# Beyond Label Semantics: Language-Guided Action Anatomy for Few-shot Action Recognition

Supplementary Material

## 7. Prompt Design

In our approach, to exploit the rich prior knowledge embedded in semantic space, we decompose each action label into an ordered sequence of atomic action descriptions with Visual Anatomy Module. Specifically, we leverage a large language model (GPT-4o) to transform action labels into atomic descriptions enriched with spatiotemporal context, as described in Sec. 3.2. To ensure accurate parsing, we carefully design the prompt and extend it with additional input-output templates. The complete prompt as following:

**Prompt:** *Deduce the scene description and three sub-action descriptions from an action label. The scene description should include possible scene elements, such as humans, objects, and background. The scene description must consist of visible elements, not abstract descriptions like atmosphere, mood or social setting. The sub-action descriptions must follow strict temporal order, focusing on the posture of the people involved, relevant elements in the scene, and potential interactive objects. Ignore object textures and dismiss any unlikely or invalid sub-actions, as well as unnecessary emotional descriptions. Keep the sub-action descriptions brief and clear and avoid the abstract descriptions such as enjoying the performance. Provide a concise answer for both the scene description and the three sub-action descriptions, following the example below:*

*Example: Input: jumping into pool. Output: "Action Label": "Jumping into poo", "sub-action description": ["A photo of a person stands at the edge of a pool, preparing to jump in.", "A photo of a person leaps off the edge, mid-air over the pool.", "A photo of a person enters the water, creating a splash as they dive in."] Your analysis should be thorough and accurate, considering all relevant aspects of the action to support your deductions effectively. Once I provide the action label, please deduce the scene description and three sub-action descriptions accordingly.*

The output examples are shown in Fig. 5. The atomic action descriptions generated by the LLM effectively characterize the actions across different datasets.

## 8. Implementation Details of Experimental

### 8.1. Network Parameters

The hyperparameters of our methods in each dataset are shown in Tab. 7. In this table, 'lr' means the learning tate, 'st_iter' indicates the number of iteration per step, 'steps' refers to the number of steps to change the learning rate when using the multistep scheduler. $\mathbf{M}$ and $\alpha$ means the attention mask weight in Fine-grained Multimdoal Fusion Module and the visual weight in Multimodal Matching Module, respectively.

Table 7. The settings of hyperparameters in each dataset.

| Dataset | lr | st_iter | steps | warm_lr | $\alpha$ |
|---|---|---|---|---|---|
| HMDB51 | 5e-6 | 700 | [0, 4, 6] | 1e-6 | 0.0250 |
| Kinetics | 1e-5 | 250 | [0, 6, 9] | 5e-6 | 0.0625 |
| UCF101 | 2e-6 | 600 | [0, 4, 6] | 1e-7 | 0.1125 |
| SSv2-Small | 1e-5 | 2000 | [0, 4, 6] | 1e-6 | 0.2 |
| SSv2-Full | 2e-6 | 600 | [0, 4, 6] | 1e-7 | 0.2 |

### 8.2. Effect of the atomic action number on the generated descriptions

In Sec. 4.3, we have explored the impact of the number of atomic actions on the model. In this section, we further analyze how the number of atomic actions affects the generated atomic action descriptions.

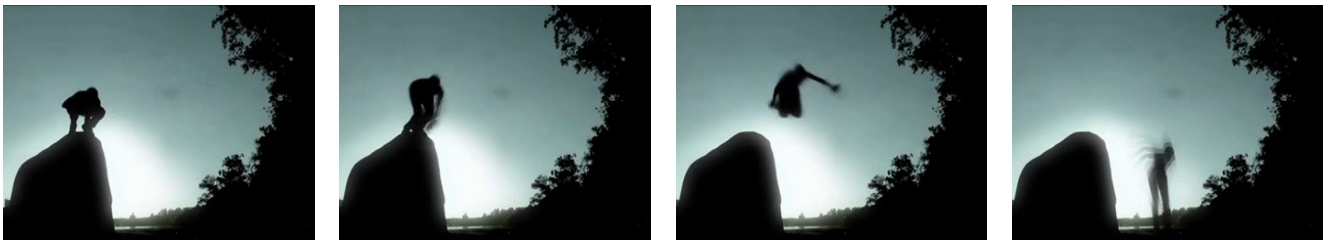
As shown in the Fig. 6, when the number of atomic actions is set to 3, the generated atomic action descriptions correspond well with the action samples. However, when the number is increased to 4, redundant atomic action descriptions are often generated. This redundancy can lead to misalignment between visual and textual features, which degrades model performance, as shown in the Tab. 6.

### 8.3. Prompt Design of VLM Experiments

In Sec. 4.3, we have explored the performance of pre-trained VLM models on FSAR tasks. Specifically, for LLaVA-1.5 and Qwen-2.5, we prompted the models with: *"Below is a sequence of images showing an action. What action is being performed?"* to guide them in describing the query action. Finally, we classify the query action by measuring the distance between the semantic features of the description generated by the VLM and the labels of the support set.

Furthermore, for LLaVA-1.5† and Qwen-2.5†, we incorporate support labels directly into the prompt to guide the VLM in classifying actions with support set labels. Specifically, we use the following prompt: *"Classify the action in the image sequence. Choose from: {Action label1}, {Action label2}, {Action label3}, {Action label4}, {Action label5}. Only output the action name from the list."* In this way, the VLM can classify actions within a limited set of labels, effectively mitigating the potential hallucinations associated with generating action descriptions.

**"Arm wrestling" from Kinetics**
Two people sitting at a table, gripping each other's hand, preparing to arm wrestle.
Two people intensely arm wrestling, with one person's arm starting to bend backward.
The conclusion of an arm wrestling match, with one person's arm pinned down to the table while the other raises their arm in victory.



**"Jump" from HMDB51**
A person bending their knees, preparing to jump off the ground.
The person in mid-air, arms raised and legs extended, as they jump.
The person landing back on the ground, knees slightly bent to absorb the impact.


**"Bowling" from UCF101**
A person stands at the start of the lane, holding a bowling ball and preparing to bowl.
The person steps forward, swinging the ball back before releasing it down the lane.
The ball rolls down the lane towards the pins, with the person watching as it approaches the target.

**"Pushing [something] so that it falls off the table" from SSv2-Small**
A person with their hands placed on an object on a table, preparing to push it.
The person pushing the object towards the edge of the table.
The object falling off the table as the person continues to push.

Figure 5. Atomic action descriptions generated from different datasets along with their corresponding action samples.

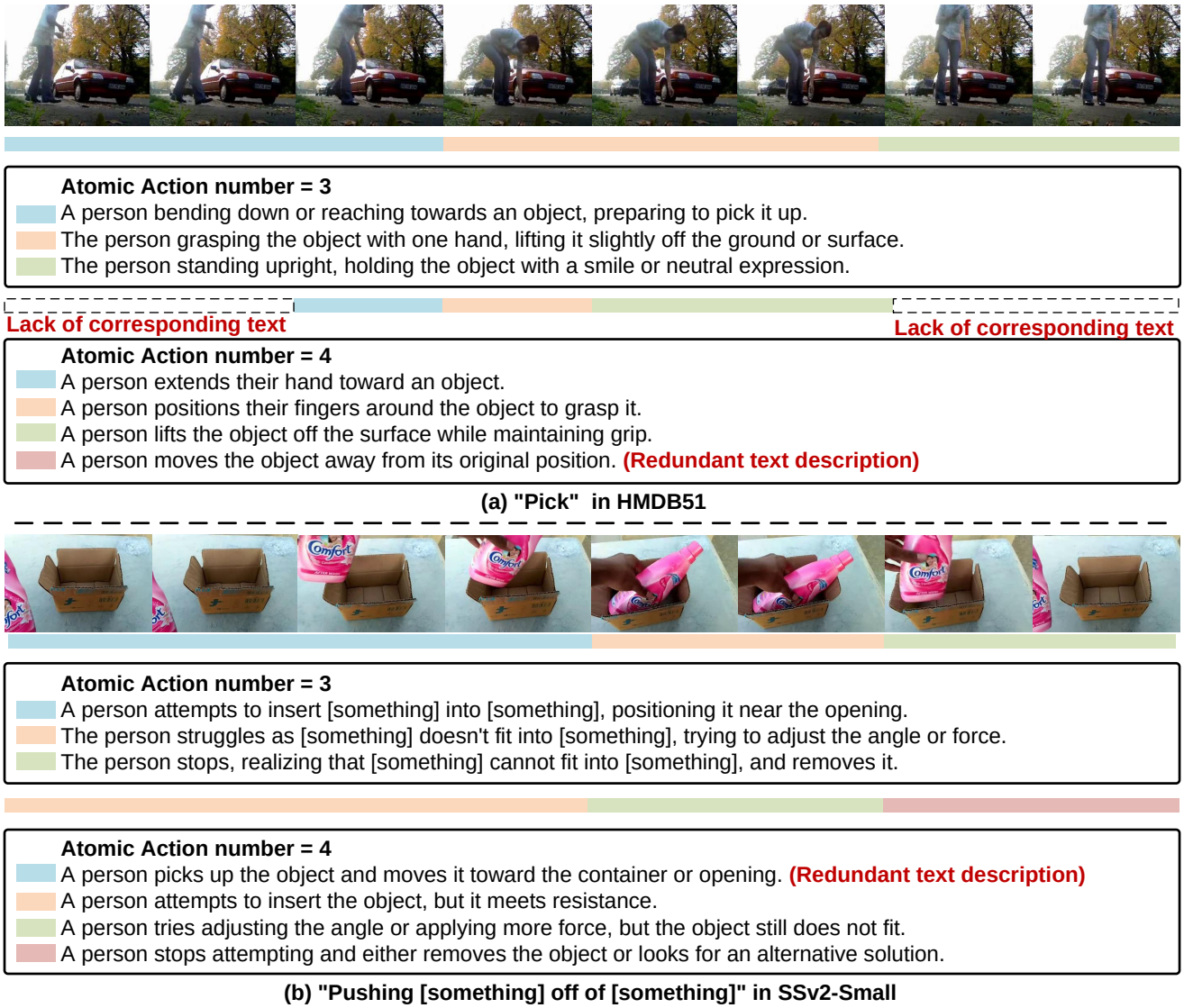



Figure 6. Atomic action descriptions generated with different numbers of atomic actions. (a) The action 'pick' from HMDB51. (b) The action 'Pushing [something] off of [something]' from SSv2-Small.

# 9. Additional Experiment Results

## 9.1. Visualization of LGA

To further evaluate LGA's ability to perceive the different temporal phases of actions, we visualize how each frame contributes to the overall distance between the video and the action prototype in HMDB51 [17]. As shown in Fig. 7, compared to baseline methods, key frames from distinct temporal phases contribute more significantly to prototype matching. For example, the frame depicting "lowering the body" in the "Push up" sample contributes more significantly than baseline methods do. This confirms that LGA selectively focuses on critical action phases rather than treating all frames uniformly.

## 9.2. Analysis of the number of support set samples

In this section, we further evaluate the performance of our method under different shot settings. As shown in Tab. 8, our method achieves greater improvements over the baseline method (CLIP-FSAR [42]) in lower-shot scenarios; specifically, it outperforms the baseline by 9.7% in the 1-shot setting compared to 1.6% in the 5-shot setting. We attribute this to the fact that enriched semantic cues are particularly effective when visual information is limited.

## 9.3. Analysis of the number of input video frames

To ensure a fair comparison with previous methods [3, 14, 40, 43], we uniformly sample 8 frames from each video to

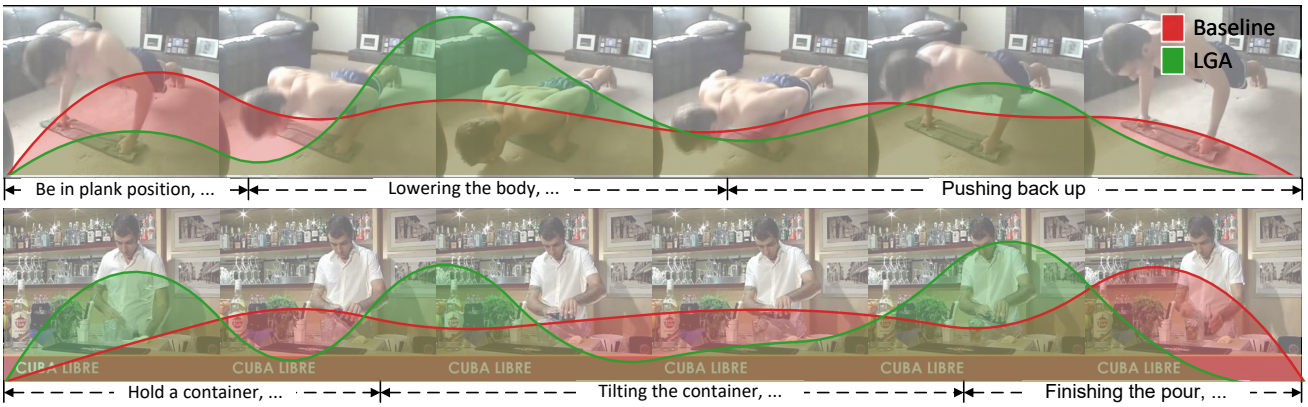



Figure 7. Temporal attention visualization of our LGA on HMDB51 [17].

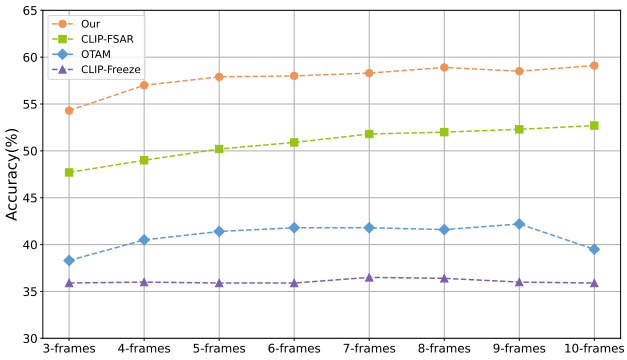



Figure 8. Performance comparison with different numbers of input video frames under 5-way 1-shot setting on SSv2-Small datasets.

Table 8. Comparison with recent state-of-the-art few-shot action recognition methods on the HMDB51 dataset with 5-way k-shot setting. The best results are **bolded** in black, and the underline represents the second best result.

| Method | Pre-training | HMDB51 | | |
|---|---|---|---|---|
| | | 1-shot | 3-shot | 5-shot |
| OTAM [3] | CLIP-ViT-B | 72.5 | 81.6 | 83.9 |
| CLIP-Freeze [30] | CLIP-ViT-B | 58.2 | 72.7 | 77.0 |
| CapFSAR [44] | BLIP-ViT-B | 65.2 | - | 78.6 |
| CLIP-FSAR [42] | CLIP-ViT-B | 77.1 | 84.1 | 87.7 |
| MA-FSAR [51] | CLIP-ViT-B | 83.4 | - | 87.9 |
| Task-Adapter [2] | CLIP-ViT-B | 83.6 | - | 88.8 |
| EMP-Net [47] | CLIP-ViT-B | 76.8 | - | 85.8 |
| Kumar*et al.* [18] | DINOv2 | 60.0 | 71.8 | 77.0 |
| CLIP-CPM$^2$C [10] | CLIP-ViT-B | 75.9 | - | 88.0 |
| TSAM [20] | CLIP-ViT-B | 84.5 | - | 88.9 |
| Our | CLIP-ViT-B | **86.8** | **89.2** | **89.3** |

construct its visual representation in our experiments. In addition, to comprehensively analyze the influence of the number of frames on the model performance, we perform an ablation study on the HMDB51 and SSv2-Small datasets by varying the number of input frames. As shown in Fig. 8, the performance starts to increase and gradually saturates as the number of input frames increases. Remarkably, our method outperforms previous methods across different frame number settings, demonstrating its effectiveness and robustness.

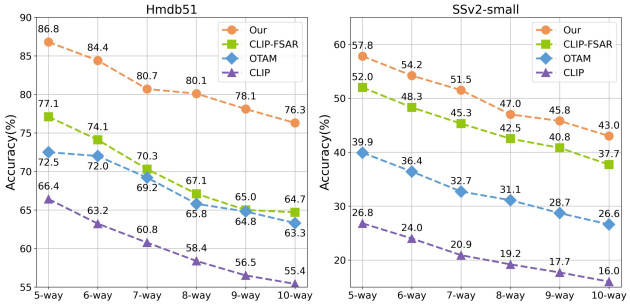



Figure 9. $N$-way 1-shot results of our method and other baseline methods with $N$ varying from 5 to 10.

## 9.4. Analysis of N-way classification

We also investigate the effect of varying $N$ on the few-shot performance. We further perform ablation experiments to evaluate the $N$-way 1-shot accuracy, where $N$ ranges from 5 to 10. As shown in Fig. 9, we can observe that as $N$ increases, the classification becomes more challenging, resulting in a noticeable drop in performance. Specifically, on the SSv2-Small dataset, our method experiences a 14.8% drop in accuracy when moving from the 5-way 1-shot setting to the 10-way 1-shot setting. Despite this performance drop, our approach consistently outperforms all baseline methods in all evaluated settings, highlighting the robustness and superiority of our method.