

Appendix

VTC: Improving Video-Text Retrieval with User Comments

Laura Hanu¹, James Thewlis¹, Yuki M. Asano², and Christian Rupprecht³

¹ Unitary Ltd.

² University of Amsterdam

³ University of Oxford

1 Appendix

1.1 Qualitative Examples

In Figure 1 we adapt the text branch, similar to Fig. 1 of the main paper. The example in the second row of Figure 1 shows how our Context Adapter Module can leverage the comments to learn that the content is indeed about parrots, as opposed to dogs. The fourth row shows that without comments, the title alone can be extremely ambiguous while comments can again guide the model to retrieve relevant videos of drumming.

We provide examples where the video branch has been adapted in Figure 2. In most cases, the retrieved titles are broadly related to the video thumbnails. However, when provided with the comments, the retrieved titles become more specific to the videos. For example, in the example from the second row of a screenshot from Mario Kart, the retrieved titles are generally about games e.g. The Castle or Sun Haven, whereas when adapting the video with the comments, the model retrieves titles specifically about Mario Kart. Similarly, in the example from the last row, the model seems to get confused about the content of the video when deprived of the comments, which provide the necessary context about feeding a fish.

Finally, in Figure 3, we show the saliency of comments with regards to a given video and title. For this, we use the approach of masking out each comment in turn, allowing us to visualise the effect of each individual comment on the network output. We compare the output descriptor when including all comments to the descriptors with a comment masked, using the inner product as a score of similarity, and present the comments sorted from lowest to high, the expectation being that an uninformative comment will not cause a large shift in the descriptor (so will still have high similarity when excluded) whereas a salient comment will cause a larger shift (and so a lower similarity when excluded). We show results for adapting both the text branch (left) and visual branch (right), and observe that, as expected, uninformative comments such as “*That was great!*” and “*Possibly?!?!?! Lol*” cause little change to the descriptor, whereas comments related to objects in the video cause a larger shift. This demonstrates that the method is able to pick out and filter the relevant information.

Table 1: Video results - adapting the text branch We try adapting the text branch rather than the video branch for the video experiment. In this setting, the addition of comments seems to transfer less well to other datasets. Showing Recall@10

inference	#frames	VTC		KineticsComms		LiveBotEN	
		VTR	TVR	VTR	TVR	VTR	TVR
video	1	29.1	28.6	49.1	46.7	48.0	52.0
video+comments	1	33.2	33.5	47.8	45.6	49.0	52.0
video	8	28.6	27.8	57.5	55.3	68.0	71.0
video+comments	8	33.7	33.3	57.3	53.7	67.0	67.0

1.2 Additional Results

Similar to Tab.7 in the main paper, in Table 2 we evaluate zero-shot generalization of our video model on MSRVT and MSVD (w/o comments) and compare to CLIP which has been shown to generalize very well [4].

Table 2: Zero-Shot Generalization. Comparison of zero-shot generalization (without using comments). Results are TVR@10.

	MSRVT	MSVD
CLIP	60.7	65.27
Ours	63.8	76.93

Additionally, we perform a baseline experiment by removing all visual information for retrieval in Table 3. As expected, using the video with comments results in improved results over text-only retrieval.

Table 3: Text-only baseline. Comparing retrieval performance without any visual information.

	R@1	R@10
Title from Comments	20.3	41.3
Comments from Title	20.0	42.2
Title from Video	28.2	51.2
Video from Title	25.1	49.9

1.3 Dataset Statistics

In this section, we report some statistics of our dataset in order to give a sense of common topics and general distributions. We show word clouds of the most frequent words in the captions and comments in figures 4-5. In figure 6, we

plot a histogram of most common subreddits based on the number of videos, with "Minecraft" having the largest proportion, followed closely by "aww". The distribution of the number of comments per post can be seen in 7. In figures 8-9, we show the distribution of captions and comment lengths, measured in number of words.

1.4 Dataset Curation

We use the GPU implementation of the FAISS similarity search toolkit [3] to efficiently deduplicate the dataset by indexing the video thumbnail embeddings obtained from a ResNet18. These indices are then used to discard video entries with a high similarity to other posts.

1.5 Training Details

The majority of experiments were conducted on a rented 4xA100-40GB GPU server costing approximately 170USD per day, over the course of three months. Image models (using batch size 128) and video models without comments (using batch size 50) could train on a single 40GB GPU. For TimeSformer models the visual branch was processed on a separate GPU (when training with CAM and batch size 50) or pair of GPUs (for finetuning on video benchmarks with batch size 128). Pretraining the adapter on images takes approximately one hour per epoch. Training the full video model with CAM takes approximately 6 hours per epoch. For the video experiments, we first train the CAM for 5 epochs with the backbone frozen, and then train the rest of the network for one epoch, with the backbone modified to have temporal attention. We use the CAM with 5 comments, and adapt the visual branch of the model.

We use both photometric and temporal data augmentation. For photometric augmentation we employ random crops (0.5 – 1.0 scale), random horizontal flipping, and colour jitter (brightness, contrast, saturation, hue). For temporal augmentation, we first temporally subsample the input frames (which are often 30fps) according to a random stride selected uniformly from (4, 8, 16, 32) and then choose a random 8-frame segment uniformly. We normalise inputs using the same preprocessing as Clip (ImageNet mean and standard deviation, 224×224 input size).

At evaluation time we use a temporal stride of 16 and split the video into 8-frame chunks, taking the average of the descriptors of the chunks.

We randomly mask out comments with probability 0.5. We randomly skip adding the residual from the adapter with probability 0.5, which ensures that unadapted descriptors are also used in the loss and so the backbone network can still be used without the adapter.

All retrieval experiments are GPU accelerated using the FAISS⁴ library.

⁴ <https://github.com/facebookresearch/faiss>

1.6 Kinetics Comments

In this section we will describe the details for the additional comments we retrieve for the Kinetics-700 dataset. In Table 4 and Figure 10 we show the distribution of the number of comments in the dataset. We collect a maximum of 10 comments and exclude videos without comments, which leaves us with 111 920 videos of the originally 650 000 video clips. The majority of videos has one or two comments available.

Table 4: Comments per video statistics for the KineticsComments dataset.

#comments	1	2	3	4	5	6	7	8	9	10
#videos	50322	21847	11946	7960	5596	4311	3220	2671	2245	1852

1.7 Additional failure cases

In Figure 11 we show additional failure cases. We find that very vague comments “Why” or generic expressions “Ain’t his fault” can distract the model from the title. In the last example, the model does not capture the concept of a sad dog due to the mention of “happy” in the comments.

2 Model Diagram

Figure 12 shows a diagram of the model.

3 Datasheet for VTC dataset

In this section, we answer the questions proposed by Gebru et al. in [2], which were introduced as a way of documenting new datasets.

3.1 Motivation

For what purpose was the dataset created? The dataset was created strictly for research purposes. More specifically, this dataset addresses the research problem of using a weakly informative modality (user comments) in conjunction with other learning signals such as titles and videos for learning multi-modal representations.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset is created by VGG, a research group at the University of Oxford and Unitary AI, a company that’s developing AI to automate content moderation.

Who funded the creation of the dataset? The creation of dataset has not been funded directly. The individual researchers are funded by Amazon Machine Learning Awards (MLRA) and Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI).

3.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The dataset is comprised of links to videos, titles, and comments. Each video-title pair corresponds to a post on reddit.com. The dataset we share does not contain the data itself but hyperlinks to the data.

How many instances are there in total (of each type, if appropriate)? There are 339k video-title pairs with an average of 14 comments per video.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? This dataset is a sample of a larger, unfiltered version of the original dataset that we have collected. From the initial version, we handpicked a list of "safe" subreddits and removed posts if: 1) they had the "NSFW" or "over_18" tags; 2) the videos contained faces or the captions contained toxic or offensive text.

What data does each instance consist of? Each instance consists of: - "reddit_id" - "post_url" - "comment_ids" - "subreddit" - "video.length"

Is there a label or target associated with each instance? No, there are no labels provided.

Is any information missing from individual instances? If a user decides to remove a post, the link to the post will become invalid and thus not accessible anymore.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? Instances that have the same subreddit are likely to share semantic meaning.

Are there recommended data splits (e.g., training, development/validation, testing)? We will release the data splits we have used in our experiments with our code.

Are there any errors, sources of noise, or redundancies in the dataset? Although we have tried to remove most bot-generated text, it is likely that some noise will still exist due to the nature of this data. Similarly, a small proportion of posts might still contain identical or highly similar videos post-deduplication.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- In order to preserve user privacy, this dataset relies on links to reddit posts and comment ids. a) The links will no longer be valid if a user decides to delete their post. b) It would be possible to find the metadata of each post, as well as the link to the media file, on the Reddit archive. c) All links are accessible to everyone and are likely to remain so in the future.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? No, all data shared links to public posts.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? The dataset is still likely to contain a small proportion of offensive data. Due to the size of the dataset, we were not able to verify each video and each comment manually. However, we have tried to minimize the number of unsafe posts by: - manually filtering the subreddits included; - using Reddit metadata such as the "NSFW" and "over_18" tags to remove unsafe posts; - using automatic machine learning models to remove posts containing faces and toxic text.

Does the dataset relate to people? The dataset relates to people in the sense that each post is created by a person. In order to minimise the content related to people, we used a public face detector model to remove most instances of videos containing faces.

Does the dataset identify any subpopulations (e.g., by age, gender)? The dataset does not explicitly identify any subpopulations. However, some titles, user comments or image contents may identify individuals as part of a subpopulation.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Yes. Our dataset contains links to posts where the Reddit username will be visible and some of them might have identifying information contained in their profile such as personal images or information. This information is, however, already publicly available on Reddit.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? While we believe this is highly unlikely (as we only use already public posts and comments) – we cannot rule this out with absolute certainty. We will actively maintain this dataset after its release and ensure that if such information is included, that it is removed swiftly.

3.3 Collection process

How was the data associated with each instance acquired? The data was already available on Reddit.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? The dataset was collected via Reddit’s own API (<https://www.reddit.com/wiki/api>).

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? NA

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? NA

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? The dataset was collected between May 2020 and July 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section. The dataset related to people in so far that the dataset creators are individual users of reddit and posts can contain people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The dataset was collected via Reddit’s API. Thus, only public posts and data was downloaded.

Were the individuals in question notified about the data collection? NA.

Did the individuals in question consent to the collection and use of their data? NA.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? NA.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? NA.

3.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? The released dataset was preprocessed using an automated pipeline. This pipeline was taken from [1] and was used to removed videos that contain human faces using a publicly available face classifier.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes.

3.5 Uses

Has the dataset been used for any tasks already? This dataset has only been used for the experiments in this paper.

Is there a repository that links to any or all papers or systems that use the dataset? Google scholar will be able to track which papers have built upon this dataset/idea.

What (other) tasks could the dataset be used for? This dataset can be used for multi-modal representation learning or video-text retrieval.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Not that we are aware of.

Are there tasks for which the dataset should not be used? This dataset should not be used for tasks that might disclose the identity of the users or directly or indirectly harm them.

3.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will have a website and GitHub repository and be downloaded as a csv file containing links to the data points.

When will the dataset be distributed? The dataset will be published together with this paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset will be distributed under a research license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? NA.

3.7 Maintenance

Who is supporting/hosting/maintaining the dataset? The authors will maintain the dataset. In particular, Laura Hanu (

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The website of the dataset will contain all information to contact the authors and or maintainers of the dataset.

Is there an erratum? No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes, the website will contain a mechanism to version and update the dataset in case of errors.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

Will older versions of the dataset continue to be supported/hosted/maintained? Yes through versioning on GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, on the website.

Will these contributions be validated/verified? Yes, by the authors and maintainers of the dataset.

4 Dataset Examples

In figures Table 5, Table 6, and Table 7 we show random examples of the dataset with two comments (or less if a video only received one comment).








Video	Title	Comment	Comment
	Beerus and Whis are still a deadly combo! (Zenkai 3)	So do I, by far one of my favorite units. Thanks!	Really well played, love seeing Beerus in action!
	This little demon falls out of bed multiple times per night and just keeps snoozing.	It's pretty funny, I have dozens of these at this point ...!	
	Anyone knows what this is? I was playing on an private Nitrado Minecraft Server with my 2 Friends, we were playing on the earliest version. When we were building, we realized that on a random mountain this skull just appeared, we couldn't destroy it and it just spawned every kind of mob	idk man, some kind of glitch maybe?	we were on peaceful
	I know it's not that crazy but I'm still really proud:)	Pressure and I didn't want to fly all the way up again, I also was supposed	Why ender pearl?
	Gotta KNEAD the dough	That sound ahh so cute-lol	
	Anyone know why my iron farm won't work? They have beds	You built a java iron farm on bedrock lol	Do they have work stations?
	What am I looking at here? ',:/	That looks like some sort-of flightless fly. It kind of looks like a bat fl	

Table 5: A set of random samples from the dataset, showing title and up to two comments per video. (Included here since the guidelines only allow pdf/mp4 supplement)

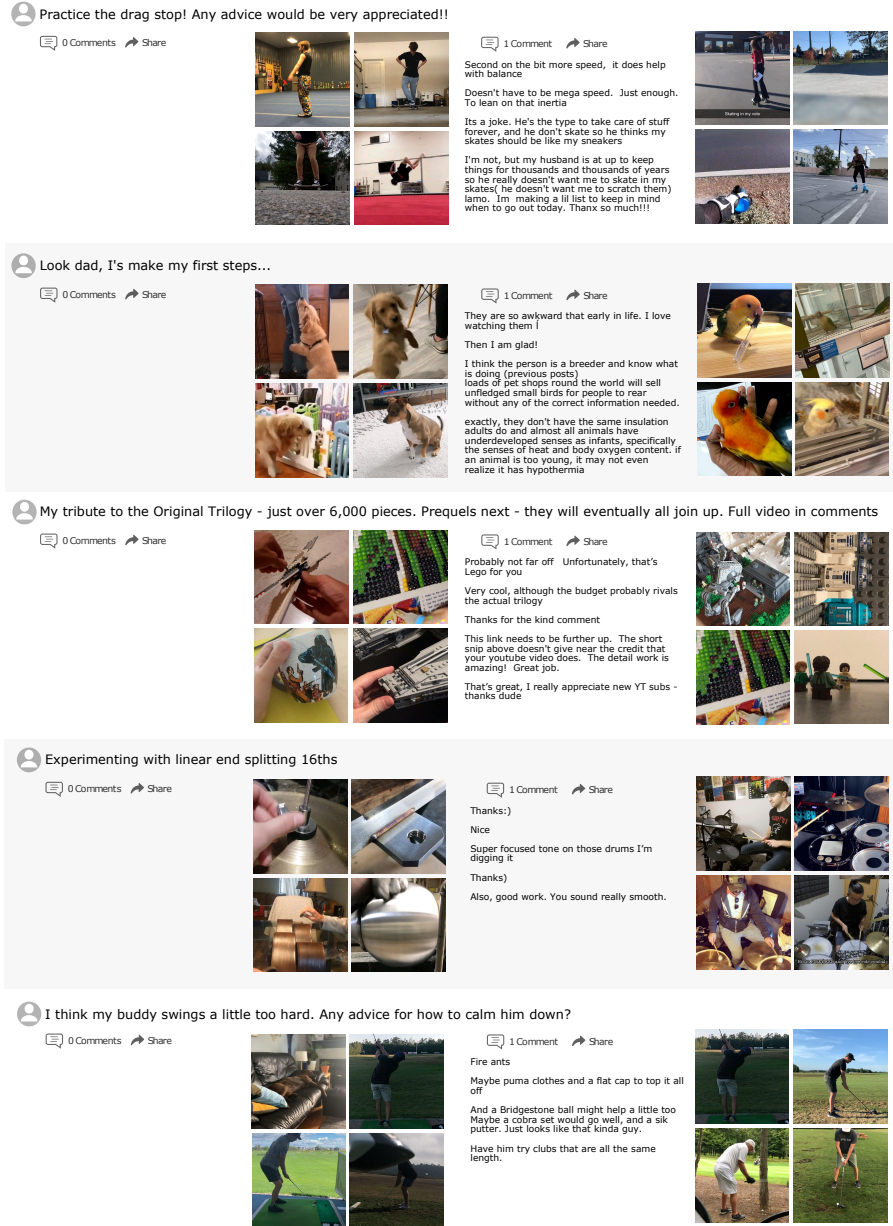


Fig. 1: Examples of retrieved video thumbnails when adapting the text branch.

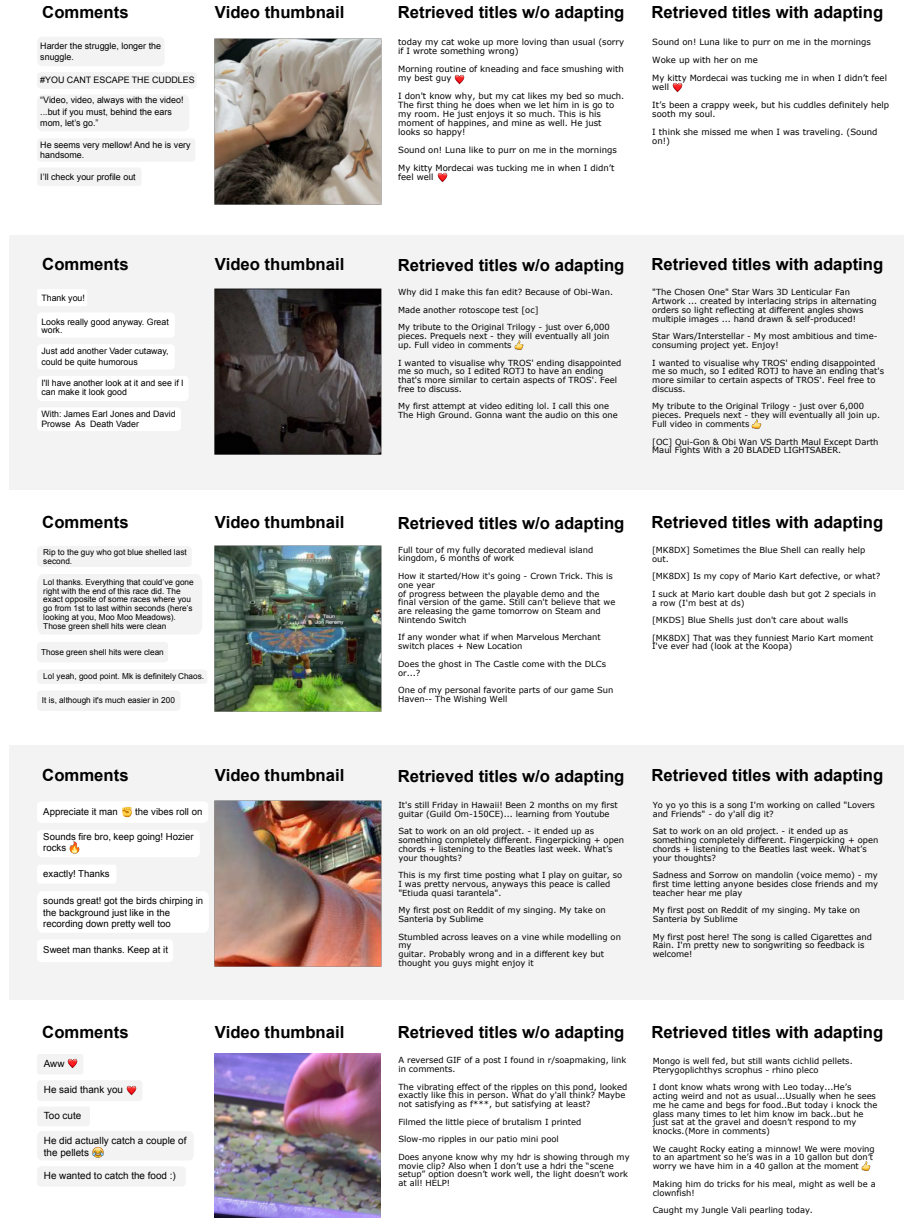
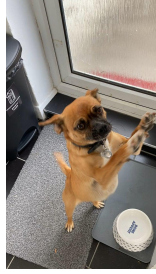
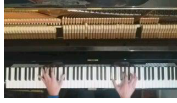


Fig. 2: Examples of retrieved titles when adapting the visual branch.

Title: She does this every time we feed her.

Lo! when I saw the bowl I was like of course
I've got the same breed with the same reaction
Which breed?
I'm so excited for her !!
I'm not sure she can eat it up there
what dog? all i see is a kangaroo
FOOD DANCE
That amazing!
Yup Pahaha
Lmao, She's DEFINITELY a dog where the "please eat slower, I
#give me the food master
Looks like it's bone appetit!
In my head I can hear the little Mario spring every time she
sproing
Totes going to edit that on, thanks for the inspo!

what dog? all i see is a kangaroo
Lo! when I saw the bowl I was like of course
#give me the food master
I'm not sure she can eat it up there
I've got the same breed with the same reaction
FOOD DANCE
Yup Pahaha
That amazing!
Looks like it's bone appetit!
I'm so excited for her !!
Which breed?
Lmao, She's DEFINITELY a dog where the "please eat slower, I
In my head I can hear the little Mario spring every time she
sproing
Totes going to edit that on, thanks for the inspo!

Title: Here's Giorno's theme from JJBA.

fire, piano has a great tone too
bro HOW BIG ARE UR HANDS? Those octaves seem like nothing to
Dude low key that'd actually be a sick stand
Beautifully played
damn I was trying to learn it but now I lost all motivation
Sheets?
SICK !!!
Would be very epic :))
This is the fonsi m arangement you can find the cover on You
Yes please ☺
Actually genial
That was great!
Same here, please
yes please!!
I love it! Expertly played. A lot of people will get the tem

fire, piano has a great tone too
bro HOW BIG ARE UR HANDS? Those octaves seem like nothing to
This is the fonsi m arangement you can find the cover on You
Dude low key that'd actually be a sick stand
Would be very epic :))
Sheets?
damn I was trying to learn it but now I lost all motivation
SICK !!!
Beautifully played
I love it! Expertly played. A lot of people will get the tem
yes please!!
Actually genial
Same here, please
That was great!
Yes please ☺

Title: Car battery problem? No Problem!

I am...electric man, Zap
Asgard's mechanic has really assimilated on Earth.
Car batteries are 12 volt direct current. Nothing happens as
I see youre a man of culture
So can someone tell me if this is healthy, safe, or painful
Electroboom already debunked this video a long ago
This made me laugh
Impossible and possibly deadly
Please enlighten us. What principle would allow this to work
There are absolutely 0 repercussions to trying.
You know volts arent the primary concerns right...
This makes my fillings hurt
If only this were possible.
Possibly?!?!?!?! Lol
Lol i understand the principle behind how this works... but

Car batteries are 12 volt direct current. Nothing happens as
I am...electric man, Zap
Electroboom already debunked this video a long ago
So can someone tell me if this is healthy, safe, or painful
Asgard's mechanic has really assimilated on Earth.
I see youre a man of culture
This made me laugh
Impossible and possibly deadly
You know volts arent the primary concerns right...
This makes my fillings hurt
Please enlighten us, What principle would allow this to work
There are absolutely 0 repercussions to trying.
Lol i understand the principle behind how this works... but
If only this were possible.
Possibly?!?!?!?! Lol

Fig. 3: Visualising comment saliency. We show the title and thumbnail for three videos, and show the ranked saliency of comments when adapting using the Text branch (left) and Image branch (right). Comments mentioning topics relevant to the title or image are ranked highly, while irrelevant comments are lower.

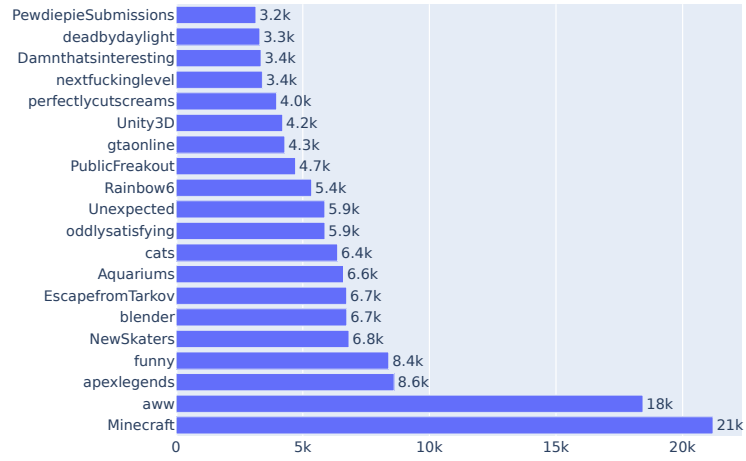


Fig. 6: Top 20 subreddits according to the number of videos.

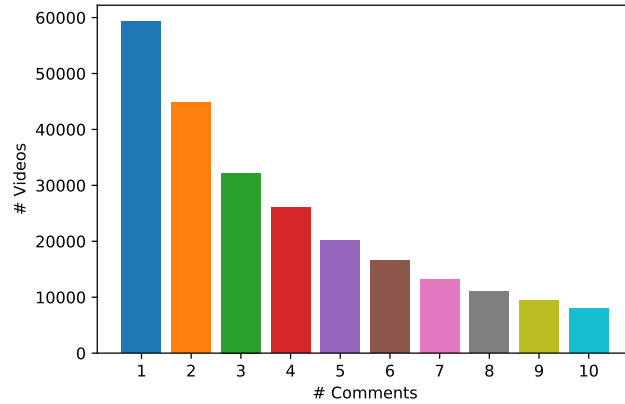


Fig. 7: We show a histogram of comment statistics on VTC.

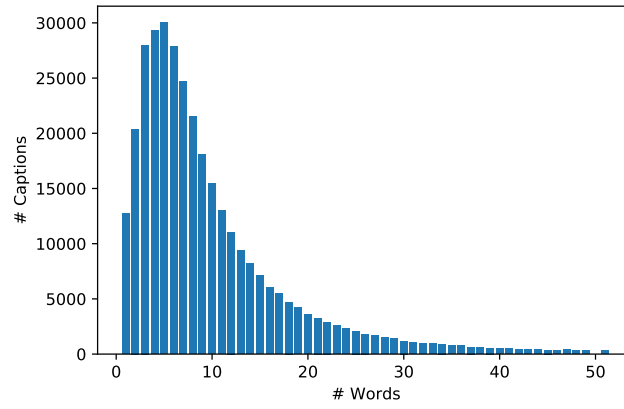


Fig. 8: Caption length distribution.

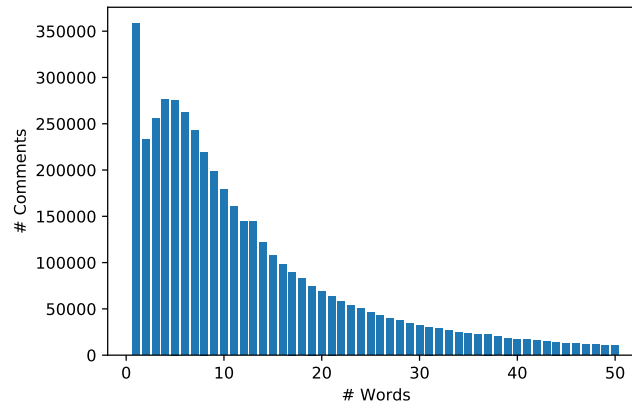


Fig. 9: Comment length distribution.

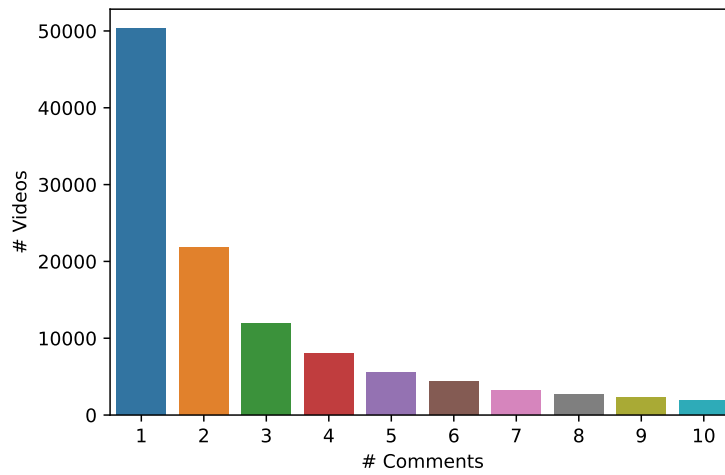


Fig. 10: We show a histogram of comment statistics on KineticsComments.

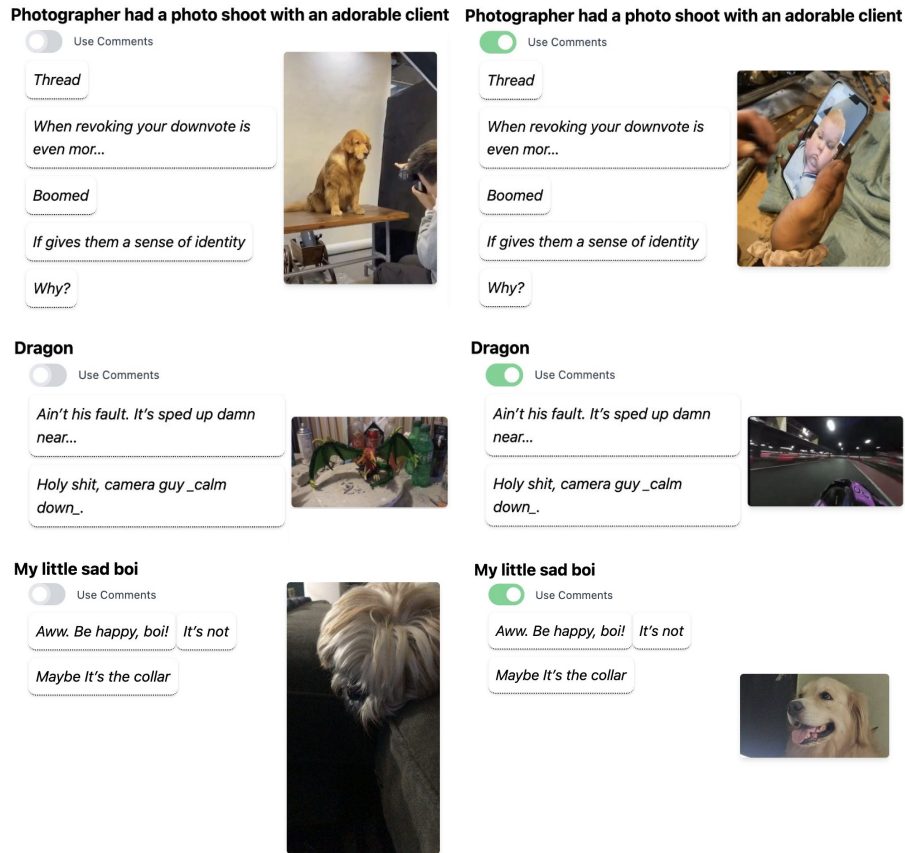


Fig. 11: Examples of failure cases where using comments confounds the model and leads to a more mismatched retrieved thumbnail.

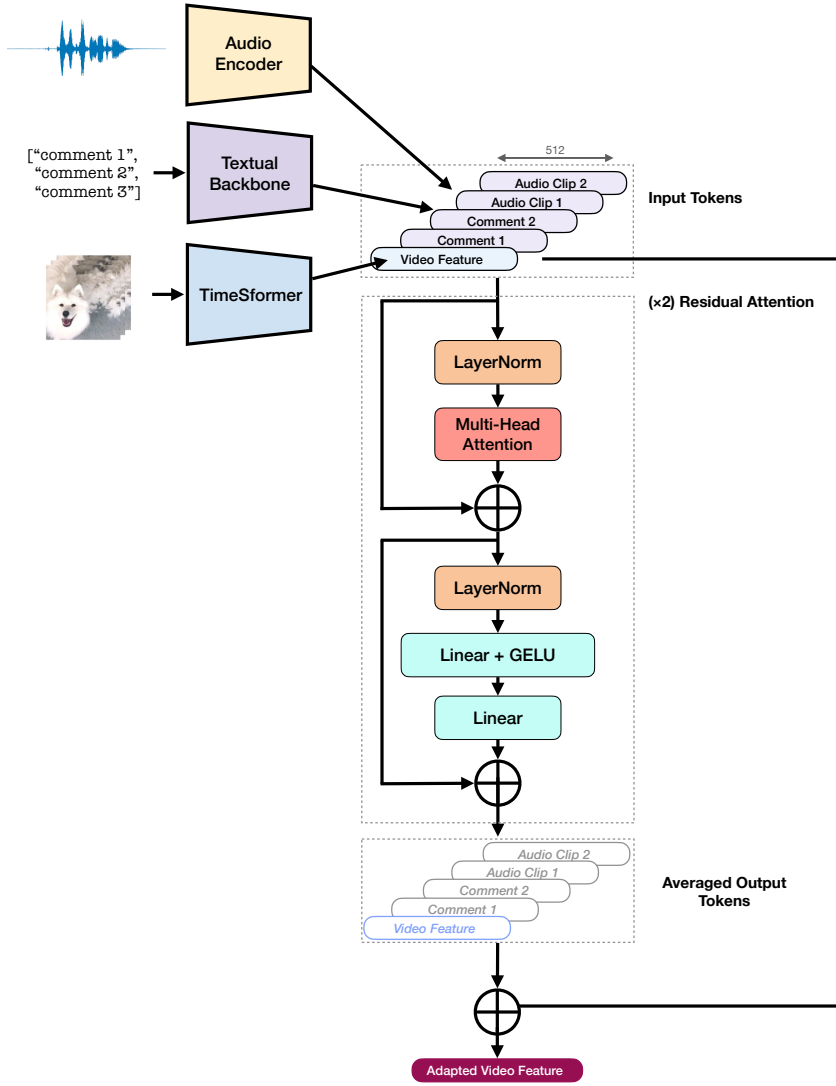


Fig. 12: We show a diagram of the feature extraction and Context Adapter Module for the case of adapting the Video Feature. Multi-Head Self-Attention is performed on the input tokens (which are themselves any combination of video, audio or textual features) as part of a transformer architecture consisting of two Residual Attention blocks. Finally the output token corresponding to the Video Feature is passed through a final linear layer and added to the original feature in a residual fashion.


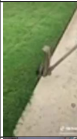

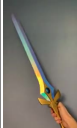






Video	Title	Comment	Comment
	This happened to me while running around as a dog, and I feel like the music is fitting	what is the music i like- that	
	A cutting edge Tik Tok!	How does one obtain this power?	That's really steady
	Didn't notice the change in physics until trying this gap again	Yeah, finding a lot of spots with strange physics, but it's mostly because	
	[self] I made my own She-Ra Sword using iridescent vinyl!	Its a PLA 3D print with aluminium core	What's underneath the vinyl?
	3 minutes of hill sprints in 17 seconds at Don Valley Park East this morning. Can you feel the burn???	Funny. I guess being corrected counts as "attitude" now.	Oh my...the contempt! Lmao....From your attitude I bet no one tells you any
	"Hidden Pools" 1	That tnak is beautiful- I love the plants and colours!	
	Bad egg from Walmart	Rough crowd.	Did you read his comment until the end?
	First trip around the track	Those were my guesses as it seems to be on the smaller side of the 4-6-0 cl	A Manor or a Hall?
	Ra Ra Rasputin Russia's greatest rage machine	Hello everyone! We have opened new	Irrelevant title but the meme is ok
	Anyone else hate having to ADS?	My friends and I love Marty Robbins XD thank you for this	"And the stranger's aim was deadly with the big iron on his hip."

Table 6: A set of random samples from the dataset, showing title and up to two comments per video. (Included here since the guidelines only allow pdf/mp4 supplement)







Video	Title	Comment	Comment
	Have had 6 cherry shrimps for 2 months, always can't find them and thought they are dead / been eaten. Surprised to see this baby shrimp today! Pencil for scale in video.	That's how they hide from me! Thanks for your advice.	They are very good at hiding. If there are holes or pits in your substrate
	The preferable option	Ugh.	Aaaieee!
	Is he strong? Listen, Bud! He's got radioactive blood. Can he swing from a thread? Take a look overhead.	Is that the house from Courage the Cowardly Dog	Oh, wheat!
	Pog gloryhole shot sorry I got excited I killed him and I finished my punisher pt3 in that raid as well	KomodoHype	-
	Why does this keep happening? (I am a noob)	On the rigidbody attached to your object select freeze rotation of the axis	When I was trying to fix it yesterday I found out that I didn't have a mesh
	I'm sure you've all seen this one, but just incase you haven't	A bug that flinches when hit by laser.	-

Table 7: A set of random samples from the dataset, showing title and up to two comments per video. (Included here since the guidelines only allow pdf/mp4 supplement)

References

1. Asano, Y.M., Rupprecht, C., Zisserman, A., Vedaldi, A.: Pass: An imagenet replacement for self-supervised pretraining without humans. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
2. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. *Communications of the ACM* **64**(12), 86–92 (2021)
3. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734* (2017)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021)