

Fighting FIRE with FIRE: Assessing the Validity of Text-to-Video Retrieval Benchmarks

Pedro Rodriguez,*
Mahmoud Azab, Becka Silvert, Renato Sanchez,
Linzy Labson, Hardik Shah, Seungwhan Moon
Meta AI

Abstract

Searching troves of videos with textual descriptions is a core multimodal retrieval task. Owing to the lack of a purpose-built dataset for text-to-video retrieval, video captioning datasets have been re-purposed to evaluate models by (1) treating captions as positive matches to their respective videos and (2) assuming all other videos to be negatives. However, this methodology leads to a fundamental flaw during evaluation: since captions are marked as relevant *only* to their original video, many alternate videos *also* match the caption, which introduces false-negative caption-video pairs. We show that when these false negatives are corrected, a recent state-of-the-art model gains 25% recall points—a difference that threatens the validity of the benchmark itself. To diagnose and mitigate this issue, we annotate and release 683K additional caption-video pairs. Using these, we recompute effectiveness scores for three models on two standard benchmarks (MSR-VTT and MSVD). We find that (1) the recomputed metrics are up to 25% recall points higher for the best models, (2) these benchmarks are nearing saturation for Recall@10, (3) caption length (generality) is related to the number of positives, and (4) annotation costs can be mitigated through sampling. We recommend retiring these benchmarks in their current form, and we make recommendations for future text-to-video retrieval benchmarks.

1 Introduction

Text-to-video retrieval (TVR) is a challenging multimodal retrieval task (Hu et al., 2011) with practical applications ranging from web search to organizing media collections (Lew et al., 2006). To measure TVR model improvement—despite a dearth of purpose-built TVR benchmarks—researchers created benchmarks by re-purposing video captioning datasets such as MSR-VTT (Xu et al.,

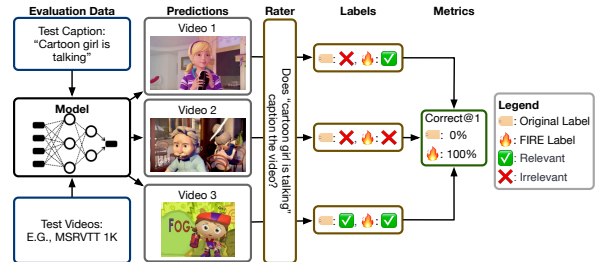


Figure 1: MSR-VTT and MSVD have one positive video per caption (each video’s caption). Captions often match multiple videos, leading to false negatives. When models rank false negatives highly, model quality is understated (full example in Appendix Figure 5). This leads to evaluations where reported metrics do not reflect their true value and are therefore not internally valid (§2.2.1).

2016), MSVD (Chen and Dolan, 2011), and ActivityNet (Heilbron et al., 2015; Krishna et al., 2017). Early work established an evaluation paradigm that treated captions as search queries over the collection of captioned videos (Zhang et al., 2018; Yu et al., 2018; Gabeur et al., 2020); each caption and their corresponding video are positives (relevant) during retrieval, and all other caption-video pairs are negatives (irrelevant).

However, even a cursory inspection of videos and captions reveals many additional positive caption-video pairs (§2). In current benchmarks, *true positives* that are not the video’s original caption are falsely assumed to be *negatives*. Wray et al. (2021) first identified this fundamental, false-negative problem in TVR evaluation; our work builds on this by quantifying the absolute metric differences that false negatives induce (see discussion in §6). Accurate absolute metrics are crucial in industrial settings where deployment criteria are often defined by minimum quality targets. These **False Implicit Relevance** labels introduce measurement error—e.g., CLIP4CLIP’s (Luo et al., 2021) Recall@1 is underestimated by 25% points (§2.2). We estimate measurement error by

*Correspondence to me@pedro.ai

annotating 683K additional caption-video pairs, which we call the FIRE 🔥 dataset (§3).¹

A core measurement principle is that operationalized metrics should strongly correlate to the quantity they intend to measure (Mathison, 2004; Liao et al., 2021). For example, Recall@K operationalizes the intent to measure retrieval quality. Label errors are a common way that measurements are invalidated (Bowman and Dahl, 2021; Northcutt et al., 2021). Our work shows that since TVR metrics are computed with false negative label errors, Recall@K does not accurately reflect retrieval quality, which negates the measurement’s validity. In the remainder of this paper, we posit rationales of why models gain different score boosts (§4.1) and estimate how useful the FIRE dataset is for evaluating future models (§4.2 and §4.3).

To conclude, we review the implications of our findings. Looking to the past, retrieval effectiveness has been understated for some models, which gives an overly pessimistic view of recent advances (Bowman, 2022). Critically, our results also suggest that the MSR-VTT benchmark is nearing saturation and should be retired soon in favor of a purpose-made benchmark. Looking outward, we identify structurally similar benchmarks—such as photo retrieval—that likely also have the same **False Implicit Relevance** problem. A successful benchmark should avoid the pitfalls we identify in this paper, be faithful to the real-world user task it targets (Rowe and Jain, 2005; de Vries et al., 2020), improve reproducibility, and evolve (§7).

2 Text-to-Video Retrieval Evaluation

This section reviews current TVR evaluation practices using two concepts: *internal validity* (Campbell, 1957, §2.2.1) and *construct validity* (Tague-Sutcliffe, 1992, §2.2.2). *Internal validity* refers to whether an evaluation reliably establishes a cause-effect relationship between the measured dependent variable and the independent variable to be estimated (Brewer and Crano, 2014; Liao et al., 2021). In TVR evaluations, false negatives confound model quality and label errors (i.e., is the model wrong or is the label wrong?) which makes *reliably* establishing cause (model quality) and effect (retrieval score) difficult. *Construct validity* “pertains to the degree to which the measure of a construct sufficiently measures the intended concept” (O’Leary-Kelly and J. Vokurka, 1998)—in

¹Data and Code: pedro.ai/multimodal-retrieval-evaluation.

TVR evaluations, an important intended concept is real-world search quality. *Construct validity* asks: can we expect that measuring retrieval quality with the benchmarks at hand generalizes to real-world search quality? This section argues that TVR evaluations are not internally valid or construct valid.

2.1 Model Evaluation

Multimodal retrieval evaluations typically focus on two tasks: text-to-video and video-to-text retrieval. The first task’s goal is—given a text query—to retrieve videos that match; the second task’s goal is—given a video—to retrieve the matching queries. The applications of text-to-video search are straightforward: it is useful for searching the web and personal media.² Since the applications of TVR are clear, and the false-negative problem is present in both tasks, here we focus on TVR.

The MSR-VTT and MSVD Datasets: It is standard for TVR evaluations (Zhang et al., 2018; Yu et al., 2018; Gabeur et al., 2020) to report on MSR-VTT and MSVD, so in the interest of comparability, we use these benchmarks too. Although these datasets were originally meant for evaluating video captioning models, they have been repurposed for TVR (Zhang et al., 2018; Gabeur et al., 2020). In this paper, we focus our investigation on MSR-VTT and MSVD since they are the most prevalent in prior work. MSR-VTT consists of 10K videos, 1K of which are in the test split. Each video has twenty captions, but for evaluation, only one (arbitrarily chosen) caption is used. MSVD contains 1,970 videos, 960 of which are in the test split. Videos have about forty captions; unlike with MSR-VTT, retrieval quality for each caption is evaluated.

Fundamentally, both MSR-VTT and MSVD are video captioning datasets—not retrieval datasets. MSVD addressed the lack of standard benchmarks for paraphrasing (Chen and Dolan, 2011). In the original task, annotators selected short clips from YouTube, watched the clip, and wrote a sentence describing its contents. The process was repeated for each video, with each sentence being written by a new annotator. This conditional independence—given the video—resulted in a diverse set of captions. MSR-VTT captions were collected similarly: independent annotators captioned the same video. Videos were sourced from the output of a commercial video search engine (Xu et al., 2016). In both

²The applications of video-to-text retrieval—that are not simply captioning—are not clear to us.

datasets, video captions are used as search queries and labeled relevant to the original video.

Metrics: Previous TVR work (Zhang et al., 2018; Yu et al., 2018; Gabeur et al., 2020; Luo et al., 2020; Zhu and Yang, 2020; Li et al., 2020; Xu et al., 2021; Park et al., 2022) reports Recall@K (R@K)³ and sometimes supplemental metrics such as median or mean rank of the first correct result. However, R@K in TVR work differs from the textbook information retrieval definition (Manning et al., 2008, p. 155) where

$$\text{R@K} = \frac{\# \text{ retrieved positives in top K}}{\# \text{ total positives in collection}}. \quad (1)$$

In TVR work, query retrieval results are scored one if a relevant video is in the top K and zero otherwise. The traditional definition of Recall@K *only* reduces to this when there is *exactly* one positive in the collection but is not comparable when there are multiple positives per caption—as in this case.

With the difference now salient, we avoid confusion by defining a new quantity Correct@K (C@K) which is 1 if at least one positive is in the top K and 0 otherwise. Correct@K naturally reduces to Recall@K—as defined in prior work—when there is exactly one positive, but handles the additional positives in our work. We recommend reporting Correct@K as well as mean average precision (Su et al., 2015; Mitra and Craswell, 2018, MAP), a metric widely used in Information Retrieval.

The drawback of Correct@K—shared by median (or mean) rank to first positive—is that it does not directly factor in rank order when there are multiple positives in retrieved results, only coarsely factoring in rank via K value. MAP (Mitra and Craswell, 2018, p. 19) is calculated by taking the mean of

$$\text{AvgPrec}_q = \frac{\sum_{\langle i,v \rangle \in R_q} \text{Prec}_{q,i} \times \text{rel}_q(v)}{\sum_{v \in V} \text{rel}_q(v)} \quad (2)$$

for each test query q where i is a video’s position in the ranked list R_q of videos, v is a video in collection V , and $\text{rel}_q(v)$ denotes whether query q is relevant to video v . Intuitively, this translates to calculating the mean of Precision@K for every K where a positive occurs in ranked predictions R_q . In all experiments, we report Correct@K and MAP.

2.2 Questioning the Validity of Evaluations

In this section, we experimentally argue that current TVR evaluations are not *internally valid*. Then

³Typical K values include 1, 5, 10, and 50.

we argue that they are not *construct valid* by considering actual use-cases for video search.

2.2.1 Internal Validity

If an evaluation metric is internally valid (Liao et al., 2021), then model effectiveness (cause) should be *accurately and reliably* reflected in metrics (effect) (Brewer and Crano, 2014). A central hypothesis of this paper is that the prevalence of false negatives invalidates the cause-effect relationship between measured model effectiveness and actual effectiveness—i.e., that correcting false negatives will significantly change metrics.⁴

To test this hypothesis, we build the FIRE dataset, which Fixes Implicit Relevance Errors. We detail the dataset later (§3), but in short, we take strong retrieval models from the past few years and annotate their top ten predictions on both MSR-VTT and MSVD. This process—called system pooling—has been used for decades in information retrieval (Spark-Jones, 1975) and, by construction, eliminates implicit false negatives.⁵ For MSR-VTT, we collect annotations from TeachText (Croitoru et al., 2021), Support-Set Bottlenecks (Patrick et al., 2021, SSB), and CLIP4CLIP (Luo et al., 2021) models; for MSVD, we collect annotations from TeachText and CLIP4CLIP models.^{6,7} Next, we compute model scores using the original positives and compare them to scores calculated with *both* the original positives *and* the new positives in FIRE.

Table 1 clearly demonstrates that FIRE annotations reveal large metric differences in both MSR-VTT and MSVD. For example, the C@1 score of CLIP4CLIP is understated by 25% points, and its C@10 score arguably saturates the benchmark at 95.7%. Even “small” differences such as those for TeachText and SSB are on par with the differences used to claim state-of-the-art results. False negatives directly cause high measurement error, which invalidates the internal validity of the benchmark.

⁴We do not see rank changes in our three models, but score differences suggest that ranks may change with more models.

⁵By implicit, we mean false negative from the lack of labeling and presuming non-positives are (implicitly) negative. There may still be false negatives arising from human error during annotation.

⁶We prioritize models that are (1) publicly available and (2) have sufficient documentation to reproduce.

⁷Annotating MSR-VTT predictions translates to $1,000 * 10 = 10\text{K}$ annotations since only one caption per video is used. This is easy compared to MSVD annotation, which uses tens of captions per video.

Dataset	Metric	TeachText	SSB	CLIP4CLIP
MSR-VTT	C@1	24.1 (23.3 + 0.800)%	27.3 (26.8 + 0.500)%	67.4 (42.4 + 25.0)%
MSR-VTT	C@5	53.2 (50.9 + 2.30)%	55.9 (54.5 + 1.40)%	90.7 (70.4 + 20.3)%
MSR-VTT	C@10	67.0 (64.8 + 2.20)%	68.9 (66.3 + 2.60)%	95.7 (80.2 + 15.5)%
MSR-VTT	AP	36.1 (35.8 + 0.296)%	39.3 (39.2 + 0.0374)%	69.5 (54.9 + 14.7)%
MSVD	C@1	34.7 (19.6 + 15.2)%	Not Annotated	65.3 (46.6 + 18.8)%
MSVD	C@5	64.7 (48.9 + 15.8)%	Not Annotated	89.6 (76.8 + 12.8)%
MSVD	C@10	76.1 (63.9 + 12.2)%	Not Annotated	94.0 (85.4 + 8.61)%
MSVD	AP	44.3 (33.1 + 11.2)%	Not Annotated	71.3 (59.7 + 11.6)%

Table 1: The table shows the impact of FIRE annotations on MSR-VTT and MSVD text-to-video retrieval metrics. “A (B + C)” has metrics computed with FIRE positives (A), only original positives (B), and the delta (C). The deltas emphasize the deleterious effects of false negatives: CLIP4CLIP’s C@1 on MSR-VTT is understated by 25% points.

2.2.2 Construct Validity

In addition to problems with internal validity, we posit that TVR evaluations are also not *construct valid* (Cronbach and Meehl, 1955; O’Leary-Kelly and J. Vokurka, 1998). Construct validity is related to “how closely our evaluations hit the mark in appropriately characterizing the actual anticipated behaviour of the system in the real world or progress on stated motivations and goals for the field” (Raji et al., 2021). What is the real-world use of text-to-video retrieval (or alternatively, the field’s motivations)? Consider the most straightforward answer: that such systems will be used by users to search through video collections, whether on the web or in personal collections. First, search queries issued by real users are very likely not similar to captions written by crowd annotators; this is easily observed by inspecting captions in Table 5 and Appendix Table 6. Second, the video distribution is unlikely to reflect real use-cases as they were selected by annotators or are search results from seed queries. Due to these problems, it seems unlikely that the evaluations are construct valid, and future benchmarks should improve this by building evaluations that match the intended use of models—i.e., be ecologically valid (de Vries et al., 2020).

3 FIRE Dataset Collection and Validation

Next, we describe and analyze the FIRE dataset.

3.1 Annotation Task and Dataset Collection

In the FIRE annotation task, annotators mark whether the displayed caption is relevant to the displayed video. Implicitly, the caption’s video is relevant to it, but how do we judge whether another arbitrary video is relevant? In other words, how should annotators mark whether a caption is relevant to a video? In both datasets (§2.1), the caption

must be completely consistent with the video; otherwise, it would not be an accurate caption. Therefore, we enforce the same condition in our task to preserve the original relevance semantics.⁸

Annotators are instructed to mark a caption as relevant to a video only if every element mentioned in the query could be reasonably considered present. Elements included persons, objects, locations, and activities, as well as quantifiers, qualifiers, and adjectives. Raters are given some leeway to use interpretation and inference but instructed to err in favor of not relevant if the caption is ambiguous or vague. For example, for the caption “a boy playing the violin,” the video must show a boy who is playing the violin, not a video of only violins or a video with only a boy. Screenshots of the annotation interfaces and details of sensitive category handling are in Appendix B. Complete annotation guidelines are included in supplemental materials.

To select caption-video pairs to annotate, we obtain the top ten MSR-VTT and MSVD test set predictions from three models: CLIP4CLIP (Luo et al., 2021), SSB (Patrick et al., 2021), and TeachText (Croitoru et al., 2021). For TeachText, we use model checkpoints available on their webpage. For CLIP4CLIP and SSB, checkpoints are not available, so we train new models and verify that retrieval quality is on par with the literature (see Table 1).

Table 2 summarizes the resulting FIRE dataset. During data collection, 683K labels were collected across a set of 579K unique caption-video pairs. Some duplication was intentional: we obtained a second label for 10% of annotations, and if the labels disagreed, we collected a third label to resolve the disagreement. Elsewhere, duplication was unintentional: for MSVD we did not deduplicate caption-video pairs between two models, so where the pre-

⁸Requiring complete matches makes the annotation task easier by eliminating ambiguous partial match cases.

Dataset	# Pairs	Percent	# Labels
MSR-VTT	24,183	100%	24,507
└ Agreement	24,167	99.9%	-
└ Relevant	2,855	11.8%	-
└ Irrelevant	21,312	88.2%	-
└ Disagreement	16	0.0662%	-
MSVD	555,391	100%	659,126
└ Agreement	553,832	99.7%	-
└ Relevant	39,909	7.21%	-
└ Irrelevant	513,923	92.8%	-
└ Disagreement	1,559	0.281%	-

Table 2: The FIRE dataset is composed of labels for MSR-VTT and MSVD text-video pairs. The positive-to-negative ratio is skewed, reflecting that queries do not match most videos. We multiply annotate a subset to compute annotator agreement rates and Krippendorff’s α . Agreement on MSR-VTT was .931 with $\alpha = .691$ and on MSVD was .958 with $\alpha = .798$. Appendix C disaggregates agreement rates which are consistent.

dictions overlapped, we obtained additional labels. Fortunately, this provided an unexpected opportunity to further validate dataset quality.

3.2 Dataset Quality Validation

Before, throughout, and after the collection, we took steps to collect high-quality data and validate its quality. The annotation task was completed by a team of one hundred raters specifically trained to review caption-video pairs and assess relevance. These annotators completed a 1,000 job training queue, which was reviewed by data quality leads and this paper’s authors. This allowed annotators to learn to annotate according to our guidelines, request clarification to the guidelines, and request tooling improvements. Annotators could also escalate tasks for being too ambiguous or confusing, which occurred less than 0.0001% of the time.

After the dataset was collected, we computed three measures of quality in Table 2: (1) the rate that judgments resolved to a label (Percent), (2) the degree to which examples with multiples labels agreed (Agreement), and (3) the Krippendorff alpha score amongst examples with multiple labels (Krippendorff, 2004). Caption-video pairs resolved to a label 99.9% of the time in MSR-VTT and 99.6% of the time in MSVD. Agreement in both datasets exceeded 90%, and the Krippendorff score suggests reasonable agreement as well. Based on this analysis, we see no evidence of data quality issues. The next section digs deeper into FIRE and suggests explanations for the observed phenomena.

Dataset	Models	Overlap	RBO
MSR-VTT	C4C & SSB	0.0638	0.0568
MSR-VTT	C4C & TT	0.0610	0.0509
MSR-VTT	TT & SSB	0.440	0.231
MSVD	C4C & TT	0.411	0.211

Table 3: Annotated predictions of one model boost the score of another model when predictions overlap. In MSR-VTT, there is little overlap between CLIP4CLIP and other models; there is far more overlap in MSVD.

Model	Data	C@1	C@5	C@10
CLIP4CLIP	All	0.674	0.907	0.957
CLIP4CLIP	New	0.430	0.713	0.812
TeachText	All	0.241	0.532	0.670
TeachText	New	0.239	0.527	0.663
SSB	All	0.273	0.559	0.689
SSB	New	0.271	0.553	0.679

Table 4: We compare C@K of a MSR-VTT model: (1) with all annotations (All) and (2) without the model’s annotated predictions to emulate model development (New). CLIP4CLIP exhibits large differences.

4 Analysis Experiments

The difference FIRE makes on metrics (Table 1) is striking, which begs the question: *why* are there such *large* differences? We suggest explanations for these differences (§4.1) while investigating how these metrics vary under commonplace evaluation settings such as new model development (§4.2).

4.1 Why Are Score Boosts Not Uniform?

FIRE-based metrics are interesting for at least two reasons: (1) the magnitude of difference and (2) the non-uniformity of boosts. Specifically, CLIP4CLIP has a larger boost than TeachText and SSB on MSR-VTT. First, we investigate the degree of prediction overlap between models. When predictions overlap, the models share the boost. Likewise, when they do not overlap, there is an opportunity for differing boosts. Table 3 shows this: on MSR-VTT, CLIP4CLIP and the other two models have little overlap; in contrast, TeachText and SSB have substantial overlap and their boosts are of roughly the same magnitude. Overlap is computed between the top ten predictions of each model using simple overlap and rank-biased overlap (Webber et al., 2010, RBO).⁹ As we might expect based on CLIP4CLIP

⁹If the ordering of predictions amongst the top ten did not matter, the overlap would be acceptable. However, as in most IR settings, we *do* care about the order so use a rank-aware metric like RBO.

and TeachText having large boosts on MSVD, their predictions also overlap. This mechanically explains the difference but fails to explain “why?”

We test the hypothesis that shorter queries have more positives because they are less specific (i.e., general) and speculate that differences in CLIP4CLIP and TeachText pre-training could make CLIP4CLIP fare better on general queries. Intuitively, shorter captions should be less specific and therefore match more videos, so models that handle general captions well should benefit the most. Table 5 and Appendix Table 6 validate this intuition by showing MSVD and MSR-VTT captions. The captions are sampled from the shortest 100 captions, median length captions, and longest 100 captions.

First, we empirically validate that short captions have more positive videos. Figure 2 shows that longer captions have fewer positive videos while shorter captions have more. By construction, since we find only positives if a model predicts them, these are where models make gains.

Figure 3 takes the next step and compares model accuracy as a function of caption length. For each bin of caption lengths (e.g., captions of length zero to twenty characters), we show the proportion of whether both CLIP4CLIP and TeachText are correct, neither are correct, or only one is correct. Empirically, we observe that CLIP4CLIP makes the largest gains from accounting for false negatives with FIRE when queries are short—whether this is due to short queries containing more positives or CLIP4CLIP handling these better is difficult to discern. Although it is difficult to validate, our best, educated guess at a causal reason for CLIP4CLIP finding more positives in MSR-VTT is that its image-text backbone, CLIP (Radford et al., 2021), was trained with text that contains many general captions.

4.2 Does System Pooling Generalize?

Although system pooling eliminates (implicit) false negatives, it comes with the substantial drawback that every new model must have its predictions annotated—otherwise, the results are potentially biased against the new model due to the possibility of false negatives in novel predictions (Yilmaz et al., 2020).¹⁰ System pooling has traditionally been used in synchronized shared tasks where all models are submitted by a deadline and evaluated

¹⁰If a model predicts a video that no prior model does and it is a false negative, then the model’s effectiveness will be underestimated. Yilmaz et al. (2020) study this when comparing traditional and deep learning IR systems.

at the same time, as in the Text REtrieval Conferences (TREC) in IR.¹¹ However, the trend in machine learning and NLP is for continuously running or even dynamic benchmarks (Kiela et al., 2021). Beyond benchmarking, even the development of new models is affected since gains from improved modeling may be understated. The question then is: how large is this bias, and how fast does it decrease with the number of pooled models?

The magnitude of the bias is affected by two factors: (1) the percent of model predictions that do not exist in pooled annotations and (2) the prevalence of false negatives in this subset.¹² While Table 3 captures prediction overlap between pairs of models, it does not capture the setting where some number of models have annotated predictions, and we wish to test a new model. Table 4 calculates (1) model scores when using all annotated predictions versus (2) model scores using only annotated predictions from the other two models. In this small three-model experiment, the bias is unfortunately still significant (24.4% for C@1) for the best model (CLIP4CLIP). Thus, the degree to which the FIRE dataset will mitigate the false negative problem in new model development is dependent on the similarity of new models to current ones. The generalizability also depends on the number of unknown positives, which we indirectly study by plotting the ranks of positive videos (Appendix G).

4.3 Mitigating Annotation Costs by Sampling

A limitation of our method is that until existing annotations include most positives, our method either disadvantages new models or introduces non-trivial annotation costs. Indeed, the costs of exhaustive annotation in our work are substantial, but exhaustive annotation is also excessive if the goal is only to (robustly) estimate model scores. Instead, we propose that future work need only annotate the top 10 predictions from N examples in the evaluation data. But how large should N be so that we can be confident that the difference between model scores is statistically significant? In our next experiment, we use bootstrap sampling to characterize the relationship between N and the effect size corresponding to a statistically significant difference at the 95% confidence level.

In our bootstrap sampling experiment, we treat the 27,763 MSVD test examples as a sample from

¹¹<https://trec.nist.gov>

¹²See Appendix F for analysis of the number of known positive videos per query.

MSVD Short Length Captions	MSVD Median Length Captions
playing panda	a gymnast falls off a balance beam
some work	a person is riding a horse
a man	a girl is riding a bicycle
a baby	two men are pushing an airplane
jumping dachhund	the turtle is playing with the cat
naah	piano is played by an artist
amanplaysaguitar	the girl put stickers on her face
a woman	a boy is reading a card
camp	a little boy is playing golf
plying music	a man is slicing a tomato
MSVD Long Length Captions	
a man holding an open umbrella jumps across a wooden stand in a park and then does a summersault after kicking a wall	
a man in a jail cell motions to a man in another cell who shows the first man his middle finger	
a bowling man picks up a spare in his lane and manages to knock over the one remaining pin in the lane to his right	
a woman is exercising by stepping from right to left and then from left to right while swinging her arms back and forth	
a man wearing a black cape is walking toward a group of people and a man in the group is shooting at him with a pistol	

Table 5: This table shows three sets of MSVD captions sampled from: (1) the 100 shortest captions, (2) median length captions, and (3) the 100 longest captions. As also observed in MSR-VTT captions (Table 6), short captions are general (e.g., “a man”) compared to the longest captions.

a population.¹³ We characterize the population distribution through bootstrap re-sampling of the original sample. Specifically, we estimate the absolute difference in model scores that correspond to a statistically significant effect size (i.e., score difference) at the 95% confidence level. For each sample size $N \in [500, 1000, 3000]$, we (1) re-sample N examples from MSVD evaluation data, (2) calculate scores on the re-sample, (3) repeat this 10,000 times, (4) average the scores then calculate the absolute value of the difference between the average score and score calculated with the full dataset, and finally (5) plot the distribution and score corresponding to the 95% percentile. The experimental results (Figure 4) demonstrate that annotation volumes of 1,000 detect statistically significant differences when C@1 differs by 0.029. The results demonstrate that (1) annotating a subset of test examples detects absolute differences of one absolute point, and (2) the number of annotated test predictions varies based on the metric of interest.

5 Recommendations for Benchmarks

Towards improving TVR evaluations, we make recommendations for both current and future benchmarks. This paper only investigated the effects of false negatives in MSR-VTT and MSVD. However, it is likely that other similarly constructed benchmarks exhibit the same problem, and testing this is important. Second, we show that for MSR-VTT

¹³MSR-VTT is small. To avoid convergence to the sample mean, bootstrap sizes need to stay low.

and MSVD, certain metrics such as Correct@10 are potentially saturated since improvements above CLIP4CLIP’s 95.7% and 94% are plausibly noisy. Consequently, since the remaining gains reside in re-ranking the top K, the community should consider retiring these evaluations. Third, the introduction of multiple positives and use of various K values makes mean average precision attractive since: (1) it factors in preference for correctness at higher ranks and (2) it handles multiple positives.

It is difficult to recommend that model developers exhaustively annotate model predictions. This suggests a future where query or video set size is a trade-off between annotation load and evaluation quality. For example, one might choose to trade-off annotation load with statistical power to differentiate between models (Card et al., 2020). TREC-style, annual shared tasks are one model for this (Voorhees, 2019; Church and Hestness, 2019); instead of building a monolithic benchmark that becomes overfit over time (Blum and Hardt, 2015; Anderson-Cook et al., 2019), stakeholders develop evaluations that evolve with research objectives.

Looking forward, TVR evaluation would benefit from: (1) a purpose-built benchmark that is grounded in an actual use case so as to be ecologically valid (de Vries et al., 2020) and (2) centralized evaluation by submitting runnable models to shared infrastructure such as Dynabench (Kiela et al., 2021). This would improve reproducibility, which was a limiting factor in selecting which model predictions to reproduce in this paper. This

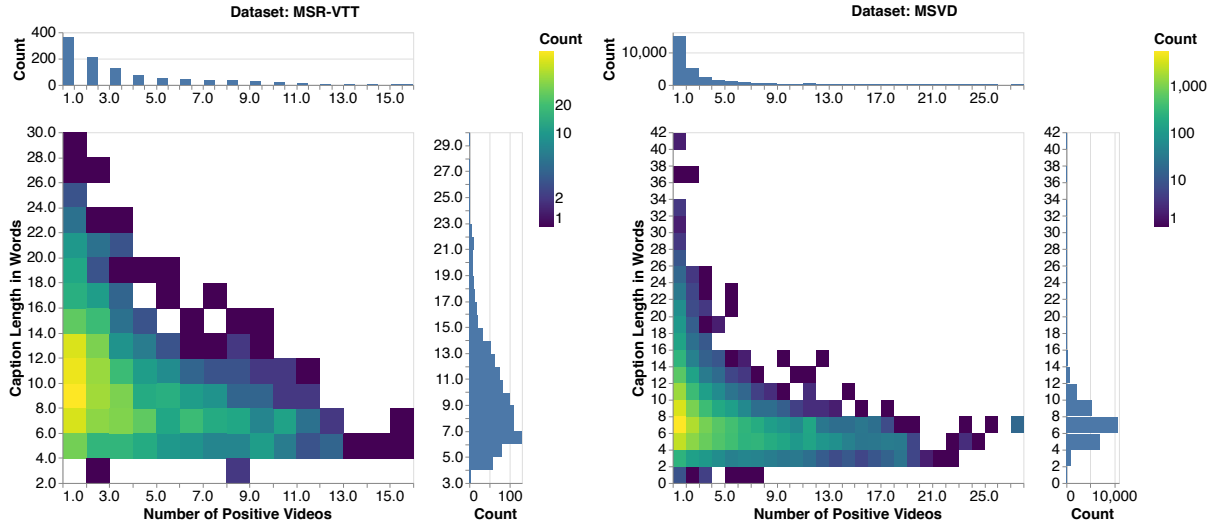


Figure 2: This figure shows the relationship between the number of positive videos and the length of captions in words for MSR-VTT and MSVD. We show a log-scale density heatmap binned by the number of positive videos and caption length; on the margins, are histograms. From this figure, we can infer that: (1) if a caption is long, it is less likely to have many positive videos, and (2) if a caption is short, then the number of positive videos can vary widely.

also makes calculating statistical tests easy (Ethayarajh and Jurafsky, 2020), which are often not reported (Dror et al., 2018; Dodge et al., 2019). TVR modeling has advanced enough to demand better benchmarks for measuring future progress.

6 Related Work

The paper draws on ideas in multimodal retrieval, information retrieval, and evaluation methodology.

Improving Benchmark Quality: Wray et al. (2021) is directly relevant to our work, and we share their motivation: to study the effects of false negatives in TVR evaluations. While we share motivation and our works are complementary, our work differs substantially in methods, contributions, and conclusions. The primary difference is this: our goal is to quantify the difference in absolute metrics that false negatives cause, even if there is no promise the data can be effectively reused in the future; Wray et al. (2021) develop automatically runnable proxy measures that improve the reliability of model rankings, but do not precisely quantify the impact of false negatives on existing metrics since automatic labeling is not equivalent to human annotation. Both these works are valuable: our work conclusively quantifies that false negatives create differences of 25% absolute points and demonstrate that new measures like those by Wray et al. (2021) are necessary for current benchmarks.

Wang et al. (2022) argue that video captioning datasets used in TVR evaluation are noisy due to

low-quality captions but differ by identifying single query tasks as the root problem (as opposed to false negatives) and recommend multi-query evaluation where users make followup refinement queries. While the multi-query problem is important, we do not agree with the assessment that single-query problems should be abandoned for multi-query problems: for example, users often have a low tolerance for voice assistant errors and abandon their query entirely after an error. Both problems are important. Fortunately, the approaches are complementary and should be combined: the multi-query setting still has false negatives, whose effects on measurement can be mitigated with our methods (§4.3). Just as we use predictions to improve datasets, Beyer et al. (2020) improve ImageNet labels by using predictions to reduce the label space which makes the annotation task easier.

Benchmarking: Across machine learning, computer vision, and natural language processing (Eger et al., 2020; Bowman and Dahl, 2021; Rogers, 2021) there is a broad effort to critically examine the benchmarks (Schlangen, 2021), data (Linzen, 2020; Thrush et al., 2022), evaluation methods (Rodriguez et al., 2021), and evaluation paradigms (Rodriguez and Boyd-Graber, 2021; Kiela et al., 2021) used in research studies. This effort goes beyond particular methodologies and extends to identifying the values prized by the community (Sculley et al., 2018; Dotan and Milli, 2020) which are subsequently operationalized in computer vision

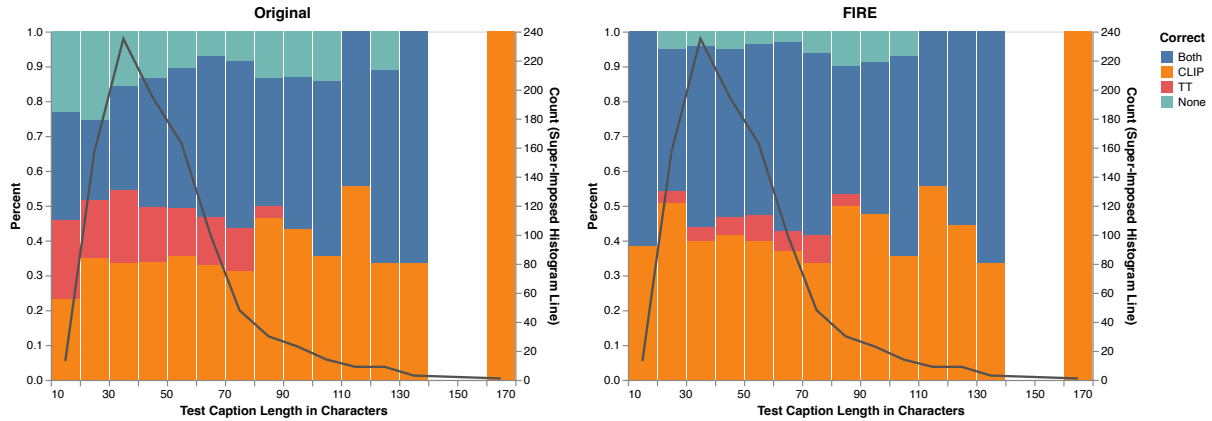


Figure 3: On MSR-VTT, we show relative model effectiveness differences (y-axis and color bars) broken down by test caption length (x-axis); we super-impose the caption length distribution (black line). Short captions tend to be more general, so they should match more videos and produce more false negatives. The gains for both models and especially CLIP4CLIP occur predominantly on this subset (reduction of “None”) as we would expect.

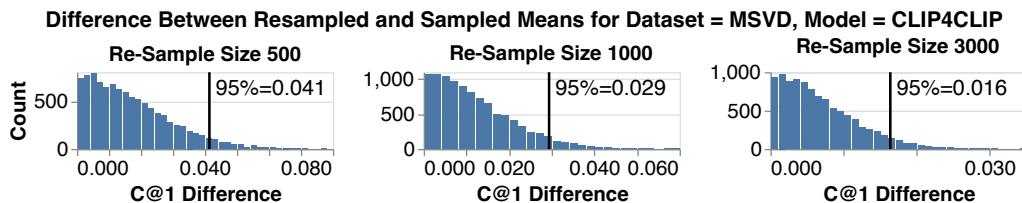


Figure 4: The distribution of absolute differences between bootstrap re-sample estimates of CLIP4CLIP C@1 and the true sample mean, by re-sample size. This estimates the number of annotations to detect an effect size at 95% confidence. Appendix E expands on this experiment by showing results for C@5, C@10, and TeachText.

datasets and benchmarks (Wu et al., 2017; Scheuerman et al., 2021). Our work is in line with this broader initiative and critically examines text-to-video retrieval evaluation methodology.

We examine internal validity (§2.2.1) and find a broken yardstick (Hernandez-Orallo, 2020). By examining construct validity (§2.2.2), we also argue that TVR evaluations should prize usefulness to ecologically valid use cases such as real-world text-to-video search (de Vries et al., 2020). Lastly, our experimental results suggest we may not be far off from retiring MSR-VTT and MSVD for TVR evaluation, something we should not be afraid to do in general (Boyd-Graber and Börschinger, 2020). An alternate approach is smaller, periodic evaluations as in TREC (Smeaton et al., 2002; Voorhees and Tice, 2000; Smeaton et al., 2009). Part of the solution is to create purpose-built datasets with clear goals (Gebru et al., 2021; Bender and Friedman, 2018) as opposed to continually re-using datasets intended for different uses (Koch et al., 2021).

Structurally Similar Tasks: TVR is not the only evaluation with the implicit false negative problem. Our critique is applicable to image retrieval bench-

marks that use caption-media pairs from image captioning datasets (Lin et al., 2014; Plummer et al., 2015) as the only positives (Karpathy and Fei-Fei, 2015; Kim et al., 2021; Singh et al., 2022).

7 Conclusion

In this work, we show that label errors (false negatives) in text-to-video retrieval benchmarks invalidate their *internal validity*—the measured metrics do not accurately reflect reality (§2). Following this, we critique the applicability of benchmarks to real-world use cases (*construct validity*). To estimate the impact of false negatives on benchmark metrics, we collect the FIRE dataset (§3) which contains 683K relevance judgements. Analysis experiments (§4) suggest explanations for why CLIP4CLIP scores higher and estimate system pooling generalization. Based on our findings, we highlight properties that future TVR benchmarks should have and outline approaches to addressing inherent challenges in retrieval evaluation (§5). Finally, we position our work in the broader effort to improve benchmarking by better aligning tasks with the intended use and improving measurement (§6).

8 Limitations

Our work has several notable limitations. First, our experiments use two representative and commonly used TVR datasets (MSR-VTT and MSVD). While we expect that our results will generalize, it is still possible that these results do not generalize. For example, both datasets are based on YouTube videos and annotator-written captions: perhaps videos and captions from alternate sources differ by too much. Similarly, our experiments use three well-known models, so while we expect our results to generalize to similar models, future models may differ substantially in ways that cause the empirical results not to hold. This said, system pooling has long been used in TREC (Voorhees et al., 2005), so we expect this to work for future models as well.

Beyond limitations in generalizability, the in-principle critiques in our work apply only to benchmarks where implicit false positives are likely to be prevalent; it does not apply to benchmarks in general. From the methods perspective, while our computational experiments are coded to be easily reproduced, the scale of our annotations is difficult to reproduce (hence limited reproducibility in this sense), but we do study sampling-based alternatives to mitigate this limitation.

9 Ethics

This section discusses potential ethical issues related to our dataset-centric work. First, we discuss data-related ethics. The FIRE dataset is built on MSR-VTT and MSVD. We distribute the minimal amount of data related to these datasets necessary to reproduce our experiments: triplets of caption identifiers, video identifiers, and annotated labels. Section 3 and Appendix B describe how the data was collected. All annotators were compensated, and the data collection was reviewed before starting. Potential risks due to the use of our dataset are limited by the additional labels we provide for an existing dataset. We thoroughly discuss the risks associated with negatively influencing benchmark reliability (i.e., prediction overlap with future models), and these risks are mitigated by our recommendation that more appropriate datasets be developed.

Our work does not directly have negative societal impacts, but it is feasible that the improved model scores we report could be used to misrepresent the capability of retrieval systems. For example, while we only claim that a model achieves a

particular measure of effectiveness on a particular benchmark, the media often inflates the importance of these metrics (Cuthbertson, 2018). In our work, we intentionally do not connect these higher metrics to more general capability and emphasize the importance of establishing construct validity.

Acknowledgements

We thank Yookoon Park, Prahal Arora, and Bernie Huang for providing code and data that were helpful to kickstart this project. We thank Daniel Haziza for infrastructure support in hosting live demos. We thank Nathan Tokala for their support with annotation infrastructure. For insightful discussion and ideas, we thank Simran Motwani, Patrick Lewis, Thomas Hayes, Joe Barrow, Xilun Chen, Chenglei Si, Ronghang Hu, Max Bain, and Jacob Kahn. For feedback on prior versions of this paper, we thank Rich James, Florian Metze, Peter Rankel, Weijia Xu, Yoo Yeon Sung, Yuandong Tian, John P. Lalor, and Kirmani Ahmed.

References

- Christine M Anderson-Cook, Kary L Myers, Lu Lu, Michael L Fugate, Kevin R Quinlan, and Norma Pawley. 2019. [How to host an effective data competition: Statistical advice for competition design and analysis](#). *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(4):271–289.
- Emily M Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. [Are we done with ImageNet?](#) *arXiv preprint arXiv:2006.07159*.
- Avrim Blum and Moritz Hardt. 2015. [The ladder: A reliable leaderboard for machine learning competitions](#). In *Proceedings of the International Conference of Machine Learning*. PMLR.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W Rae, Erich Elsen, and Laurent Sifre. 2021. [Improving language models by retrieving from trillions of tokens](#). *arXiv preprint arXiv:2112.04426*.

- Samuel Bowman. 2022. [The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Samuel R Bowman and George E Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Jordan Boyd-Graber and Benjamin Börschinger. 2020. [What question answering can learn from trivia nerds](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Marilynn B. Brewer and William D. Crano. 2014. *Research Design and Issues of Validity*, 2 edition, page 11–26. Cambridge University Press.
- D T Campbell. 1957. [Factors relevant to the validity of experiments in social settings](#). *Psychological bulletin*, 54(4):297–312.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- David Chen and William Dolan. 2011. [Collecting highly parallel data for paraphrase evaluation](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Kenneth Ward Church and Joel Hestness. 2019. [A survey of 25 years of evaluation](#). *Natural Language Engineering*, 25(6):753–767.
- Ioana Croitoru, Simion-Vlad Bogolin, Yang Liu, Samuel Albanie, Marius Lordeanu, Hailin Jin, and Andrew Zisserman. 2021. [TeachText: Crossmodal generalized distillation for text-video retrieval](#). *International Conference on Computer Vision*.
- Lee Joseph Cronbach and Paul E. Meehl. 1955. [Construct validity in psychological tests](#). *Psychological bulletin*, 52 4:281–302.
- Anthony Cuthbertson. 2018. [Robots can now read better than humans, putting millions of jobs at risk](#). *Newsweek*.
- Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. [Towards ecologically valid research on language user interfaces](#). *arXiv preprint arXiv:2007.14435*.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [Show your work: Improved reporting of experimental results](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Ravit Dotan and Smitha Milli. 2020. [Value-laden disciplinary shifts in machine learning](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors. 2020. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.
- Kawin Ethayarajh and Dan Jurafsky. 2020. [Utility is in the eye of the user: A critique of NLP leaderboards](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. [Multi-modal Transformer for Video Retrieval](#). In *European Conference on Computer Vision*.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. [ActivityNet: A large-scale video benchmark for human activity understanding](#). In *Computer Vision and Pattern Recognition*.
- Jose Hernandez-Orallo. 2020. [AI evaluation: On broken yardsticks and measurement scales](#). In *Workshop on Evaluating Evaluation of Ai Systems at AAAI*.
- Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. [A survey on visual Content-Based video indexing and retrieval](#). *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society*, 41(6):797–819.
- Andrej Karpathy and Li Fei-Fei. 2015. [Deep visual-semantic alignments for generating image descriptions](#). In *Computer Vision and Pattern Recognition*, pages 3128–3137.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit

- Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. [Vilt: Vision-and-language transformer without convolution or region supervision](#). In *Proceedings of the International Conference of Machine Learning*.
- Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. [Reduced, reused and recycled: The life of a dataset in machine learning research](#). In *NeurIPS: Datasets and Benchmarks Track*.
- Klaus Krippendorff. 2004. *Content Analysis: an Introduction to its Methodology*. Sage: Thousand Oaks, CA. Chapter 11.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. [Dense-captioning events in videos](#). In *International Conference on Computer Vision*.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. [Content-based multimedia information retrieval: State of the art and challenges](#). *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. [Question and answer Test-Train overlap in Open-Domain question answering datasets](#). In *Proceedings of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. [HERO: Hierarchical encoder for Video+Language omni-representation pre-training](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *NeurIPS: Datasets and Benchmarks Track*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. [Microsoft COCO: Common objects in context](#). In *European Conference on Computer Vision*. Springer International Publishing.
- Tal Linzen. 2020. [How can we accelerate progress towards human-like linguistic generalization?](#) In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. [UniVL: A unified video and language Pre-Training model for multimodal understanding and generation](#).
- Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. [CLIP4Clip: An empirical study of clip for end to end video clip retrieval](#). *arXiv preprint arXiv:2104.08860*.
- Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.
- Sandra Mathison. 2004. *Encyclopedia of evaluation*. Sage publications.
- Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.
- Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. [Pervasive label errors in test sets destabilize machine learning benchmarks](#). In *NeurIPS: Datasets and Benchmarks Track*.
- Scott W O’Leary-Kelly and Robert J. Vokurka. 1998. [The empirical assessment of construct validity](#). *Journal of Operations Management*, 16(4):387–405.
- Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metze, Gourab Kundu, and Kirmani Ahmed. 2022. [Normalized contrastive learning for text-video retrieval](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. 2021. [Support-set bottlenecks for video-text representation learning](#). In *Proceedings of the International Conference on Learning Representations*.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in python](#). *Journal of Machine Learning Research*, 12(85):2825–2830.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. [Flickr30k entities: Collecting Region-to-Phrase correspondences for richer Image-to-Sentence models](#). In *International Conference on Computer Vision*.

- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the International Conference of Machine Learning*. PMLR.
- Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. [AI and the everything in the whole wide world benchmark](#). In *NeurIPS: Datasets and Benchmarks Track*.
- Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Pedro Rodriguez and Jordan Boyd-Graber. 2021. [Evaluation paradigms in question answering](#). In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Anna Rogers. 2021. [Changing the world by changing the data](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Lawrence A Rowe and Ramesh Jain. 2005. [ACM SIGMM retreat report on future directions in multimedia research](#). *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(1):3–13.
- Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. [Do datasets have politics? disciplinary values in computer vision dataset development](#). *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–37.
- David Schlangen. 2021. [Targeting the benchmark: On methodology in current natural language processing research](#). In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.
- D Sculley, Jasper Snoek, Alexander B Wiltschko, and A Rahimi. 2018. [Winner’s curse? on pace, progress, and empirical rigor](#). In *Proceedings of the International Conference on Learning Representations*.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. [FLAVA: A foundational language and vision alignment model](#). In *Computer Vision and Pattern Recognition*. IEEE Computer Society.
- Alan F Smeaton, Paul Over, and Wessel Kraaij. 2009. [High-Level feature detection from video in TRECVID: A 5-year retrospective of achievements](#). In Ajay Divakaran, editor, *Multimedia Content Analysis: Theory and Applications*, pages 1–24. Springer US, Boston, MA.
- Alan F Smeaton, Paul Over, and Ramazan Taban. 2002. [The TREC-2001 video track report](#). Technical report, National Institute of Standards and Technology.
- Karen Spark-Jones. 1975. Report on the need for and provision of an ‘ideal’ information retrieval test collection. *Computer Laboratory*.
- Wanhua Su, Yan Yuan, and Mu Zhu. 2015. [A relationship between the average precision and the area under the ROC curve](#). In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. Association for Computing Machinery.
- Jean Tague-Sutcliffe. 1992. [The pragmatics of information retrieval experimentation, revisited](#). *Information processing & management*, 28(4):467–490.
- Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. 2022. [Dynatask: A framework for creating dynamic AI benchmark tasks](#). In *Proceedings of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.
- Ellen M. Voorhees. 2019. [The Evolution of Cranfield](#), pages 45–69. Springer International Publishing, Cham.
- Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, MA.
- Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky. 2022. [Multi-query video retrieval](#). In *European Conference on Computer Vision*.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. [A similarity measure for indefinite rankings](#). *ACM Transactions on Information Systems*, 28(4).
- Michael Wray, Hazel Doughty, and Dima Damen. 2021. [On semantic similarity in video retrieval](#). In *Computer Vision and Pattern Recognition*.
- Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. 2017. [Deep Learning for Video Classification and Captioning](#), pages 3—29. Association for Computing Machinery and Morgan & Claypool.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. [VLM: Task-agnostic Video-Language model pre-training for video understanding](#). In *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics.

- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. [MSR-VTT: A large video description dataset for bridging video and language](#). In *Computer Vision and Pattern Recognition*.
- Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. 2020. [On the reliability of test collections for evaluating systems of different types](#). In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA.
- Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. [A joint sequence fusion model for video question answering and retrieval](#). In *European Conference on Computer Vision*.
- Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. [Cross-modal and hierarchical modeling of video and text](#). In *European Conference on Computer Vision*.
- Linchao Zhu and Yi Yang. 2020. [Actbert: Learning global-local video-text representations](#). In *Computer Vision and Pattern Recognition*.

A Model Prediction Comparison

As part of this paper, we develop several web apps to make exploring the data more accessible. For example, Figure 5 compares the predictions of three models along with the labels in the original MSR-VTT dataset compared to augmenting them with FIRE’s labels. The source code repository provides instructions to run these web app demos.

B Annotation Interfaces

The FIRE dataset (§3) was collected using the annotation interface in Figure 6.

In addition to the previously described annotation instruction (§3.1), raters were also instructed on how to handle sensitive categories. The raters were instructed to accept the caption as accurate unless they had compelling, concrete reasons to believe otherwise (e.g., a little baby should be not considered old, and octogenarians with white hair and wrinkled skin should not be considered young); raters should not attempt to make more fine-grained distinctions. In particular, they were instructed not to make any assumptions about gender and accept the gender described by the caption.

C FIRE Data Quality

This section provides additional evidence to validate the quality of the FIRE dataset. Specifically, Figure 7 complements the agreement metrics computed in §3.2 and Table 2 by un-aggregating agreement rates.

D Shorter Captions, Their Generality, and Correlation to Model Behavior

Experiments in §4.1 establish that shorter captions have more positives and longer captions have fewer. We intuitively explain this by stating that shorter captions by nature are less specific, so will, in principle, match more videos. For example, one of the shortest captions in MSVD is “a man” (Table 5) which is less specific than one of the longest captions like “a man holding an open umbrella jumps across a wooden stand in a park and then does a summersault after kicking a wall.” Inspecting these captions also validates our construct validity critique (§2.2.2): they do seem like search queries.

In previous experiments (§4.1), we discussed how caption length is related to which models gain higher boosts. This section breaks down which models gain the most on MSR-VTT by train-test

overlap. We take inspiration from question answering and language modeling, where unintentional textual overlap between train and test sets degrades evaluation and model quality (Lewis et al., 2021; Borgeaud et al., 2021; Lee et al., 2022). Our objective is to measure the degree to which test captions in MSR-VTT are present in the training captions—be it word-for-word or approximate. To measure this, we use Scikit-Learn (Pedregosa et al., 2011) to fit a 5-gram character TF-IDF encoder to the test captions and compute the cosine similarity of each test caption to each train caption. For each test caption, we compute the mean similarity of the top train ten captions and combine this information with Correct@5 scores (Figure 8). The results suggest that TeachText overfits the train set, which may explain its comparatively better scores on the original positives—were it not overfit, train-test overlap should not matter.

Lastly, these factors are not unrelated. Since shorter captions tend to be less specific, these are also the captions that we would expect are more prevalent in the training set, whether in exact form or approximate (e.g., the phrase “a man” is likely in the train set). To test whether these factors are related, we compute the Kendall Tau correlation and Spearman Rank correlation between the train-test textual similarity score and caption length (in both words and characters). As we expect, there is a non-trivial negative correlation between caption length and similarity score (Table 7): the lower the caption length, the higher the train-test overlap score.

E Can Annotation Costs be Mitigated Through Sampling?

In our experiments, we use bootstrap sampling to estimate the number of example annotations needed to detect given effect sizes at the 95% confidence level (§4.3). Figure 4 reports these results for CLIP4CLIP since it was the best model; in practice, it represents the type of model we would test after models like TeachText. Figure 9 extends the results from C@1 to C@5 and C@10. Figure 10 replicates these results, but using the TeachText model.

F Number of Positive Videos per Text Query

The generalization of the FIRE dataset to newer models is reliant on two factors: (1) the number of positive videos per query and (2) whether the

Model Comparison Viewer for MSRVT/MSVD

Original Dataset Label, FIRE Dataset Label

Caption: cartoon girl is talking

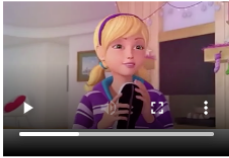
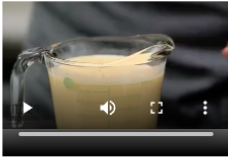
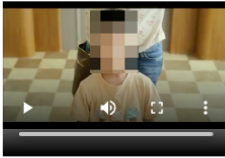
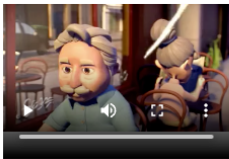


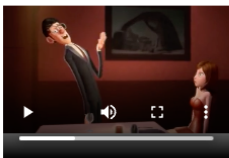

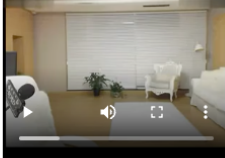
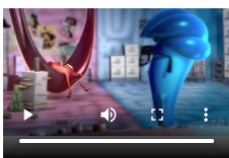
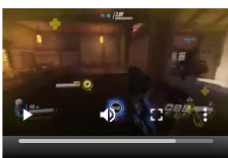
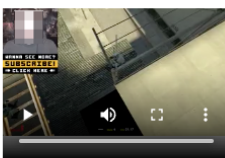

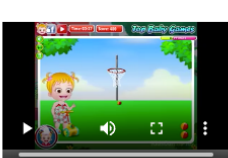

Clip4Clip	Experts	SSB
 <p>Label: <input type="checkbox"/> Irrelevant Label: <input checked="" type="checkbox"/> Relevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>
 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>
 <p>Label: <input type="checkbox"/> Irrelevant Label: <input checked="" type="checkbox"/> Relevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>
 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>
 <p>Label: <input checked="" type="checkbox"/> Relevant Label: <input checked="" type="checkbox"/> Relevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input checked="" type="checkbox"/> Relevant</p>	 <p>Label: <input type="checkbox"/> Irrelevant Label: <input type="checkbox"/> Irrelevant</p>

Figure 5: The web application shows the ranked predictions of three models: CLIP4CLIP, TeachText, and SSB. Qualitatively, CLIP4CLIP predictions better match the query by showing only cartoon videos. This is reflected quantitatively when FIRE labels are incorporated. Lastly, the ranked predictions also show some of the overlap that TeachText and SSB shared.

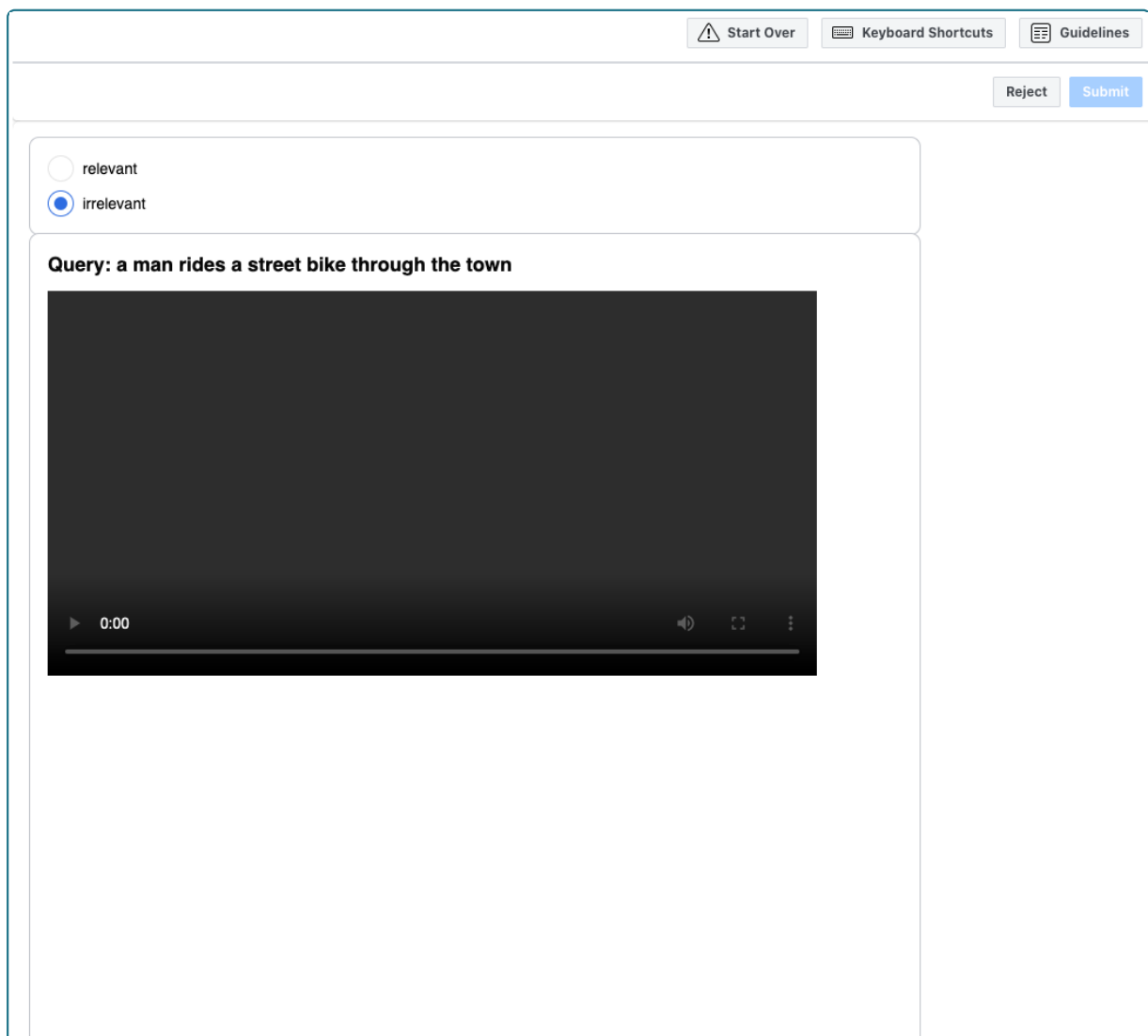


Figure 6: To annotate the FIRE dataset, raters used this annotation interface. The interface shows the candidate query (caption) and video; raters are trained to select “relevant” or “irrelevant” based on whether every component of the query matches the video.

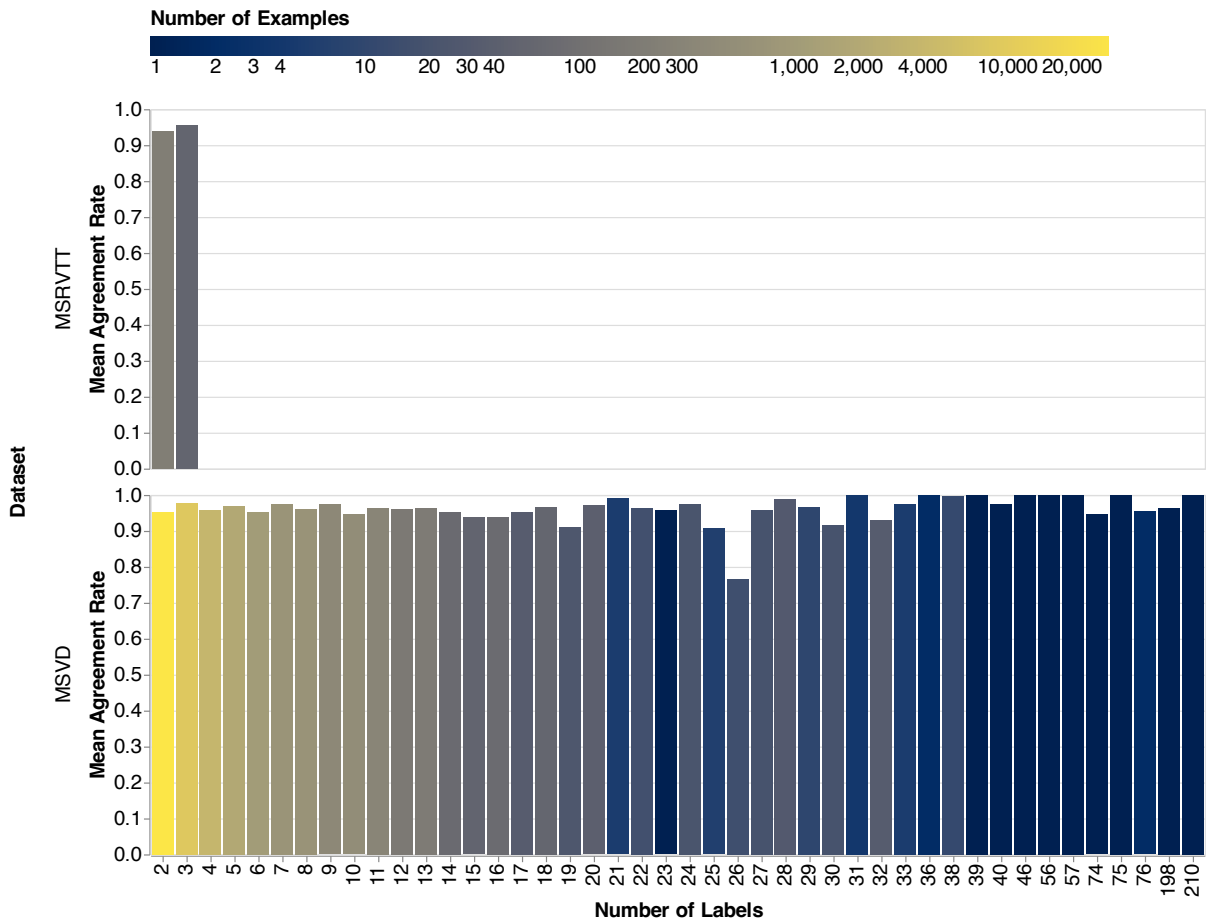


Figure 7: The agreement rate of annotators is broken down by the number of labels. For example, about 10,000 MSVD examples (text-video pairs) were annotated twice; of those, the two labels agreed on about 95% of examples. As we did with the MSR-VTT collection, our intent for the data collection was to de-duplicate text-video pairs and only annotate about 10% of the data multiple times to estimate reliability. However, we accidentally omitted this step for the MSVD collection which resulted in some examples being annotated many times. Fortunately, this provides an unplanned opportunity to further validate inter-annotator agreement.

MSR-VTT Short Length Captions

a man playing video games
anchor talking about a shows
a woman is stirring food
sports are being played
a woman holding a ribbon
a diver goes underwater
baseball player hits ball
cartoon show for kids
two women are embracing
advertisement of seat basket

MSR-VTT Median Length Captions

a man runs into the crowd when trying to catch a basketball
in a music video a man is laying with women while singing
some people video conferencing as they watch a movie
a boy is trying out for a part on the voice kids
basketball players making a shot in the last seven seconds
views of two persons working on the super computer with the head phones on
a character is jumping and floating in the air in a video game
two people playing basketball and the one with a hat makes every shot
batman is beating up bane in a scene from a batman movie
a girl being surprised with a stuffed animal by male friend

MSR-VTT Long Length Captions

a man and a woman are sitting in front of a television and addressing and audience
a woman stirs up some soup sprinkles a spice in and drops a shot of liquid into it
a man is filming as he and a woman watch the news where it shows an area filled with smoke
flight is shaken and the pilots trying to land the flight while they opened the air
the chef adds fish sauce and fish paste to a large stainless steel cooking pot
a girl wearing a dress stands to the side of the screen while lyrics to a song playing in the background appear on the other side
the man is giving an informational speech to a group of people about telling someone something
a girl in blue color dress wearing sitting speaking and television screen with black shirt man beside still image displaying on screen
a man plays a video game where the player has a first person perspective and shoots other characters
a man playing a video game character that is carrying a sword and killing animals with it

Table 6: This table shows three sets of captions from MSR-VTT sampled from: (1) the 100 shortest captions, (2) the median length captions, and (3) the 100 longest captions. As we argued by intuition (§4.1), inspecting these samples validates that the shorter captions are more general (e.g., “sports are being played”) and longer captions are very specific (e.g., “a woman stirs up some soup sprinkles a spice in and drops a shot of liquid into it”).

models we studied in this work predict all the true positives. Estimating the number of true positives per query without exhaustive annotation is difficult at best. However, we can at least characterize how many positives there are when including FIRE annotations. Figure 11 shows a histogram of the number of positive videos per query across MSR-VTT and MSVD. For example, about 350 MSR-VTT queries have only one known positive, which implies that the other 650 have more than one known positive. Unfortunately, even estimating the upper bound would require annotating all the videos for each query in a representative sample (e.g., for a sample of MSR-VTT 200 queries, exhaustive annotation would include $200 * 1,000 = 200,000$ query-video annotations).

G Rank of Positive Videos

While we follow prior work in measuring models based on metrics computed from their top ten predictions, this still leaves open the question: ignoring prior work, is ten predictions the right choice? If ten is the correct choice, then we should see a clear trend that positives are primarily distributed below ten. Figure 12 plots the rank of positive videos in CLIP4CLIP predictions (i.e., 1 is top-ranked) versus their count. As expected, the number of positives drops dramatically before rank 10 (especially for MSVD), although not to zero; the ranks of the original positives suggest there is a long tail of undiscovered positives. Note that the steep dropoff at 10 is due to annotating only the top 10; positives beyond this are either from the original dataset or predicted by other models. From

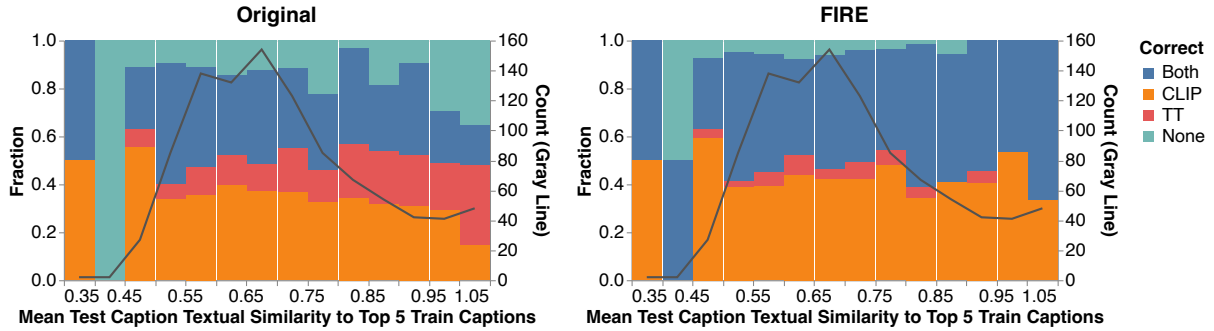


Figure 8: Why are the MSR-VTT score differences between CLIP4CLIP and TeachText when using FIRE large? We test the hypothesis that TeachText (comparatively) overfits textual training data. We compute the textual similarity of each test caption to each train caption with a 5-gram character model; for each test caption, we calculate the mean similarity of the ten most similar captions. The plot shows whether both models score a point on Correct@5, binned by train-test similarity (the overall histogram is shown as the super-imposed line) when using original versus FIRE annotations. On the original annotations, CLIP4CLIP fares much better compared to TeachText when similarity is not nearly 1.0 (i.e., not overfitting).

Length	Spearman	Kendall
Word	-0.419	-0.296
Character	-0.479	0.334

Table 7: This table shows the Spearman and Kendall rank correlations between the train-test textual similarity score used to measure train-test overlap (Figure 8) and the length of captions in both words and characters. The results support our hypothesis that caption length and train-test overlap are correlated.

this, we conclude that although most positives have likely been collected, there likely remain more past rank 10, especially in MSR-VTT.

H Computational Resources

This paper was developed using two types of computational resources. To rerun text-to-video retrieval models, we trained and evaluated on a single AWS p4d compute node which has 96 vCPUs, 1152GB of RAM, and eight Nvidia A100 GPUs.¹⁴ All other experiments were run locally on a 16 inch, 2019 Macbook Pro with a 2.4GHz 8-core Intel Core i9 CPU and 32GB of RAM.

¹⁴<https://aws.amazon.com/ec2/instance-types/p4/>

Absolute Difference Between Resampled and Sampled Metric Means for Dataset = MSVD, Model = CLIP4CLIP

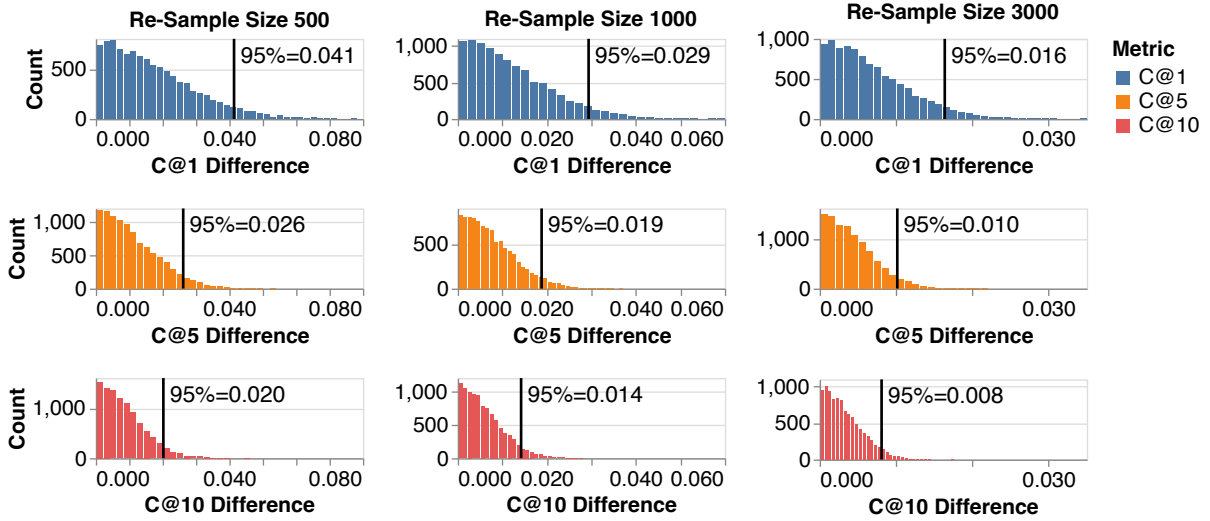


Figure 9: This figure replicates the C@1 results from Figure 4 but adds results for C@5 and C@10. The additional results are consistent in showing that differences of about 1 point are already detectable with 1,000 annotations.

Absolute Difference Between Resampled and Sampled Metric Means for Dataset = MSVD, Model = TeachText

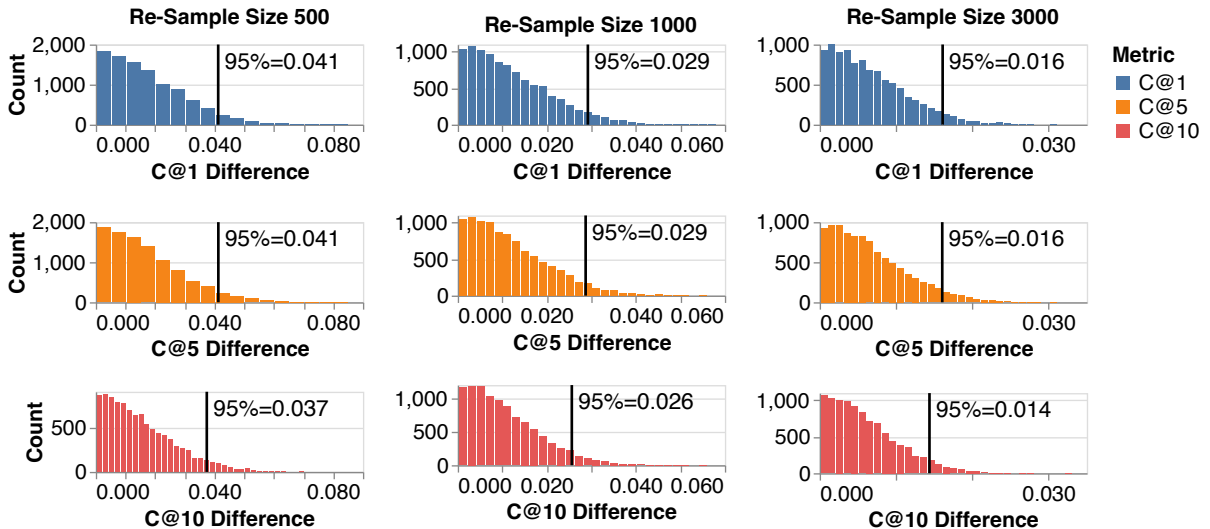


Figure 10: Similar to Figure 4, this figure shows the distribution of absolute differences between bootstrap re-sample estimates of TeachText C@1, C@5, and C@10 scores and their true sample mean (i.e., scores on the full test set). Compared to CLIP4CLIP, statistically significant differences are marginally harder to detect.

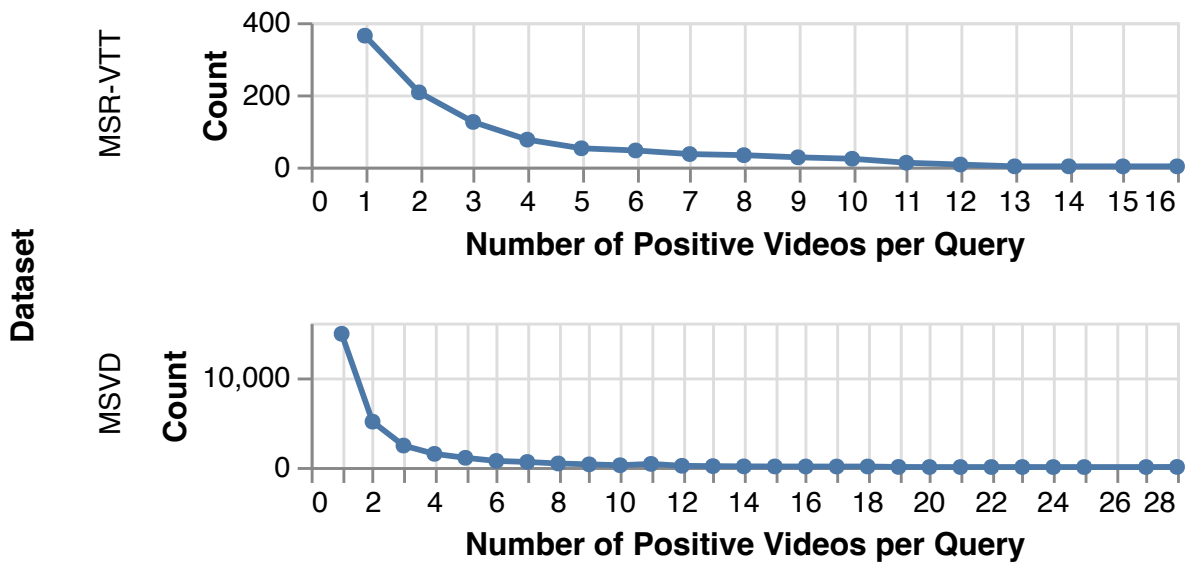


Figure 11: For MSR-VTT and MSVD, we plot the number of positive videos per test set query. While many queries across both datasets have only one known positive, many others have more than that.

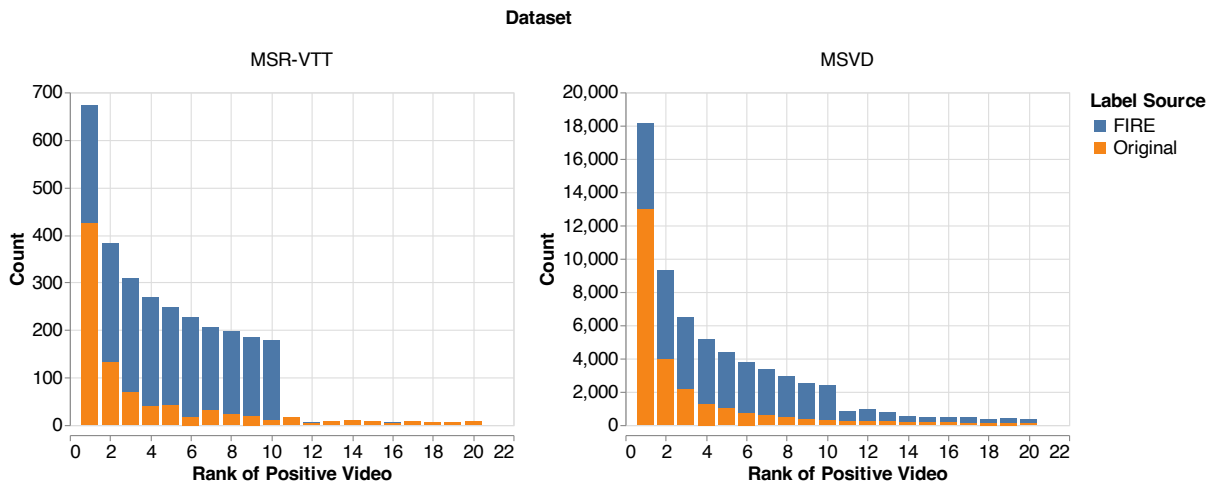


Figure 12: The figure plots the rank of positive video predictions from CLIP4CLIP versus their count. The plot displays MSR-VTT and MSVD separately, and it breaks down the source of each positive (from the original dataset versus from FIRE). While the distribution suggests most positives are found within the top 10, the long tail suggests that there are still unknown positives.