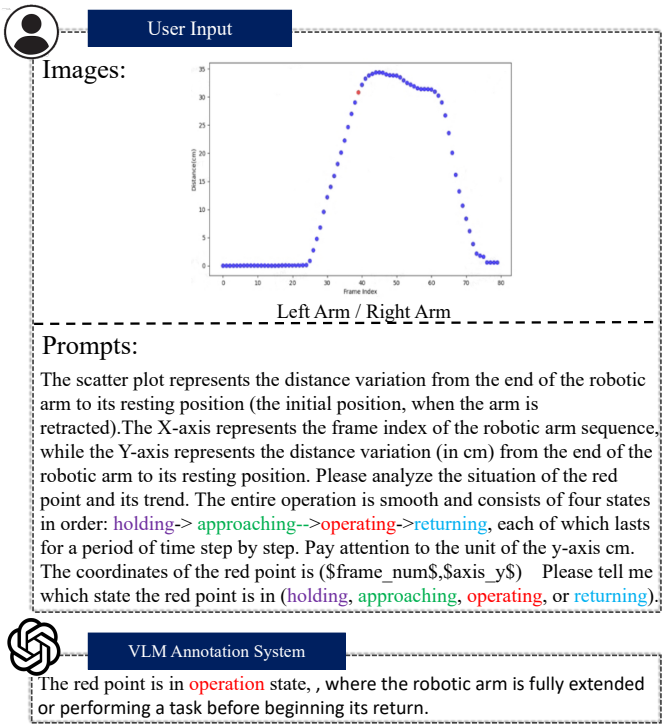


Fig. 4. System Prompt of VLM Scoring System. Given the plotted images and system prompt, the VLM system outputs current state of the robotic arm.

Overall, the low success rate of the baseline models is primarily due to the inability to locate positions accurately and precisely control the gripping. This is because, when performing fine-grained manipulation, multiple views introduce excessive redundant information, which makes it challenging for baseline methods [1], [4] to accurately and generally learn the mapping pattern from visual features to future actions. To validate this reasoning, we analyzed the failure cases of ACT [1] and RDT [4]. Both methods exhibit similar failure patterns when generalization requirements are introduced, such as varying box positions and random grasp points on deformable objects. In these cases, both policies show significant deviations in estimating critical grasp points, which ultimately hinder progression to subsequent stages.

Moreover, the two baselines differ in the types of failures. ACT [1] performs well in basic tasks but fails during fine manipulation due to insufficient trajectory refinement. For example, in the *Unzip bag* task, after securing the bag, the right gripper often misaligns with the zipper, leading to failure. On the other hand, RDT [4] experiences minor yet persistent trajectory errors across all stages, causing issues like bimanual coordination failures. For instance, in the *Fold towel* task, one gripper may lift the towel's corner while the other fails to grasp it properly, yet continues the folding motion. All these observed failure behavior aligns with our hypothesis regarding the limitation of baseline methods.

In contrast to these issues, our method effectively resolves the challenges through view selection and feature fusion, and can accurately identify the gripper point even in generalized settings, while autonomously correcting errors during the manipulation process, as demonstrated in Tab. II.

D. Ablation Study

In this section, we present an ablation study on ACT-BFA to investigate the impact of various design choices for fusing and the number of views during the manipulation. We focus on the fine-grained manipulation task of "Unzip Bag" and "Open box" as the experimental setting for our ablation analysis.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

To reveal the importance of continuously adjusting the fusion weights during manipulation process, we employed four different fusion strategies which named "Mean", "Reweight Concat", "Best Feature" and "w/o Score Loss" as shown in the Tab. III. For "Mean", We simply average the visual features from multiple perspectives as defined by $\hat{f} = \frac{\sum_i^N f_i}{N}$.

For "Reweight Concat", we multiply the scores obtained from the scoring network with the visual features, then directly concatenate them and pass the result to the policy network as defined by $\hat{f} = [f_1 \times s_1, ..., f_N \times s_N]$. For "Best Feature", we use the scores from the scoring network to select the visual feature with the highest score, which is then passed to the policy network following $\hat{f} = f_{argmax(s_1,...,s_N)}$. For "w/o Score Loss," we rely solely on policy loss for supervision, enabling the model to learn an importance score without the need for human-provided ground truth.

These four methods represent other commonly used fusion approaches. The results is presented in Tab. III. Our fusion strategy demonstrated significantly higher success rates on the unzip task compared to other methods, which proves the effectiveness of our fusion strategy.The success rate of the 'Mean' fusion method and the baseline remains consistent in the end.

TABLE III
ABLATION STUDY ON FEATURE FUSION.

Fusion Method	Unzip Bag			Open Box	
	Grab	Pinch	Slide	Align	Lift
Mean	90%	60%	30%	50%	30%
Best Feature	**100%**	80%	40%	80%	60%
Reweight Concat	80%	60%	40%	50%	40%
w/o Score Loss	70%	40%	20%	80%	70%
Ours	**100%**	**90%**	**70%**	**100%**	**90%**

Moreover, it is worth noting that the latter three methods which utilize the important score ultimately outperform the baseline, indicating that the score network effectively guides the model to better learn from the human demonstrations. Since the 'Best Feature' approach directly selects the view with the highest importance score, it results in a significant loss of information, which negatively impacts both generalization and policy performance. Moreover, 'Best Feature"s discrete feature weight selection results in large fluctuations in unseen scenarios, whereas our method's continuous weight adaptation ensures smoother, more stable decision-making. Moreover, 'Reweight Concat' approach fails to effectively fuse features as our method does. It merely applies weighted multiplication without true integration, offering only minimal improvement over the baseline. Our approach, however, leverages prior knowledge by pre-setting signal-to-noise ratios for each view, enabling more efficient weighting and better utilization of available information. In comparison, Reweight Concat shifts the fusion responsibility to the policy itself, which can be less effective. Finally, the "w/o score loss" method allows the model to learn the importance score independently, and it is much less effective than our approach. The important score remains fixed across manipulation stages, such as $[0.4, 0.2, 0.4]$ for unzip bag and $[0.4, 0.6]$ for open box. This approach is essentially a simple weighted feature fusion with fixed weights, lacking our method's ability to adapt to different manipulation stages. Its score does not fluctuate violently but only varies within a narrow range; the backpropagated gradient of the action loss is not strong enough to enable it to fuse useful features effectively.

TABLE IV
ABLATION STUDY ON VIEW SELECTION.

Viewpoint	Unzip Bag			Open Box	
	Grab	Pinch	Slide	Align	Lift
Top-view	50%	20%	10%	30%	10%
Left-wrist	0%	0%	0%	-	-
Right-wrist	30%	20%	10%	30%	10%
Baseline (3-views)	**100%**	60%	30%	50%	30%
Ours (3-views)	**100%**	**90%**	**70%**	**100%**	**90%**

We then conducted an ablation study on the baseline's view selection as shown in the Tab. IV. The results indicate that using only one camera performs worse than using all three views. This suggests that simply reducing the number of views is not a viable strategy—for instance, relying solely on the top view results in limited accuracy during manipulation, and using only the wrist camera fails to achieve precise positioning at close range. Moreover, our method significantly outperforms the configuration that uses all three views, thereby demonstrating its effectiveness.

TABLE V
THE COMPARISON OF DIFFERENT ANNOTATION METHOD

Method	**Unzip bag**	**Close box**	**Play Chess**	**Open Box**	**Fold towel**	**Avg**
Human	1.00	1.00	1.00	1.00	1.00	1.00
Rule-based	0.92	0.90	0.84	0.92	0.88	0.89
VLM-based	0.99	0.98	0.97	0.99	0.98	0.98

To validate our annotation method, we compared different annotation approaches against manual human annotations. Due to cost constraints, we evaluated 10 episodes per task. We developed a rule-based method using gradients of distance from both left and right grippers. At the same time, human annotators only labeled dual-arm manipulation states, and both the manual annotations and rule-based method adhered to the same hard mapping rules outlined in Table I. Using manual annotations as the ground truth, our comparison revealed that the VLM-based approach achieves better alignment with human annotations than the rule-based method, as demonstrated in Tab. V.

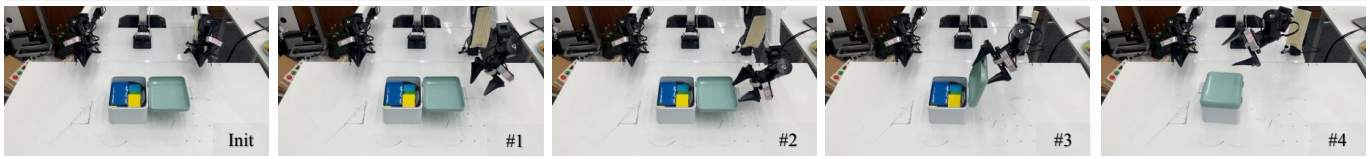
TABLE VI
COMPARISON OF SUCCESS RATE AND COMPUTATIONAL COST.

Method	Suc.(%)	FLOPs (G)	# params. (M)
ACT	32	16.34	**106.22**
ACT-BFA (Ours)	**78**	**12.96**	106.90
RDT	20	4356.99	166.23
RDT-BFA (Ours)	**42**	**3805.66**	**162.50**

To further demonstrate the effectiveness of our BFA strategy, we report the overall performance, the computational cost and the parameters (see Tab. VI). It shows our BFA only adds marginal parameters, while greatly reducing the computational cost and improving the success rate. The BFA reduces the

Play Chess : The robot need to Going to Pick (#1), Picking precisely (#2), Moing to target (#3), Targeting (#4). The last image shows it succeeds when the chess piece land on the red-cross position.

Open box : The robot need to Going to Align (#1), Aligning accurately (#2), Lifting (#3), Lifted (#4). The images show the task requires precise edge alignment for lid opening while maintaining robustness under box displacement.



Close box : The robot need to Going to Align (#1), Aligning accurately (#2), Fliping (#3), Fliped (#4). The images show that the task requires precise edge alignment for fliping while maintaining robustness under box displacement.

Fold towel : The robot need to Going to Grab (#1), Grabing accurately (#2), Folding (#3), Folded (#4). The task emphasizes precise gripper placement below towel and synchronized dual-arm control for deformable object.

Unzip bag : The robot need to Going to Grab (#1), Grabing accurately (#2), Moving to Pinch (#3), Pinched and Sliding (#4), Slided (#5). The task demands dual-arm balance to accurately identify and manipulate a 2.5cm zipper on a randomly positioned pencil bag.

Fig. 5. **Visualization of robotic manipulation sequences for five tasks**: playing chess, opening a box, closing a box, folding a towel and unzipping a bag. Each sequence highlights the robot's precise control in alignment, grasping, flipping, and dual-arm coordination, demonstrating the effectiveness of our method in fine-grained manipulation tasks.

information redundancy of multi-view images, providing the informative visual features for policy networks.

V. CONCLUSIONS

In this paper, we propose the Best-Feature-Aware fusion strategy for multi-view fine-grained manipulation. Such strategy achieves the dynamic view fusion during manipulation. It greatly reduces the visual information redundancy and computational costs while significantly improving the success rate of complex fine-grained manipulation tasks. To implement the strategy, we further introduce the VLM-based scoring system to generate the multi-view ground-truth of importance score, as the supervision of the introduced light-weight scoring network. The proposed BFA strategy provides 22%-46% improvements on fine-grained manipulation tasks. In the future, we will extend our method to VLA (Vision-Language-Action) based approaches [27], [32]–[35]. Since most tokens of VLA in these method are image tokens, we anticipate our method's effectiveness and its potential to significantly reduce large computational resources.

REFERENCES

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, Eds., 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.016

[2] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 4565–4573.

[3] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: a survey of learning methods." *ACM computing surveys*, vol. 50, 2017, cOMPLETED – Issue in progress, but article details complete; not marked as pending following upload 10.05.2017 GB – Now on ACM website, issue still in progress 9/5/2017 LM – Not on journal website 24/2/2017 LM – Info from contact 10/2/2017 LM ADDITIONAL INFORMATION: Elyan, Eyad – Panel B. [Online]. Available: http://hdl.handle.net/10059/2298

[4] S. Liu *et al.*, "RDT-1B: a diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[5] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y. Chao, and D. Fox, "RVT: robotic view transformer for 3d object manipulation," in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*,

ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 2023, pp. 694–710.

[6] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y. Chao, and D. Fox, "RVT-2: learning precise manipulation from few demonstrations," in *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. [Online]. Available: https://doi.org/10.15607/RSS.2024.XX.055

[7] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, Eds., 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.026

[8] R. Zeng, Y. Wen, W. Zhao, and Y. Liu, "View planning in robot active vision: A survey of systems, algorithms, and applications," vol. 6, no. 3, 2020, pp. 225–245.

[9] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments," vol. 30, no. 11. SAGE Publications Sage UK: London, England, 2011, pp. 1343–1377.

[10] E. Arruda, J. L. Wyatt, and M. S. Kopicki, "Active vision for dexterous grasping of novel objects," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 2881–2888.

[11] J. Lv, Y. Feng, C. Zhang, S. Zhao, L. Shao, and C. Lu, "SAM-RL: sensing-aware model-based reinforcement learning via differentiable physics-based simulation and rendering," in *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, Eds., 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.040

[12] I. Chuang, A. Lee, D. Gao, M.-M. Naddaf-Sh, and I. Soltani, "Active vision might be all you need: Exploring active vision in bimanual robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2409.17435

[13] M. Breyer, L. Ott, R. Siegwart, and J. J. Chung, "Closed-Loop Next-Best-View Planning for Target-Driven Grasping," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 1411–1416.

[14] G. Wang, H. Li, S. Zhang, D. Guo, Y. Liu, and H. Liu, "Observe then act: Asynchronous active vision-action model for robotic manipulation," *IEEE Robotics Autom. Lett.*, vol. 10, no. 4, pp. 3422–3429, 2025.

[15] R. Cheng, A. Agarwal, and K. Fragkiadaki, "Reinforcement learning of active vision for manipulating objects under occlusions," in *Conference on Robot Learning*. PMLR, 2018, pp. 422–431.

[16] Jinghuan Shang and Michael S. Ryoo, " Reinforcement Learning under Limited Visual Observability," in *Advances in Neural Information Processing Systems*, 2023.

[17] X. Chen *et al.*, "Transferable active grasping and real embodied dataset," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3611–3618.

[18] Y. Zaky, G. Paruthi, B. Tripp, and J. Bergstra, "Active perception and representation for robotic manipulation," *arXiv preprint arXiv:2003.06734*, 2020.

[19] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[20] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, Eds., 2024. [Online]. Available: https://doi.org/10.15607/RSS.2024.XX.067

[21] T. Z. Zhao *et al.*, "ALOHA unleashed: A simple recipe for robot dexterity," in *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 2024, pp. 1910–1924.

[22] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4788–4795.

[23] X. Ma, S. Patidar, I. Haughton, and S. James, "Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 081–18 090.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[25] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[26] A. Hurst *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[27] K. Black *et al.*, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[28] W. Mao *et al.*, "Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world," 2025. [Online]. Available: https://arxiv.org/abs/2412.00171

[29] D. Ghosh *et al.*, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, Eds., 2024. [Online]. Available: https://doi.org/10.15607/RSS.2024.XX.090

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[31] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99.

[32] M. J. Kim *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 2024, pp. 2679–2713.

[33] A. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and RT-X models : Open x-embodiment collaboration," in *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 6892–6903.

[34] J. Wen *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2409.12514

[35] R. Zheng *et al.*, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.