# VideoGEM: Training-free Action Grounding in Videos

## Supplementary Material

The supplementary material is organized as follows: first, we show a comparison between GEM and our proposed weighting mechanism in Section 7. Then, we extend our evaluation to include the BLIP backbone [12] in Section 8. Next, we present additional experiments on extending image models to video data in Section 9 and offer a more detailed analysis of the effects of static weights, dynamic weights, and their combination in Section 10, and finally, we provide qualitative evaluation in Section 14.

## 7. Weighted GEM vs. GEM

VideoGEM contains two new concepts that we introduced in the main paper: prompt decomposition and a weighting mechanism for GEM. While prompt decomposition has already been illustrated in Figures 1 and 2, we also want to focus on the difference between GEM and our proposed weighting mechanism for GEM (weighted GEM) in Figure 5. The proposed weighting mechanism is illustrated on the left. Static weights and dynamic weights can be applied independent of each other. While static weights can be set heuristically or via hyperparameter search based on the general pipeline performance, dynamic weights are adapting to the importance of the different transformer layers individually and with respect to each prompt as described in Section 3.3. Standard GEM does not use any weights, which equals to always using $1-$weights in our formulation. This results in one weight for the initial self attention output that is the first input into GEM as well as one weight for the output of each self-self attention block. A layer output is multiplied by its corresponding weight, producing the final output as the sum of weighted outputs.

## 8. Additional backbone

We extend the evaluation of our VideoGEM to another backbone. Namely, we consider the image-text BLIP [12] model finetuned on the instructional video-text HowToCaption dataset [25] (Table 6). The HowToCaption dataset is based on the HowTo100M [20] dataset and provides captions for instructional videos. We apply both, the original GEM as well as the proposed VideoGEM to this backbone and evaluate it on on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets. Table 6 shows that also with BLIP as a backbone, VideoGEM outperforms GEM by more than $10\%$ on average, outperforming GEM individually on each dataset. Since our proposed prompt decomposition method relies on good predictions for the individual verb, object, and action prompts we assume that the proposed method benefits more from a finetuned backbone with improved individual predictions via prompt decompostion than the original GEM . Moreover, Tables 8 to 10 show that BLIP also largely benefits from dynamic and static weights.

| Setting | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|
| GEM | 67.79 | 69.00 | 34.77 | 37.97 | 52.38 |
| VideoGEM | **77.20** | **72.04** | **51.57** | **53.83** | **63.66** |

Table 6. **Experimental evaluation using the BLIP backbone fine-tuned on the video-text HowToCaption dataset.** We report the model's performance on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets.

| Model | Set | Data | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|---|---|
| OpenCLIP | base | img | **68.28** | **74.05** | **56.87** | **32.91** | **58.03** |
| | | vid | 65.20 | 70.68 | 48.67 | 28.15 | 53.18 |
| | ours | img | **76.42** | **80.32** | **60.05** | **45.33** | **65.53** |
| | | vid | 74.26 | 76.55 | 52.93 | 42.10 | 61.46 |
| CLIP | base | img | **67.79** | **78.52** | **50.08** | **36.92** | **58.33** |
| | | vid | 67.55 | 76.31 | 46.10 | 34.82 | 56.20 |
| | ours | img | **76.90** | **84.53** | **52.57** | **47.46** | **65.37** |
| | | vid | 75.15 | 82.68 | 50.36 | 45.79 | 63.50 |

Table 7. **Experimental evaluation of extending image-text backbones to video input.** CLIP and OpenCLIP are extended to video input in a training-free manner. Their accuracy with image and video data as input is compared for VideoGEM (ours) and standard GEM (base) on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets.

## 9. Video input for image-language models

Since image-language models, such as CLIP and OpenCLIP, are limited to processing single images as input, we conduct additional experiments where we extend these models to handle video input.

**Setup.** Let $F = \{f_1, ..., f_T\}$ be a video with $T$ frames. For an input frame $f_i$, the first layer $l_0$ of a backbone encodes the input image into $N$ patch tokens, before they are jointly processed through the transformer layers. In order to utilize video data we apply the first layer $l_0$ individually to all frames $f_i, 1 \leq i \leq T$ resulting in $T * N$ patch tokens. Since the transformer layers are independent of the number of patch tokens, the $T * N$ patch tokens are processed by the transformer layers without any architectural change. For a given input sequence, the patch tokens from all frames are concatenated into a single sequence $X \in \mathcal{R}^{(T*N)\times d}$,

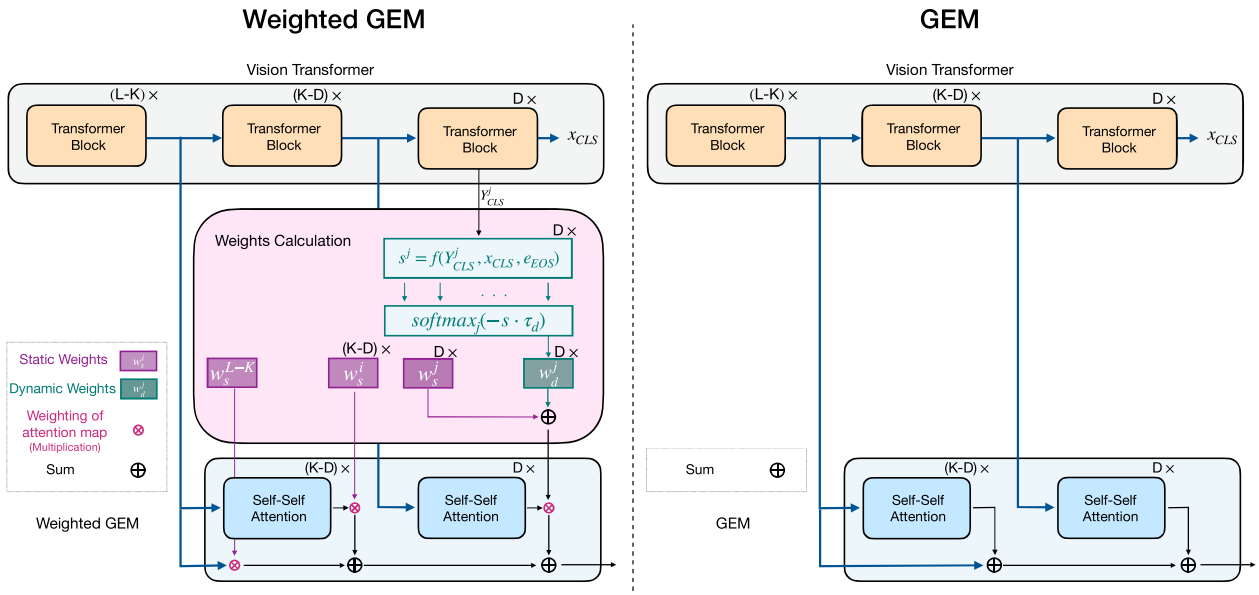


Figure 5. **Comparison of GEM and our proposed weighting mechanism.** The proposed weighting mechanism is illustrated on the left. Static weights and dynamic weights can be applied independent of each other. While static weights can be set heuristically or via hyperparameter search based on the general pipeline perfromance, dynamic weights are adapting to the importance of the different transformer layers individually and with respect to each prompt as described in Section 3.3. Standard GEM does not use any weights, which equals to always using $1-$weights in our formulation.

where $d$ is the embedding dimension and further processed by GEM as described in Section 3.2.

**Results.** We evaluate the performance of GEM and VideoGEM with CLIP and OpenCLIP as backbones for video input with $T = 8$ frames compared to image input in Table 7. The frame sampling for video input is the same as for ViCLIP in previous experiments according to Section 4.2. Independent of the backbone or the setting (GEM, or VideoGEM) video input decreases the accuracy compared to image input. We attribute this to the fact, that the backbones are not trained to handle video input and therefore are not capable of utilizing the spatial and temporal relationships that are inserted into the input data by using videos, but are distorted by the different input instead. This assumption is supported by the results for ViCLIP in Table 5. ViCLIP has a similar architecture to CLIP besides its first embedding layer which embeds a video input into patch tokens. ViCLIP is specifically trained with video input and its accuracy benefits from using videos as input as shown in Table 5, suggesting that video-specific pretraining is needed to leverage the spatial and temporal relationships in videos.

## 10. Ablation for static and dynamic weights

In this section, we extend our analysis of the effects of static, dynamic, and combined weights (elaborating on the experimental evaluation in Section 4.4). We investigate their effects separately for verb prompts (Table 8), object prompts (Table 9), and action prompts (Table 10). Note, that compared to Table 2 in the main paper, Tables 8 to 10 don't apply prompt decomposition but only the weighting strategies. We include ViCLIP on video and image data, as well as CLIP, OpenCLIP, and BLIP which is finetuned on HowToCaption, as backbones. First, we observe that independent of the prompt (verb, object, or action) averaged over datasets and backbones, dynamic weights and static weights improve over using no weights. The combination of static and dynamic weights further improves over both, static weights or dynamic weights on their own, showing the capabilities of static and dynamic weights as well as their additive effects. Moreover, OpenCLIP and BLIP improve around $0.5 - 1\%$ with dynamic weights across different prompts, while the accuracy of ViCLIP and CLIP only shows minor changes. This confirms the assumption that the benefit of dynamic weights highly depends on the backbone.

| Model | Data | Set | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|---|---|
| ViCLIP | vid | base | 63.33 | 75.62 | 33.84 | 29.40 | 50.55 |
| | | dyn | 63.27 | 75.82 | 34.00 | 29.62 | 50.68 |
| | | stat | 64.66 | 76.20 | 36.17 | 30.96 | 52.00 |
| | | s+d | **64.72** | **76.21** | **36.54** | **31.26** | **52.18** |
| ViCLIP | img | base | 64.17 | 75.87 | 32.64 | 28.40 | 50.27 |
| | | dyn | 64.17 | 76.03 | 32.60 | 28.62 | 50.36 |
| | | stat | **65.44** | **76.35** | 34.16 | 29.67 | 51.41 |
| | | s+d | 65.38 | 76.32 | **34.24** | **30.00** | **51.49** |
| CLIP | img | base | 69.90 | 79.71 | 29.30 | 23.29 | 50.55 |
| | | dyn | 70.14 | 79.79 | 29.14 | 22.57 | 50.41 |
| | | stat | 72.32 | 82.06 | **30.43** | **26.44** | **52.81** |
| | | s+d | **72.80** | **82.48** | 29.74 | 25.69 | 52.68 |
| OpenCLIP | img | base | **70.87** | 78.65 | 41.16 | 17.75 | 52.11 |
| | | dyn | 70.57 | 78.42 | 41.72 | 18.43 | 52.29 |
| | | stat | 69.36 | **80.69** | 43.09 | 20.59 | 53.43 |
| | | s+d | 69.60 | 80.24 | **43.45** | **21.60** | **53.72** |
| BLIP | img | base | 63.63 | 64.75 | 33.00 | 31.18 | 48.17 |
| | | dyn | 64.66 | 65.45 | 33.20 | 31.41 | 48.68 |
| | | stat | 68.15 | **68.23** | 44.86 | 40.41 | 55.41 |
| | | s+d | **69.00** | 67.91 | **45.54** | **40.49** | **55.74** |

Table 8. **Ablation study on the effect of weights for verb prompts.** Accuracy for verb prompts for vanilla GEM (base), dynamic weights for the last three layers (dyn), static weights (stat), and the proposed combination of static and dynamic weights (s+d), evaluated on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets.

| Model | Data | Set | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|---|---|
| ViCLIP | vid | base | **62.48** | 66.44 | 53.70 | 45.72 | 57.09 |
| | | dyn | 62.00 | 66.53 | 53.82 | 45.75 | 57.03 |
| | | stat | 62.24 | **68.73** | 54.86 | 46.56 | 58.10 |
| | | s+d | 62.30 | 68.71 | **54.94** | **46.78** | **58.18** |
| ViCLIP | img | base | 62.06 | 66.83 | 53.86 | 44.39 | 56.81 |
| | | dyn | **62.18** | 66.92 | 53.98 | 44.63 | 56.93 |
| | | stat | 60.92 | 68.59 | 54.14 | 45.03 | 57.17 |
| | | s+d | 61.10 | **68.62** | **54.30** | **45.17** | **57.30** |
| CLIP | img | base | 65.74 | 76.27 | 46.50 | 36.24 | 56.19 |
| | | dyn | **66.47** | 76.78 | 46.02 | 35.70 | 56.24 |
| | | stat | 65.68 | 78.70 | **47.35** | **38.06** | **57.45** |
| | | s+d | 65.56 | **79.45** | 47.19 | 37.52 | 57.43 |
| OpenCLIP | img | base | 67.55 | 76.43 | 49.68 | 28.35 | 55.50 |
| | | dyn | **68.52** | 75.67 | 50.76 | 30.91 | 56.47 |
| | | stat | 64.90 | **77.86** | 50.96 | 32.03 | 56.44 |
| | | s+d | 65.44 | 77.00 | **51.21** | **34.20** | **56.96** |
| BLIP | img | base | 59.95 | **59.85** | 32.68 | 34.17 | 46.66 |
| | | dyn | 60.62 | 59.59 | 33.76 | 35.57 | 47.39 |
| | | stat | **62.48** | 58.09 | 45.34 | 40.27 | 51.55 |
| | | s+d | 62.06 | 57.94 | **46.38** | **42.15** | **52.13** |

Table 9. **Ablation study on the effect of weights for object prompts.** Accuracy for object prompts for vanilla GEM (base), dynamic weights for the last three layers (dyn), static weights (stat), and the proposed combination of static and dynamic weights (s+d), evaluated on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets.

## 11. Ablation for weights

For prompt decomposition the same weights of $w_{verb} = 0.2, w_{obj} = 0.2, w_{act} = 0.6$ are applied in all experiments. To analyze the importance of the weighting scheme we compare 4 different weighting schemes in Table 11. One that prioritizes objects (W1), one that prioritizes verbs (W2), one that treats all three prompts equally (W3), and our original weighting scheme (org). We observe similar accuracy for the different weighting schemes, while our original weights perform the best.

## 12. Comparison to multimodal LLMs

We extend our comparison to State-of-the-art methods to multimodal large language models, namely Qwen-VL [1]. Note that Qwen is not directly comparable, as it utilizes location information during training and differs significantly in both the number of parameters and the amount of training data. We provide a comparison of Qwen-VL and VideoGEM in Table 4. We use "Generate grounding for " as a prompt template for Qwen-VL. We use a different prompt for Qwen-VL compared to VideoGEM to ensure the highest rate of bounding box predictions for Qwen-VL. The center of the predicted bounding box is the prediction for Qwen-VL. The results indicate that our training-free method VideoGEM outperforms Qwen-VL on average.

## 13. Ablation for CLIP prompting

We compare the prompt decomposition of VideoGEM with the prompting technique proposed by CLIP in Table 13. We apply the 80 prompt templates that are provided by the CLIP paper to the test labels. Two different settings are used to obtain the final prediction. In the first setting, the text embeddings are averaged (*avg. txt.*) and then the combined weights of VideoGEM are applied to the single averaged text embedding resulting in a single heatmap. In the second setting, the combined weights are applied to each of the 80 prompts individually resulting in 80 heatmaps that are averaged pointwise (*avg. heat.*). The location of the maximum logit in the heatmap is the predicted location for both settings, similar to VideoGEM. VideoGEM significantly outperforms CLIPs prompting technique. This shows, that the benefit of our proposed prompt decomposition comes not only from the majority voting but the prompt decomposition itself.

## 14. Qualitative analysis

We present qualitative examples of predictions from our VideoGEM model, using ViCLIP as the backbone with video inputs on the V-HICO and GroundingYouTube datasets in Figures 6 and 7 respectively. The ground truth

| Model | Data | Set | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|---|---|
| ViCLIP | vid | base | 65.08 | 73.75 | **53.62** | 51.28 | 60.93 |
| | | dyn | 64.84 | 73.81 | **53.62** | 51.25 | 60.88 |
| | | stat | **66.53** | **74.23** | 52.97 | 51.93 | **61.42** |
| | | s+d | 65.68 | 74.17 | 52.97 | **51.99** | 61.20 |
| ViCLIP | img | base | 65.20 | 74.00 | 52.17 | 48.80 | 60.04 |
| | | dyn | 64.66 | 74.04 | 52.37 | 48.91 | 60.00 |
| | | stat | **65.86** | **74.62** | 52.65 | 49.51 | 60.66 |
| | | s+d | 65.80 | 74.50 | **52.85** | **49.66** | **60.70** |
| CLIP | img | base | 67.79 | 78.52 | 50.08 | 36.92 | 58.33 |
| | | dyn | 67.43 | 78.47 | 49.68 | 36.83 | 58.10 |
| | | stat | 68.76 | **80.91** | **51.37** | **39.65** | **60.17** |
| | | s+d | **68.94** | 80.53 | 50.84 | 39.31 | 59.91 |
| OpenCLIP | img | base | 68.28 | 74.05 | 56.87 | 32.91 | 58.03 |
| | | dyn | **68.64** | 74.24 | 56.47 | 35.58 | 58.73 |
| | | stat | 66.10 | **75.50** | **58.68** | 36.18 | 59.12 |
| | | s+d | 66.77 | 74.87 | 57.36 | **38.38** | **59.35** |
| BLIP | img | base | 67.79 | 69.00 | 34.77 | 37.97 | 52.38 |
| | | dyn | 69.00 | 68.86 | 37.02 | 39.47 | 53.59 |
| | | stat | 69.30 | **71.86** | 46.26 | 45.52 | 58.24 |
| | | s+d | **70.02** | 70.68 | **47.99** | **46.53** | **58.81** |

Table 10. **Ablation study on the effect of weights for action prompts.** Accuracy for action prompts with vanilla GEM (base), dynamic weights for the last three layers (dyn), static weights (stat), and the proposed combination of static and dynamic weights (s+d), evaluated on the V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT) datasets.

| Setting | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|
| W1 | 73.52 | 77.27 | **56.55** | 56.43 | 65.94 |
| W2 | 75.57 | 78.38 | 52.37 | 54.99 | 65.33 |
| W3 | **77.80** | **78.74** | 52.33 | 54.98 | 65.96 |
| org | 75.75 | 78.25 | 55.10 | **57.21** | **66.58** |

Table 11. **Influence of different weighting schemes**. VideoGEM with prompt decomposition and combined weights is evaluated with different weighting schemes: $w_{verb} = 0.1, w_{obj} = 0.3, w_{act} = 0.6$ (W1), $w_{verb} = 0.3, w_{obj} = 0.1, w_{act} = 0.6$ (W2), $w_{verb} = \frac{1}{3}, w_{obj} = \frac{1}{3}, w_{act} = \frac{1}{3}$(W3), $w_{verb} = 0.2, w_{obj} = 0.2, w_{act} = 0.6$ (org). Where *W1* prioritizes objects, *W2* prioritizes verbs, *W3* treats all prompts equally, and *org* are the original weights employed in the main paper. The weighting schemes are evaluated with ViCLIP on video input on V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT).

bounding box is green and the final VideoGEM prediction is white. The individual prompt predictions and heatmaps are shown for action, object, and verb prompts in red, blue, and yellow respectively. We observe that the heatmaps and the predicted locations for verb prompts differ the most from action and object prompts. Mostly all three predicted locations are close together, especially for actions with small spatial scale such as *"spread butter"*. However, when the action is bigger as for *"unpacking suitcases"* VideoGEM

| Model | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|
| Qwen-VL | **84.20** | 63.05 | 28.49 | **57.69** | 58.36 |
| VideoGEM | 75.74 | **78.25** | **55.10** | 57.21 | **66.58** |

Table 12. **Comparison to multimodal LLMs**. VideoGEM is compared to a multimodal large language model, namely Qwen-VL [1]. For Qwen-VL the prompt template "Generate grounding for " is applied. The center of the predicted bounding box is the prediction of Qwen-VL. VideoGEM is applied with ViCLIP as a backbone on video input. Both models are evaluated on V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT).

| Setting | VH | Daly | YC | gYT | avg |
|---|---|---|---|---|---|
| avg. txt. | 63.33 | 63.10 | 51.09 | 54.37 | 57.97 |
| avg. heat. | 63.33 | 63.04 | 51.09 | 54.41 | 57.97 |
| VideoGEM | **75.75** | **78.25** | **55.10** | **57.21** | **66.58** |

Table 13. **Comparison to CLIP prompting**. The prompt decomposition of VideoGEM is compared to the prompting technique of CLIP. The 80 prompt templates that are provided by the CLIP paper are used to generate 80 prompts. These prompts are used in two different settings. In the first setting, the 80 prompts are averaged, resulting in one text embedding. The combined weights of VideoGEM are then applied (*avg. txt.*). In the second setting, the combined weights are applied to each prompt individually resulting in 80 heatmaps (*avg. heat.*). The final prediction in both settings is obtained by taking the location of the maximum logit in the heatmap. VideoGEM is applied with ViCLIP as a backbone on video input. Both models are evaluated on V-HICO (VH), Daly, YouCook-Interactions (YC), and GroundingYouTube (gYT).

centers the action between its components. Moreover, if predictions are slightly off, such as in *"catching fish"*, or *"pulling son"*, where the action prompt initially is outside the ground truth bounding box focusing only on the object, VideoGEM drags the prediction back into the bounding box, leading to more robust and accurate predictions.

## 15. Limitations

VideoGEM is designed for spatial action grounding. It considers the temporal context of videos only implicitly via its video backbone. VideoGEM's predictions are spatial locations for the given input images/videos without temporal predictions. Moreover, the proposed prompt decomposition only works for action grounding in its current state. However, this could be adapted to more general settings.

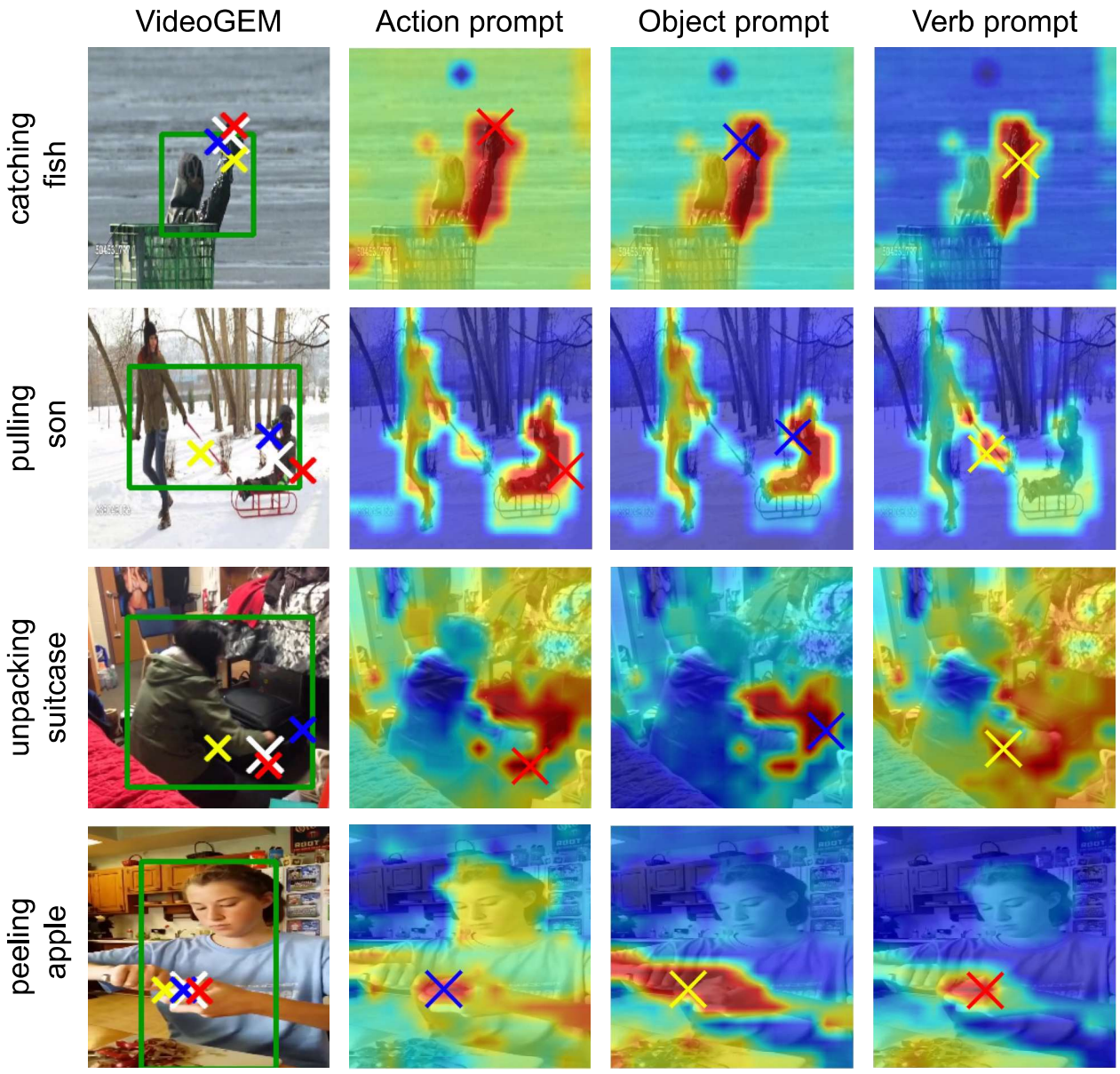


Figure 6. **Qualitative examples for VideoGEM on V-HICO**. VideoGEM is applied with ViCLIP on video data. The main frame with its label is shown. The green bounding box is the ground truth and the white cross is the final prediction of VideoGEM. Besides that are the heatmaps for the action, object, and verb prompt. The individual predictions for the action, object, and verb prompt are shown by red, blue, and yellow crosses respectfully. The ground truth label of the image is shown on the left.

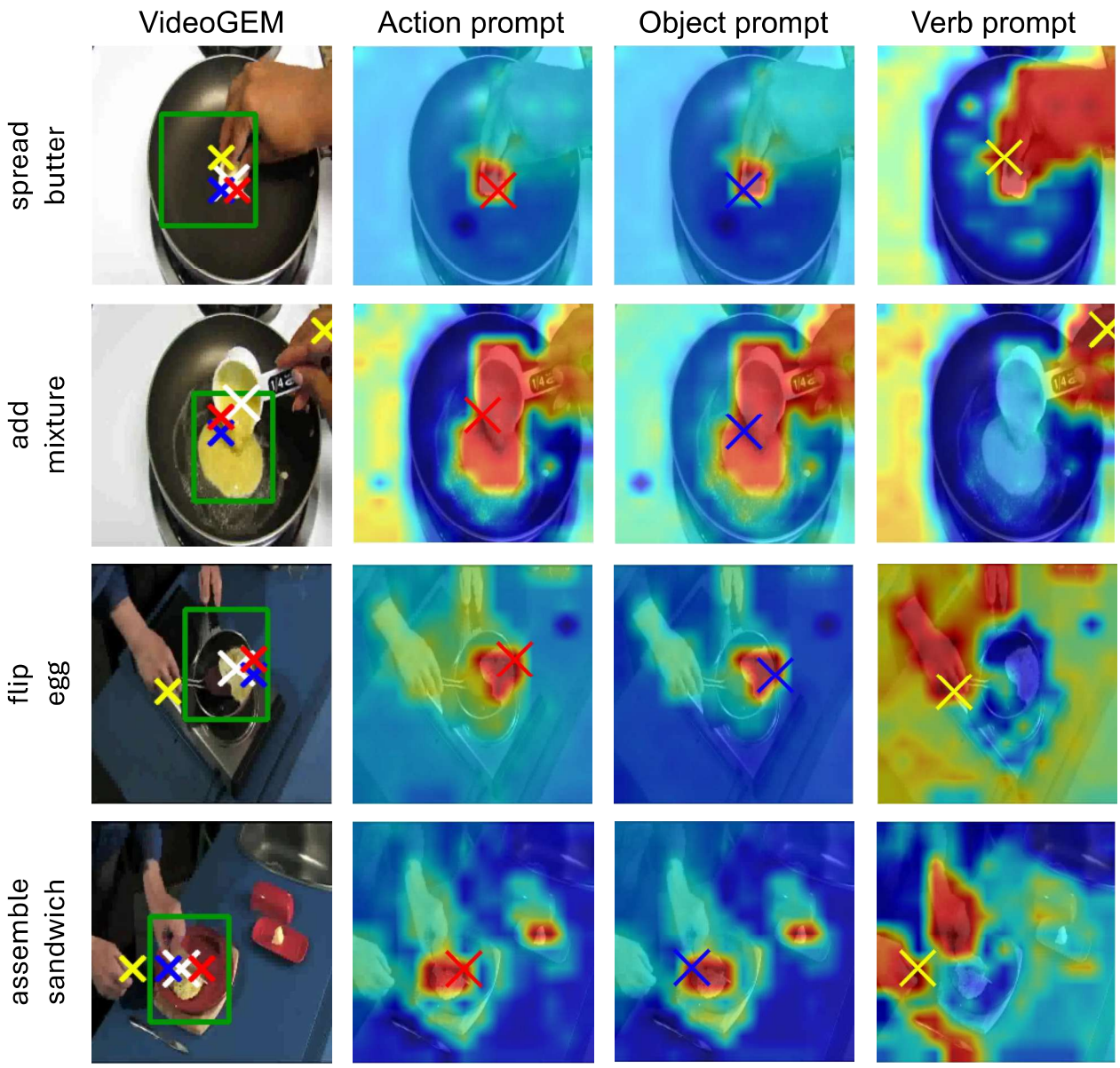


Figure 7. **Qualitative examples for VideoGEM on GroundingYouTube**. VideoGEM is applied with ViCLIP on video data. The main frame with its label is shown. The green bounding box is the ground truth and the white cross is the final prediction of VideoGEM. Besides that are the heatmaps for the action, object, and verb prompt. The individual predictions for the action, object, and verb prompt are shown by red, blue, and yellow crosses respectfully. The ground truth label of the image is shown on the left.