A Video is Worth 10,000 Words: Training and Benchmarking with Diverse Captions for Better Long Video Retrieval

Supplementary Material

8. GPT-3.5 Details

8.1. Prompts and Costs

We share prompts for summarization, simplification, and the combination of the two (joint). In the main paper, summarization is denoted as s, m, l depending on length, where s has 1 word and m has 4 words for every 7 words in l. Simplification is denoted by l+e, l+i, l+u. Joint is s+e, s+i, s+u.

We reduce the cost in terms of input token counts by batching our inputs. For example, we are generating 3 different summarizations per paragraph, but the source paragraph is the same in all 3 cases. So, instead of passing the input once for each level of summarization (3 times total), we pass the input once, and ask for all summarizations to be present in the output, reducing our input tokens by a factor of 3. We do the same for simplification and joint. So, if we want to generate summarization, simplification, and joint captions for a given ground truth caption, we must make 3 calls to the API (or, if hosted locally, one would have 3 forward passes). Remarkably, the model did not generate a malformed response a single time; in every case, we received each of the 3 requested outputs, properly tagged. It is worth mentioning these could possibly all be batched for a single pass, although at the time of preparing the dataset, the model was less robust under such conditions. If using our strategy for query expansion, discussed in Section 4.2, one would ideally batch all desired axes for a single pass, for the sake of speed.

The resulting costs can be computed in terms of tokens. The summarization prompt is approximately 180 tokens, not including the paragraph. For the 14,926 ActivityNet videos we consider, whose captions are an average of 49.8 words per caption, this means we submitted approximately 3.5 million input tokens for the 3 levels of summarization. Input tokens for the other two axes can be computed similarly. If using certain proprietary models, one must also consider the cost for output tokens, which can be estimated based on the length of the input paragraph compared to the word counts we provide for each dimension in Table 3. So, our final prompts are as follows for summarization, simplification, and joint. Note the use of "primary school" to generate our "elementary" level captions, and "secondary school" to generate "intermediate" captions.

Summarization You are a helpful writing assistant, with a speciality in summarizing text-based scene descriptions. You will be asked to write 3 summaries of the scene described in the following paragraph, indicated by PARAGRAPH. Do not modify the indicated order of events. Prioritize visual details. Do not hallucinate. Do not describe objects or events that do not appear in the original paragraph.
PARAGRAPH: ⟨*ORIGINAL PARAGRAPH*⟩.
Label this summary as SUMMARY_1. For this summary, please write 10 words which summarize the scene described by the PARAGRAPH. Do not use more or less than 10 words. Without using more than 10 words, write complete sentences.
Label this summary as SUMMARY_4. For this summary, please write 40 words which summarize the scene described by the PARAGRAPH. Do not use more or less than 40 words. Without using more than 40 words, write complete sentences.
Label this summary as SUMMARY_7. For this summary, please write 70 words which summarize the scene described by the PARAGRAPH. Do not use more or less than 70 words. Without using more than 70 words, write complete sentences.

Simplification You are a helpful writing assistant, with a speciality in simplifying and rewriting descriptions for different age groups and reading levels. You will be asked to write 3 versions of the scene described in the following paragraph, indicated by PARAGRAPH. Do not modify the indicated order of events. Prioritize visual details. Do not hallucinate. Do not describe objects or events that do not appear in the original paragraph.
PARAGRAPH: ⟨*ORIGINAL PARAGRAPH*⟩.
Label this version as VERSION_primary_school. For this version, rewrite the PARAGRAPH with 70 words to make it suitable for a primary school reading level.
Label this version as VERSION_secondary_school. For this version, rewrite the PARAGRAPH with 70 words to make it suitable for a secondary school reading level.
Label this version as VERSION_university. For this version, rewrite the PARAGRAPH with 70 words to make it suitable for a university reading level.

Joint You are a helpful writing assistant, with a speciality in summarizing text-based scene descriptions. You also have a speciality in simplifying and rewriting descriptions for different age groups and reading levels.

> You will be asked to use 10 words to write 3 summaries of the scene described in the following paragraph, indicated by PARAGRAPH. Do not modify the indicated order of events. Prioritize visual details. Do not hallucinate. Do not describe objects or events that do not appear in the original paragraph.
> PARAGRAPH: $\langle$*ORIGINAL PARAGRAPH*$\rangle$.
> Label this version as VERSION_primary_school. For this version, rewrite the PARAGRAPH with 10 words to make it suitable for a primary school reading level. Do not use more or less than 10 words. Without using more than 10 words, write complete sentences.
> Label this version as VERSION_secondary_school. For this version, rewrite the PARAGRAPH with 10 words to make it suitable for a secondary school reading level. Do not use more or less than 10 words. Without using more than 10 words, write complete sentences.
> Label this version as VERSION_university. For this version, rewrite the PARAGRAPH with 10 words to make it suitable for a university reading level. Do not use more or less than 10 words. Without using more than 10 words, write complete sentences.

8.2. Automatic Analysis

We provide LF-VILA and QuerYD to complement Table 3 in Table 13 and Table 14, respectively. These are conistent with the major trends for ActivityNet10k, with the notable difference that since these captions are longer, the absolute differences are larger.

8.3. Annotator Analysis

For our sample, we recruited 15 individuals, all of whom had at least a bachelor's degree. Individuals spent between 10 and 20 minutes to answer the 15 questions on their assigned survey. For an example survey, please refer to the attached material.

9. Ablations

We share some ablations that indicate how we choose hyperparameter values. The most important thing is that the losses are used, and the change that causes the most different is training with $\eta = 0.0$, highlighting the importance of using 10k Words data while training.

10. Miscellaneous

10.1. Hallucination Prevalence Results

In Table 6 we give results computed in two ways, as the percentage of all votes which belong to a given category ("Total") and by determining the majority label for each word, then computing percentages ("Majority"). To further clarify this computation, consider the following example, with 3 voters and 3 words. For the first word, 2 voters select matches, 1 selects unsure. For the second word, all 3 voters select unsure. For the third word, 3 select different. Since there were 3 votes for different, 4 for unsure, and 2 for matches, the percentages for total would be 33.33%, 44.44%, and 22.22% respectively. For majority, since the first was majority matches, second was majority unsure, and third was majority different, these would be 33.33% each.

10.2. Training-time Improvement Details

First, we show an illustration of our data sampling approach, as a visual aid, in Figure 5.

Since part of our contribution is a data augmentation strategy, we also evaluate its performance by finetuning with different fractions of the original ActivityNet data in Figure 6. Notice that the absolute differences in recall between training with 10k data and training without remain consistent for all amounts of training data. For training on short captions the difference is around a 3% improvement while for long captions it is around 2%. By training with synthetic data, we achieve the same performance with less manually annotated data.

We also show that our findings hold when finetuning on other datasets, such as LF-VILA (Table 16).

10.3. Inference-time Improvement Details

For our ensembles in Section 4.2, for the sake of simplicity, for synthetic captions we choose the 'l' and 'l+i' captions, since we find that 'l+e' and 'l+u' have higher tendency to either reduce information (for 'l+e') or else infer unnecessary detail (for 'l+u'). Sampling short and medium length captions is less effective in this regime due to the information loss. Introducing such ambiguity into the retrievals would be counterproductive. To actually perform the retrieval, we compute the standard text-video similarity matrix, as well as a separate text-video similarity matrix for each type of 10k caption ('l' and 'l+i'). We then add these together, giving 50% weight to the standard text-video matrix, and equal weight to the remaining 2 matrices.

10.4. Information Loss and Uniqueness Details

We realize the length is not a perfect measure of information. In fact, part of the motivation of this work is that captions can be quite short but very information-dense. So, we compute information loss is 3 ways. First, we use short length divided by standard length, as given in the main paper in Figure 3. Second, we use spaCy to count entities in the short and standard captions, dividing the number in the short by the number from the source standard caption in Figure 7. Third, we get the word2vec embeddings for the entities in the short and standard captions, and compute the cosine similarities between all entities. We choose the

Table 13. Automatic dataset statistics for LF-VILA10k. We show the average change in unique nouns and verbs, as well as word count and length.

		Summarization			Simplification			Summarization and Simplification		
Metric	Source	Short	Medium	Full Length	Elementary	Intermediate	University	S and P	S and S	S and U
Δ Nouns		-11.77	-4.23	1.49	-1.40	3.75	11.14	-14.35	-13.18	-12.36
Δ Verbs		-2.60	0.90	4.32	1.96	7.63	11.71	-3.07	-2.34	-1.86
Word Count	155.40	36.06	76.30	105.43	129.52	136.58	154.13	28.94	31.85	34.83
Word Length	4.66	5.00	4.96	5.18	4.79	5.25	5.74	4.66	4.90	5.12

Table 14. Automatic dataset statistics for QuerYD10k. We show the average change in unique nouns and verbs, as well as word count and length.

		Summarization			Simplification			Summarization and Simplification		
Metric	Source	Short	Medium	Full Length	Elementary	Intermediate	University	S and P	S and S	S and U
Δ Nouns		-27.25	-20.38	-14.81	-14.83	-9.26	-2.55	-32.21	-31.16	-29.27
Δ Verbs		-12.10	-7.90	-4.46	-3.04	1.26	4.56	-14.11	-13.57	-12.83
Word Count	207.86	53.41	86.69	114.55	150.97	164.26	181.92	34.81	37.28	43.10
Word Length	5.47	5.89	5.72	5.79	5.27	5.66	6.02	5.37	5.73	5.98

Table 15. Mixing ratio ablations.

	ActivityNet		
η	Full	Short	Long
0.0	59.4	31.9	55.8
0.25	**60.1**	33.2	**56.6**
0.5	59.4	33.3	56.5
0.75	59.9	**33.5**	56.2
1.0	59.3	**33.5**	**56.6**

Table 16. LFVILA COSA finetuning. Results improve with 10k finetuning.

Finetune Method	All		Short		Long		Partial	
	R@1	Avg. R	R@1	Avg. R	R@1	Avg. R	R@1	Avg. R
Domain	77.3	86.9	65.2	78.4	90.2	95.9	73.8	84.9
Ours	**85.2**	**92.6**	**78.2**	**89.2**	**95.3**	**98.2**	**73.0**	**83.9**

best matches for the entries in the short caption, and sum the similarities, then divide by the number of entries in the short caption. Hence we use similarity between bags of words as our proxy for how much the information in the short caption overlaps the information in the standard caption, with results in Figure 8. These two alternatives confirm the findings from using length, so we opt to use length in the main paper since it is simpler.

To calculate uniqueness, we take the similarity score defined above (greedy matching of cosine similarities for word2vec embeddings of entities). We additionally compute the similarity between the short caption and the standard captions for the top 5 retrieved videos, as retrieved using the short caption, not including the standard caption for the matching video. That is, if the matching video is in the top 5 retrievals, we exclude it and additionally consider the standard caption for the video retrieved at rank 6. We average the similarities between the short caption and these 5 standard captions, and subtract it from the similarity between short and source (matching) standard caption, for a uniqueness score. This "uniqueness" score provides the color in Figure 3, Figure 7, and Figure 8.

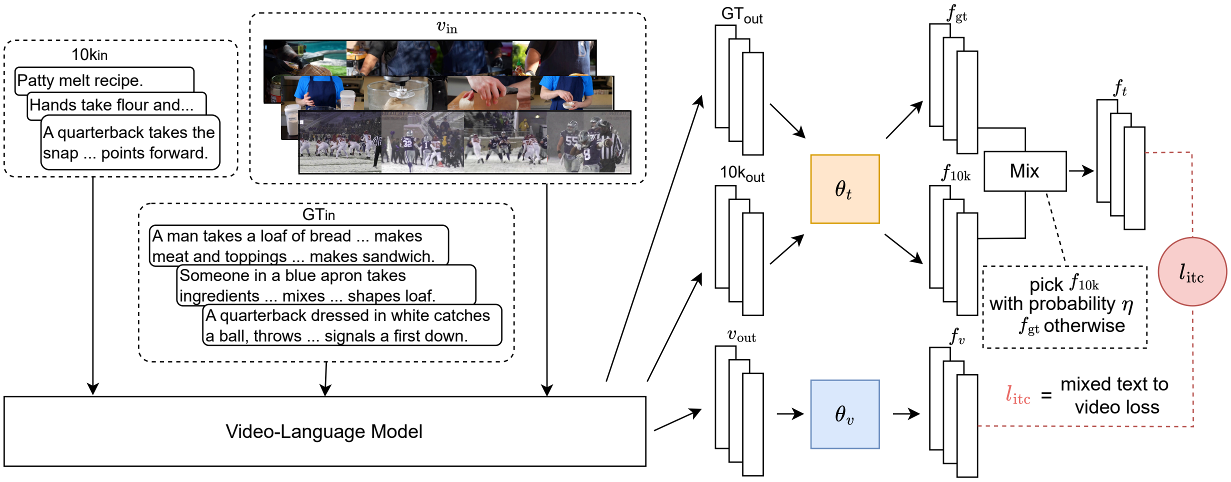
10k$_{\text{in}}$
Patty melt recipe.
Hands take flour and...
A quarterback takes the snap ... points forward.
v_{in}
GT$_{\text{in}}$
A man takes a loaf of bread ... makes meat and toppings ... makes sandwich.
Someone in a blue apron takes ingredients ... mixes ... shapes loaf.
A quarterback dressed in white catches a ball, throws ... signals a first down.
Video-Language Model
GT$_{\text{out}}$
10k$_{\text{out}}$
v_{out}
θ_t
θ_v
f_{gt}
$f_{10\text{k}}$
f_v
Mix
f_t
pick $f_{10\text{k}}$ with probability η f_{gt} otherwise
l_{itc}
l_{itc} = mixed text to video loss

Figure 5. We perform contrastive finetuning for retrieval with video-caption pairs. We propose efficient sampling of our 10k text captions for data augmentation, where we compute standard contrastive loss, but each caption is sampled randomly from the 10k captions for a given video, according to a mixing ratio, η.

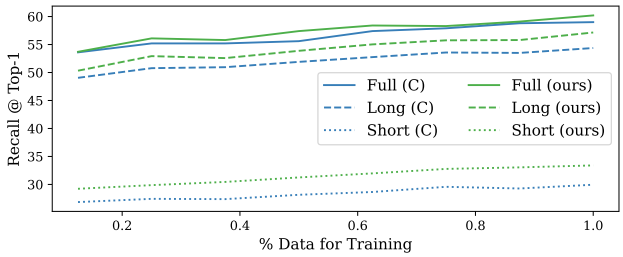

Figure 6. We measure how much our data augmentation helps in the data constrained regime, training only with the indicated amounts of data, and performing retrieval with the resulting trained models. We show that finetuning COSA with 10k data (ours) is superior to generic COSA finetuning (C) for ActivityNet10k.

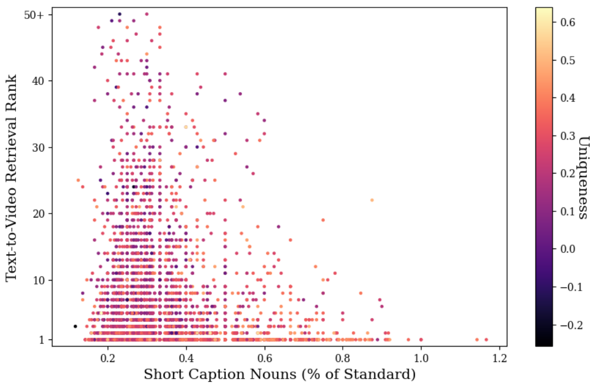

Figure 7. We measure the number of nouns and retrieval uniqueness for short caption retrieval, and find that the highest ranks correlate with captions that have lost their nouns and unique information.

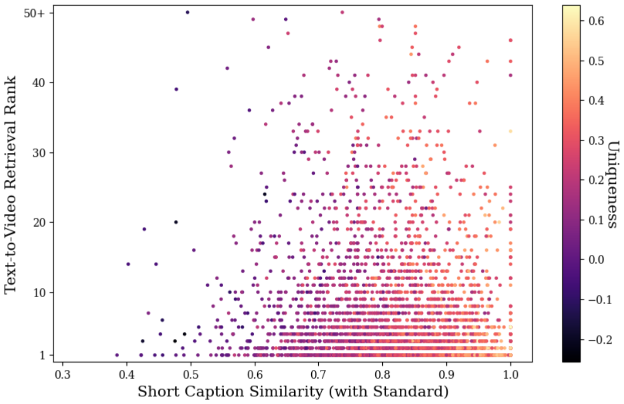

Figure 8. We measure the number of nouns and retrieval uniqueness for short caption retrieval, and find that the highest ranks correlate with captions that have lost their similarity with the source caption.