

Can CLIP text<->patch similarity focus the target object? (sim / RoboCerebra)

Zero-training probe: does CLIP's language alignment localize the named target object?

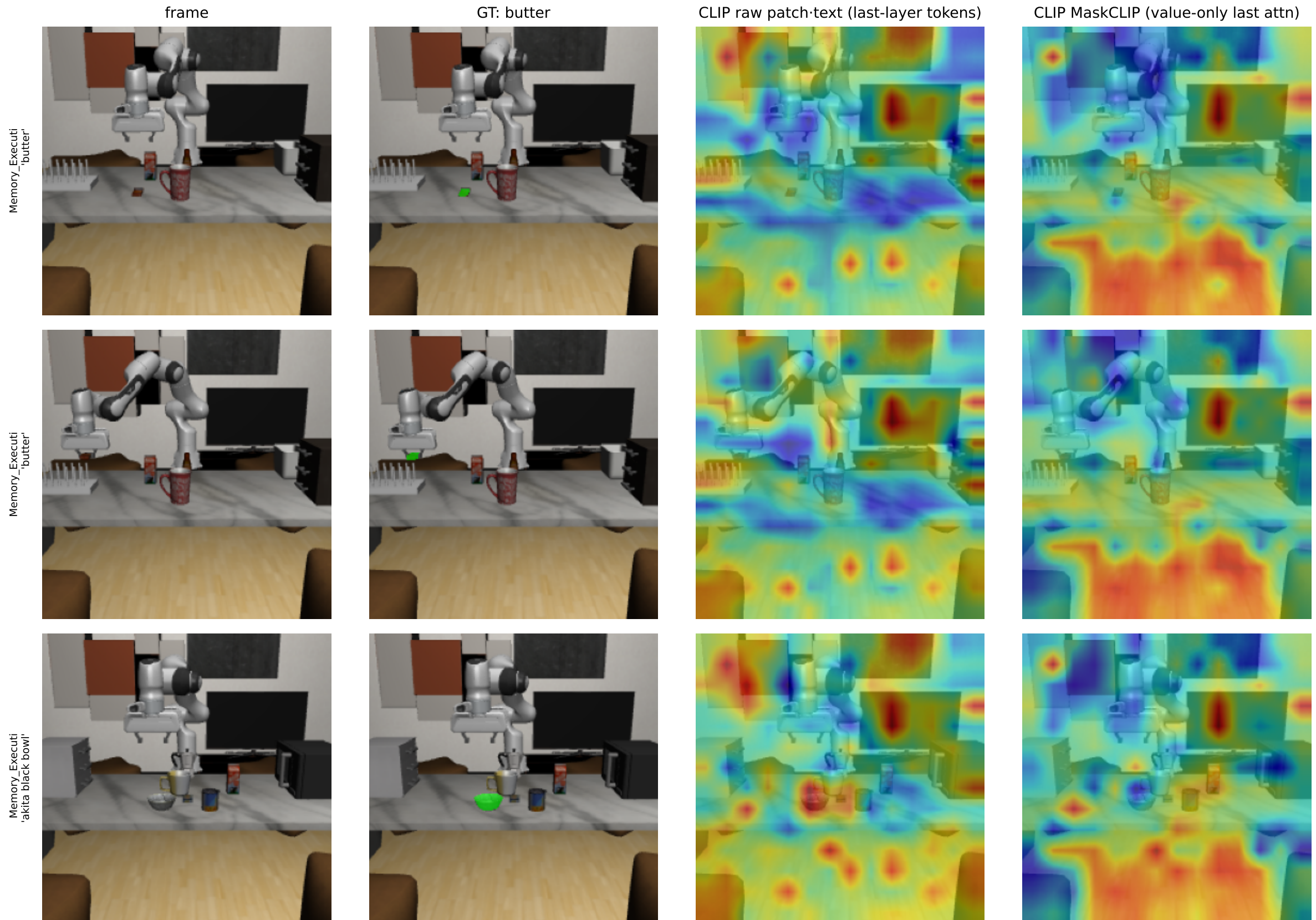
For each step we take the manipulated object's name (e.g. 'red coffee mug'), encode it with CLIP's TEXT encoder, and compute cosine similarity between that text vector and every image PATCH's CLIP feature -> a 14x14 heatmap. If CLIP grounds language, the heat should sit on the object.

Columns:

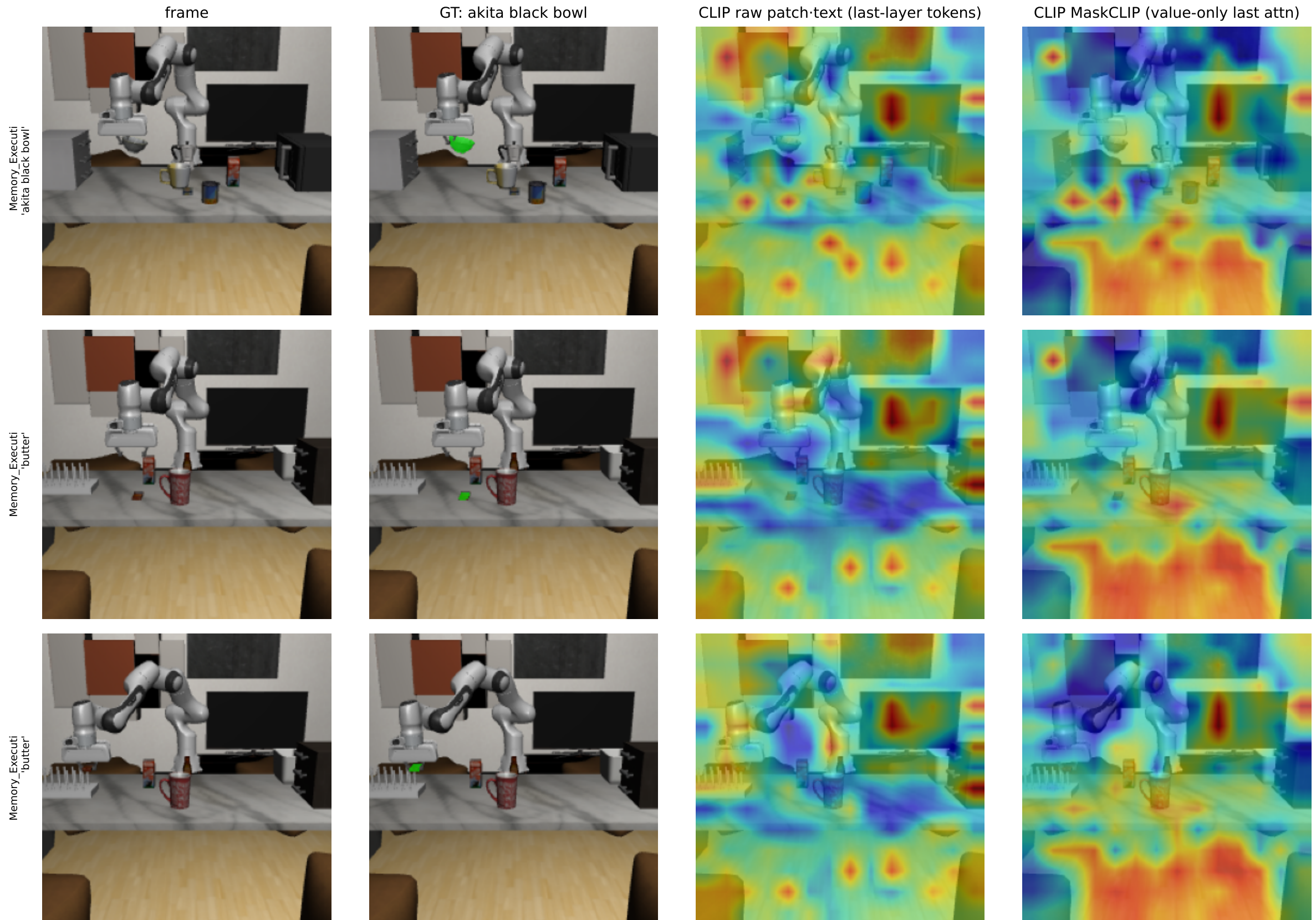
- | | |
|----------------------|--|
| frame | - the input RGB (RoboCerebra sim render, 224x224). |
| GT: <obj> | - the target object's ground-truth mask (green), from the simulator instance seg. |
| CLIP raw patch.text- | similarity using CLIP's LAST-layer patch tokens (projected to the joint space).
CLIP patch tokens are NOT trained for localization -> usually noisy. |
| CLIP MaskCLIP | - the standard fix: the LAST self-attention is made 'value-only' (no attention mixing) so each patch keeps its own local, language-aligned feature. This is the correct way to read dense/per-patch language similarity out of CLIP. |

Read: is the hot region (red) ON the green GT object? If neither raw nor MaskCLIP lands on it, CLIP cannot ground our objects here (expected cause: CLIP is trained on real photos -> sim is OOD, and some target names are odd, e.g. 'akita black bowl'). jet colormap: red=high, blue=low similarity.

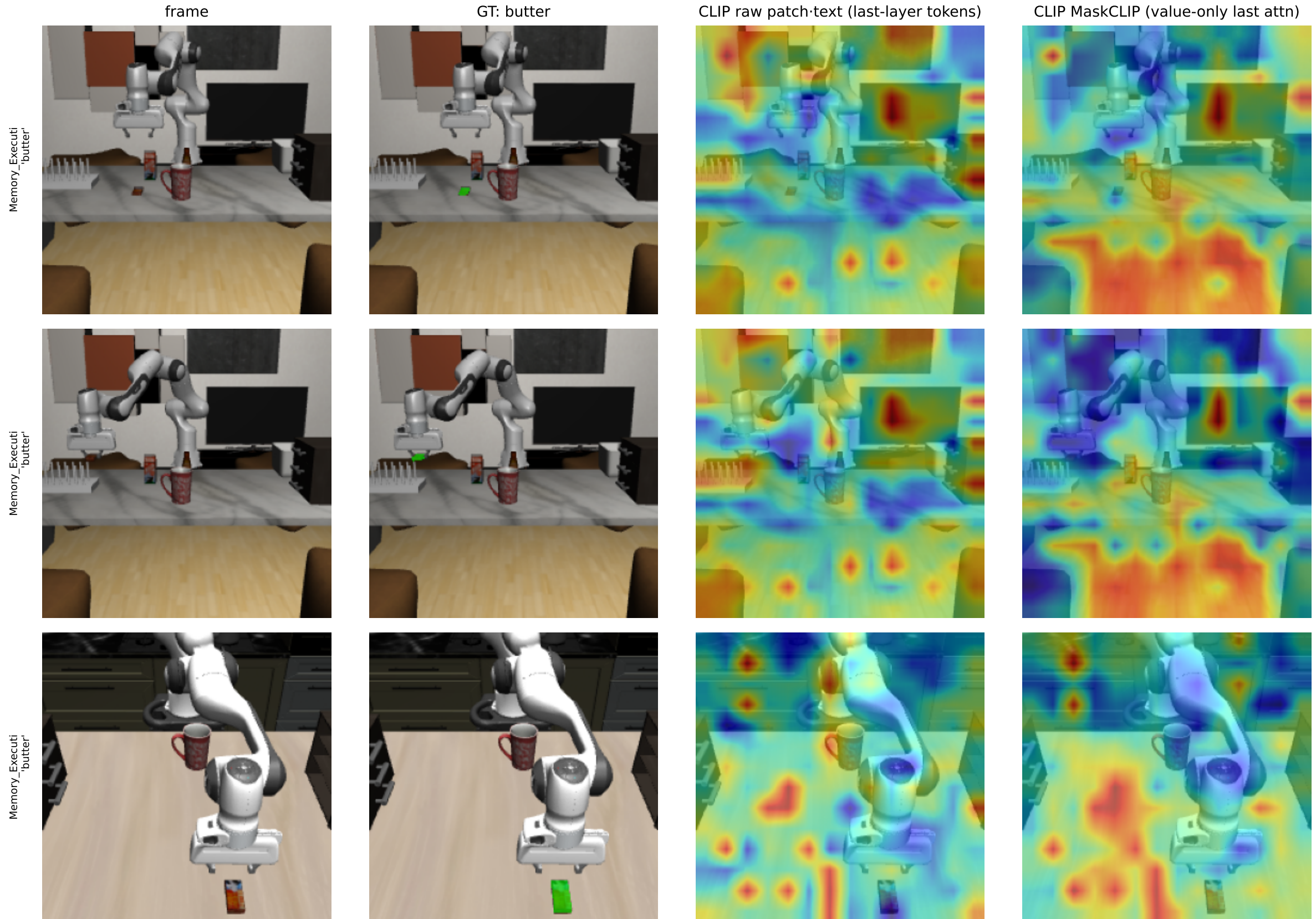
CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity



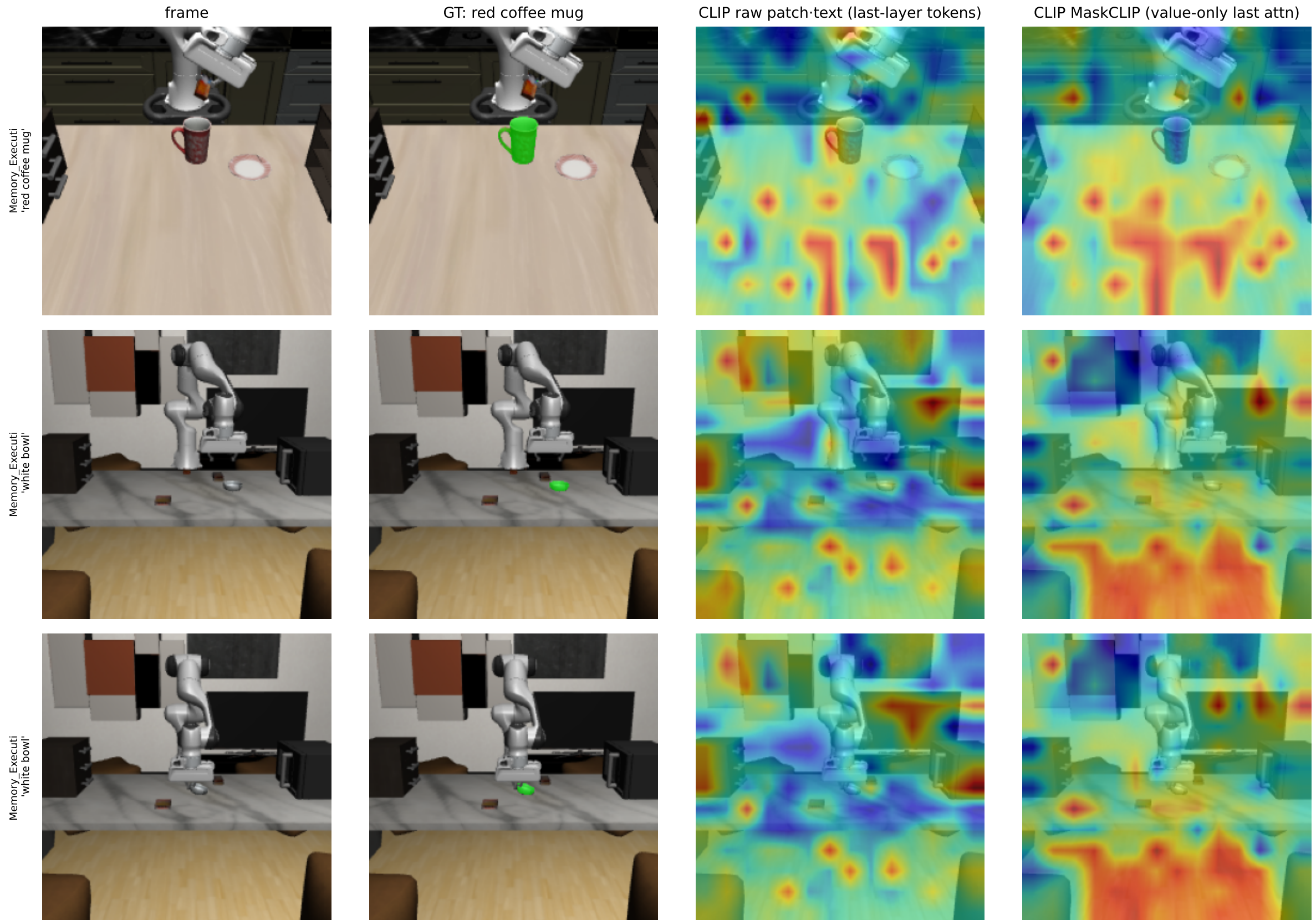
CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity



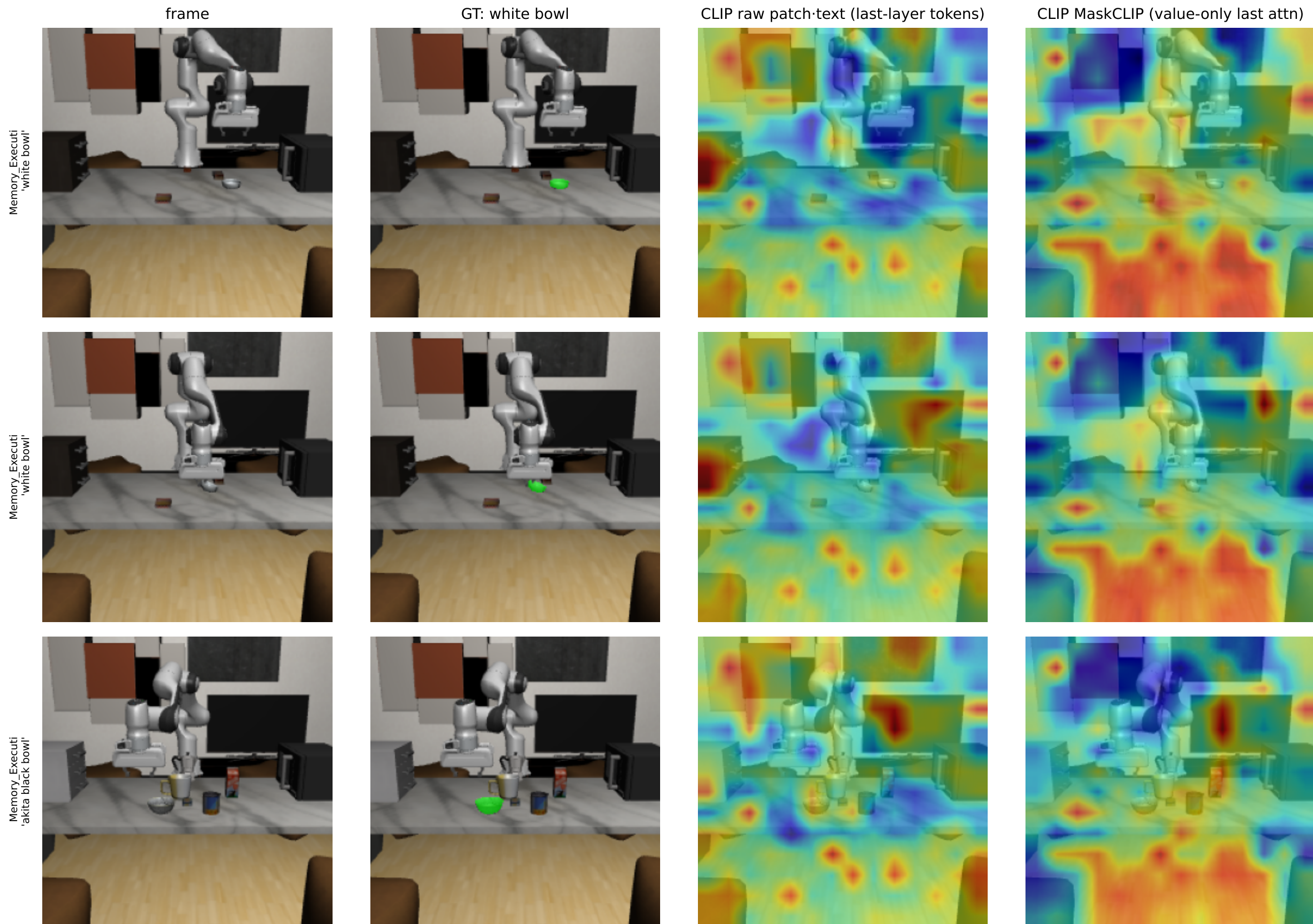
CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity



CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity



CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity



CLIP text<->patch grounding on sim | green=GT object · red=high CLIP-text similarity

Memory_Executi
'akita black bowl'

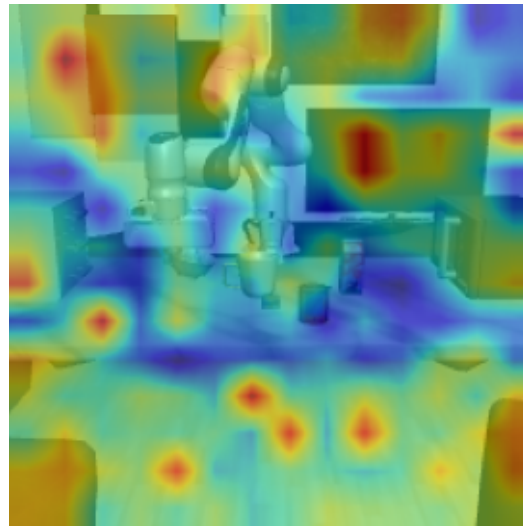
frame



GT: akita black bowl



CLIP raw patch·text (last-layer tokens)



CLIP MaskCLIP (value-only last attn)

