Scaling Multimodal Pre-Training via Cross-Modality Gradient Harmonization Suppmentary Material

1 Qualitative examples of our proposed measure based on agreement of gradients

To further validate the conjecture that *aligned video-text-audio triplets should have higher cosine similarity for g_{va} and g_{vt}, and vice versa*, we conduct a sanity check, where we started from a pre-trained VATT network and further optimize it for 200 iterations with a batch size of 64 on Youtube8M dataset, we then randomly sample video-text-audio triplets out of the samples with top 5% and bottom 5% most aligned gradient, measured by $cos(g'_{va}, g'_{va})$. In Figure 1 and 2, we visually observed that the top 5% group triplets has more semantically aligned words in the corresponding text narration (highlighted in **green**) thus enjoy a better cross-modality alignment, while the bottom 5% group triplets has fewer semantically aligned words, therefore are much more noisy in their alignments.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

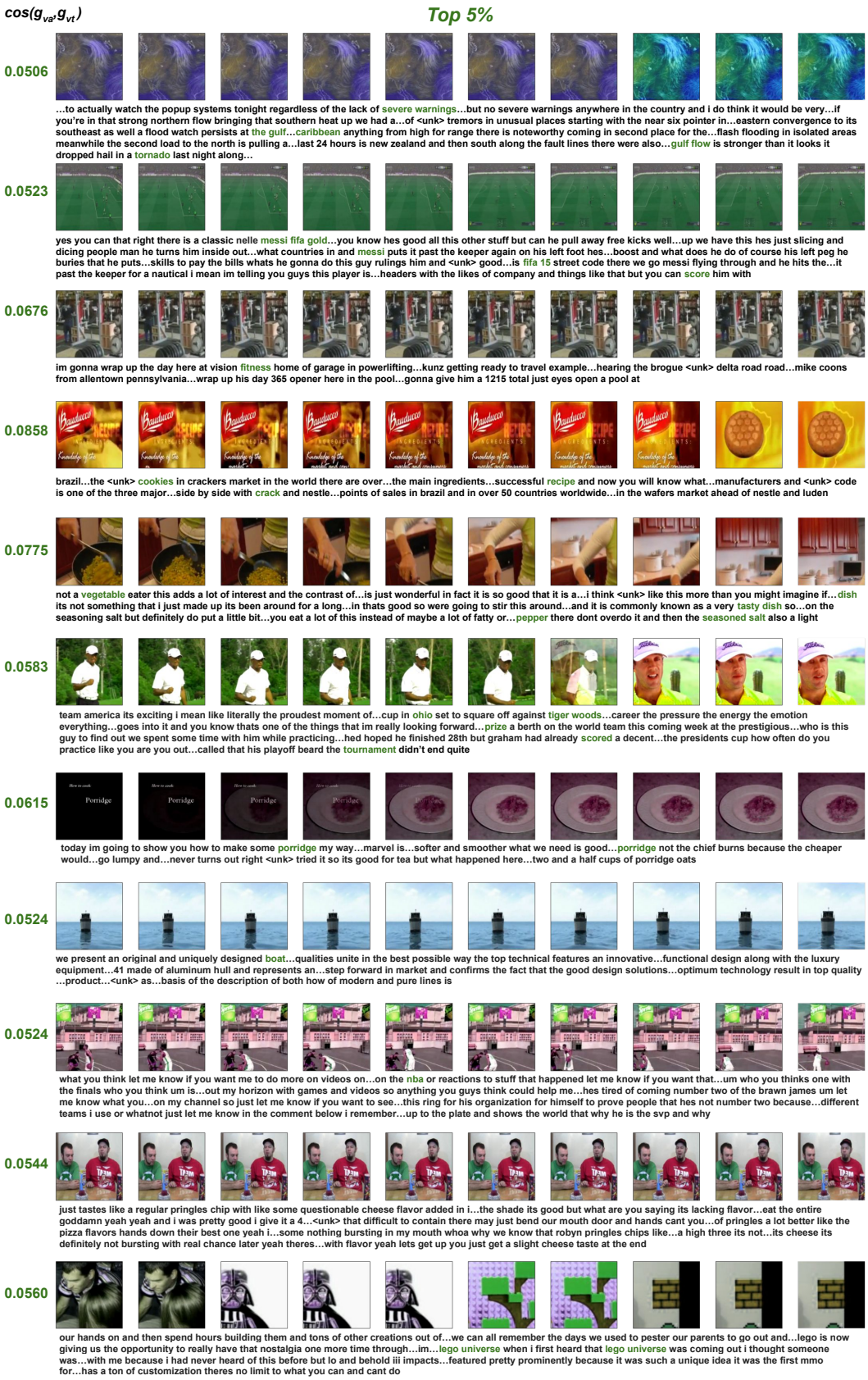
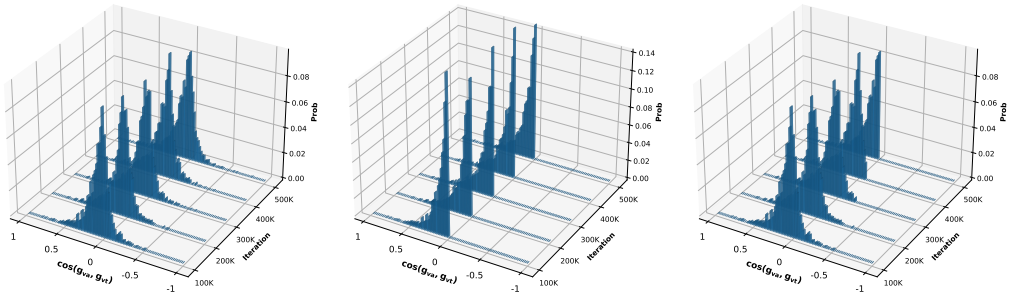


Figure 1: Visualization of Top 5% examples measured by the agreement of gradients on Youtube8M dataset, semantically aligned text are highlighted in **green**, $cos(g_{va}, g_{vt})$ reflect the agreement of between g_{va} and g_{vt}, by measuring their cosine similarity.

Figure 2: Visualization of Bottom 5% examples measured by the agreement of gradients on Youtube8M dataset, semantically aligned text are highlighted in **green**, $cos(g_{va}, g_{vt})$ reflect the agreement of between g_{va} and g_{vt}, by measuring their cosine similarity.

2 Training dynamics of our proposed measure on Youtube8M dataset

In Figure 3, we plot the distribution of cosine similarities, $cos(g_{va}, g_{va})$, across 500k iterations of pre-training, on the noisy Youtube8M dataset. We observe that $cos(g_{va}, g_{va})$ at any iteration resembles a normal distribution, and about half of the gradients g_{va} and g_{va} have misaligned directions (negative cosine similarities). We also calculate the mean of $cos(g_{va}, g_{va})$ across all 500k iterations on Youtube8M dataset, and found it to be 30% smaller than that on Howto100M dataset, indicating stronger mis-alignment in gradient directions on Youtube8M, which further verify that non-narrative video sets such as Youtube8M have more severe misalignment than the narrative ones such as Howto100M, hence challenging the CMA assumption commonly used in multimodal pre-training[**? ?**].



(a) Multimodal pre-training baseline

(b) w/ Cross-Modality Gradient Realignment

(c) w/ Gradient-based Curriculum Learning

Figure 3: Visualization of cosine similarity between gradient g_{va} and g_{vt} across 500K Iterations on the Youtube8M dataset. (a) shows the baseline setting in multimodal pre-training; (b) with only cross-modality gradient realignment; (c) with only gradient-based curriculum learning.