6 APPENDIX

Implementation Details In all experiments, training is conducted on 2 GPUs for 200 epochs. In each mini-batch, 256 samples are drawn for each GPU and in each sample, image regions are cropped from the whole image and resized to 112×112. The transformer encoder in all models has the same configuration: 4 layers, a hidden size of 384, and 4 self-attention heads in each layer. Besides, we use AdamW optimizer Loshchilov & Hutter (2017) with the learning rate 3e-5. For AdamW optimizer Loshchilov & Hutter (2017), we set the update coefficients, for averages of gradient and its square (β_1, β_2), and ϵ on denominator as 0.9, 0.999, 1e-4. During training, we mask out text tokens $\frac{1}{3}$ of the time and image tokens $\frac{1}{6}$ of the time and follow the same setting of random masking strategy with BERT Devlin et al. (2018).

During training and testing, for each target sample, we randomly select 200 remaining samples as the corresponding reference set. In our experiments, we utilize Top-K (K = 3) reference samples to get involved in analogical reasoning.

In evaluation, we calculate the accuracy based on the same random mask strategy as training process. An early stop strategy is utilized based on the Top-5 Acc. in validation that the training process will terminate if the validation Top-5 Acc. doesn't increase again.

Target-Reference Case Study From the two examples shown in Figure 7, it's easy to observe that the model can successfully retrieve relevant reference samples which contain analogical pairs, by computing their visual and language similarity to the target composition. For the correct prediction "peel carrot", the model discovers "stir carrot", "peel potato" and "cut potato" for analogical reasoning, while for "wash knife", the model retrieves "wash plate", "take knife" and "rinse knife" as reference samples.

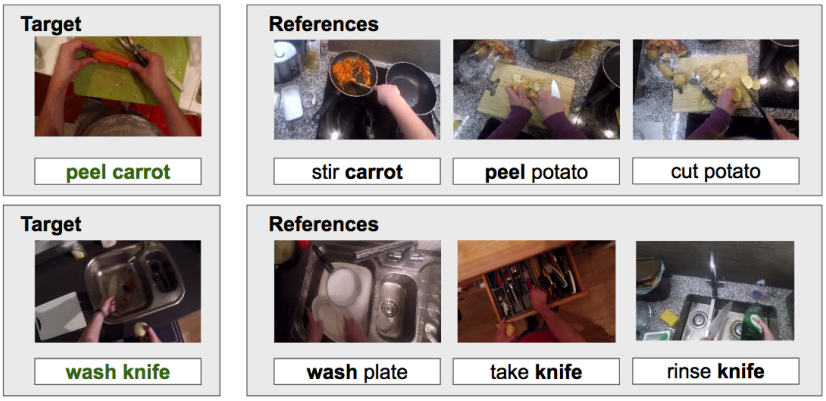

Figure 7: **Case Study:** The target sample and top samples discovered from the reference set.

Reasoning Attention Distribution over Multimodal Analogy Pairs In the four examples shown in Figure 7, we provide two examples for correct prediction and another two for wrong predictions. From the bottom two examples, the model learns compositional semantics from both visual and textual constituents. For the correct prediction of new composition "put sausage", the model learns to acquire and approximate novel composition from our multimodal reasoning. The reasoning has more attention on the textual phrase of the first reference sample "put oil in pan" and visual regions "sausage, whole image" of the second reference sample. This implies that the model learns textual and visual semantics from reference samples and compose them under similar scenarios as context. A similar phenomenon also appears in the second prediction example for seen composition "chop onions". The model is able to learn the phrase "chop onion" from different modalities. For wrong prediction results, the minor visual differences of several verbs will lead to wrong reasoning (e.g. "remove skin of garlic"), although the model can successfully retrieve relevant reference samples with aid by contextual information. While chopped garlic is not visually recognized by the adopted vision model, the attention distribution of the visual analogy pairs for the example seems to focus on the garlic but also be confused by "cream, whole image". Meanwhile, the accuracy of the reasoning

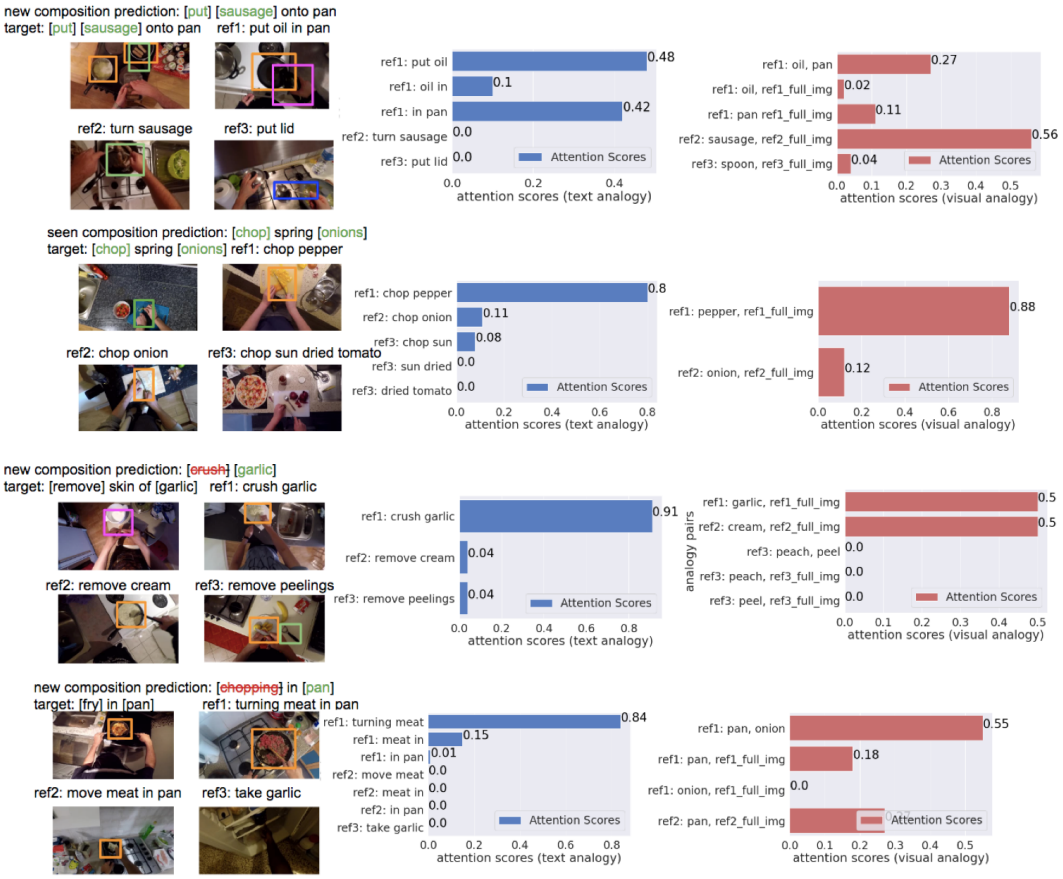

Figure 8: **Reasoning Attention Visualization over Multimodal Analogy Pairs (Correct or Wrong Predictions):** The bar charts shows attention scores in our reasoning module for textual (left chart) or visual (right chart) analogy pairs.

is also impacted by the relevant samples (e.g., "fry in pan"). When the model didn't discover "fry" in the relevant references and can not distinguish the actions ("fry" and "chop") in a target and a reference sample, the reasoning would easier to get the wrong prediction.