International Journal of Computer Vision (2026) 134:60
https://doi.org/10.1007/s11263-025-02664-4

It's Just Another Day: Unique Video Captioning by Discriminitive Prompting

Toby Perrett[1] · **Tengda Han**[2] · **Dima Damen**[1] · **Andrew Zisserman**[2]

Received: 15 May 2025 / Accepted: 6 October 2025

Abstract

Long videos contain many repeating actions, events and shots. These repetitions are frequently given identical captions, which makes it difficult to retrieve the exact desired clip using a text search. In this paper, we formulate the problem of unique captioning: Given multiple clips with the same caption, we generate a new caption for each clip that uniquely identifies it. We propose Captioning by Discriminative Prompting (CDP), which predicts a property that can separate identically captioned clips, and use it to generate unique captions. We introduce two benchmarks for unique captioning, based on egocentric footage and timeloop movies – where repeating actions are common. We demonstrate that captions generated by CDP improve text-to-video R@1 by 15% for egocentric videos and 10% in timeloop movies. https://tobyperrett.github.io/its-just-another-day

Keywords Uniqueness · Video Captioning · Egocentric · Movies

1 Introduction

Life is repetitive. So videos of daily life will inevitably contain visually similar events, places, people and activities. As a consequence, captioning video clips from similar activities will often result in identical sentences. For example, in Ego4D (Grauman et al., 2022) when using an off-the-shelf captioner (Zhao et al., 2023), 66% of clips in each video share their caption with at least one other clip, and thus do not have a unique caption. This lack of caption uniqueness impacts text based search – a user has to linearly scan all similar clips to find the desired clip. Can we do better?

The root of the problem is that currently clips are captioned *independently* (Luo et al., 2020; Seo et al., 2022; Lin et al., 2022)59,(Zhao et al., 2023; Zhang et al., 2023; Liu, 2022; Lu, 2023). Instead, if the captioner is aware of visually similar clips, then potentially it can discriminate one from the others in its description.

That is the objective of this paper: to generate concise captions which discriminate between visually similar clips. We achieve this in two ways: first, we develop a model that observes all visually similar clips, and predicts prompts that will trigger the captioner to generate a unique description for each. Second, if it is not possible to uniquely caption a clip, we increase its temporal extent until a unique caption can be found.

In particular, we show that by taking an approach similar to the 'twenty questions' game (Twenty, 2024), a lightweight model can be learnt for an already trained captioner. This allows for the direct prediction of discriminative prompts at inference, eliminating the need to explicitly test all possible prompts. Our framework is agnostic to the captioner (and visual-text embedding) used.

We focus on two sources of videos with known repetitions. One is *egocentric footage* of daily life, where similar clips occur naturally due to actions and routines occurring in familiar environments. The other is *timeloop movies*, where repetition is specifically written into the plot, with the added challenge of identical or near duplicate clips.

The ability to uniquely caption video clips will enable a number of downstream tasks: (i) When querying for a caption in a retrieval system, e.g. "Opens the fridge" in Fig. 1, the unique captions will enable an 'auto-completion', append-

✉ Tengda Han
htd@robots.ox.ac.uk

Toby Perrett
toby.perrett@bristol.ac.uk

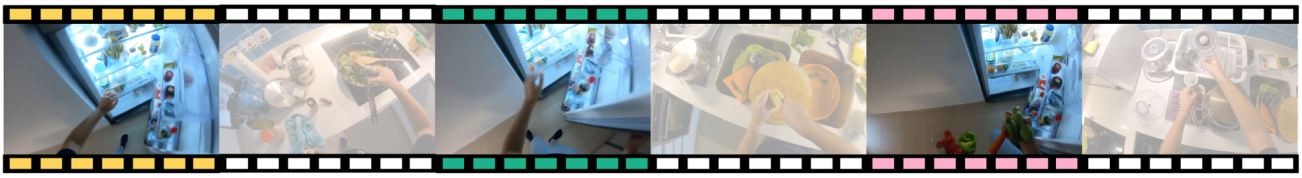
Dima Damen
dima.damen@bristol.ac.uk

Andrew Zisserman
az@robots.ox.ac.uk

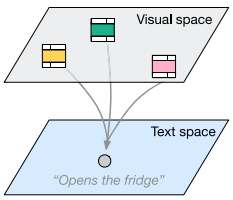
[1] University of Bristol, Bristol, UK

[2] University of Oxford, Oxford, UK

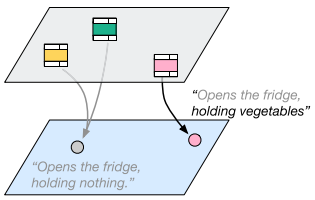
Published online: 13 January 2026



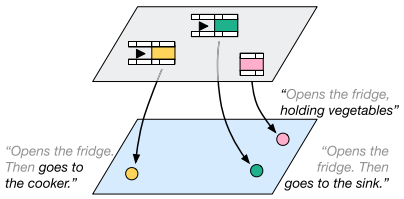
(a) A long egocentric video. The yellow/green/pink clips are captioned as "Opens the fridge".



(b) Standard captioning can generate the same caption for multiple clips.



(c) We consider clips with the same caption, to find a property that captions them uniquely.



(d) If we cannot find a unique property, we explore the following clips for an extended unique caption.

Fig. 1 Standard video captioning breaks the video into smaller clips and considers each clip independently. As a result, it is likely multiple clips from one video will have the same exact caption (a). We introduce Captioning by Discriminative Prompting (CDP), an approach for generating unique captions. considers the set of clips with the same caption (b), and predicts a discriminative prompt (*e.g.* "holding") that allows the clip to be captioned uniquely (c) When a unique caption cannot be found, we advance to the next clip ▶ to allow unique captioning based on following actions (d)

ing the distinctive aspects of clips to the text query. This way the user can directly select the clip they are interested in (e.g. "Opens the fridge, while holding vegetables in hand") without having to study all videos of opening the fridge. (ii) Further refinement of captions on clips previously captioned identically, without having to change the captioner. In summary, our contributions are:

- We introduce a framework for unique captioning, Captioning by Discriminative Prompting (CDP), based on predicting the dimension along which multiple clips differ. This dimension is represented by a discriminative prompt, which conditions the captioner to generate a unique caption for each clip.
- We introduce two benchmarks for unique captioning, from egocentric videos with identical narrations and from the repeating segments of timeloop movies.
- We show CDP improves text-to-video and video-to-text retrieval on both benchmarks.

2 Related work

Captioning. The standard practice for video captioning follows that for images (Mokady et al., 2021; Li et al., 2022, 2023), which is to take an auto-regressive language model, and condition it on the video (Luo et al., 2020; Seo et al., 2022; Lin et al., 2022)59(Zhao et al., 2023; Yang et al., 2023; Zhang et al., 2023). Captioning models have improved with more data (Sharma et al., 2018; Yang et al., 2023) and better pre-trained image (Li et al., 2023) and language models (Zhang et al., 2023). Improvements have also been found by generating longer captions with more detail (Ding, 2023), maximising mutual information (Wang et al., 2020), and incorporating synthetic captions (Betker et al., 2023). Other approaches include distilling knowledge from foundation models (Yang et al., 2020; Long et al., 2023; Wu et al., 2023), alignment with human labelling (Sun, 2023), and controlling the level of detail given by captioners (Dwibedi et al., 2024). Dense (Yang et al., 2023), hierarchical (Islam et al., 2024) and progress-aware (Xue et al., 2025) captioning models assign multiple sequential captions to a longer clip, but do not aim for unique captions. To introduce diversity into generated captions, works have sought to produce multiple captions per clip, ensuring that the concepts contained in one caption are distinct from other captions *for the same video clip* (Liu, 2022; Lu, 2023).

The problem we tackle in this paper is orthogonal to the above works. Given an already trained captioner, how can we find the differences between *multiple* video clips given a captioner's existing capability? We aim to generate concise, unique captions for all clips in a video, or a gallery of videos.

Model "blind spots". A related line of work is determining aspects of an image or video that models are blind to, and thus unable to tell apart, typically evaluated on classification and VQA tasks. Examples include compositionality (Thrush et al., 2022) in images, temporal ordering in video (Hendricks et al., 2018), and the limitations of image/language pre-trained models compared to self-supervised image representations (Tong et al., 2024). These works have aimed at evaluating

and providing benchmarks, but some also attempt to fix these shortcomings, for example by collecting data focusing on these blind spots (Bagad et al., 2023; Ventura et al., 2024), or training captioners on LLM processed auto-labels to find differences between pairs of samples (Nagarajan & Torresani, 2024; Lin, 2024). In this work, instead of trying to improve the base ability of a captioning model, we acknowledge that limitations will likely always exist. We find differences which are discernible by the chosen existing captioning model. This approach will continue to be relevant and useful as models' capabilities continue to improve.

Long video. Early computer vision studies of long video focused on footage from surveillance (Hampapur, 2005), TV (Duan et al., 2006; Xu & Li, 2003) and movies (Sivic & Zisserman, 2003). Due to computational budgets, the community shifted to understanding short clips (Carreira & Zisserman, 2017; Goyal et al., 2017). Long-video is again being studied, due to the rise of large-scale instructional (Zhou et al., 2018; Miech et al., 2019) and egocentric datasets (Damen et al., 2022; Grauman et al., 2022), and the ability of models to operate on larger temporal windows (Wu et al., 2022; Wang, 2022; Liu et al., 2023). Several studied tasks in long video would benefit from the ability to uniquely caption clips, such as summarisation (Lee et al., 2012), audio description (Han et al., 2023), VQA (Tapaswi et al., 2016) and retrieval (Ramakrishnan et al., 2023). While visual-language retrieval benchmarks (Rohrbach et al., 2015; Xu et al., 2016) report Recall@K accuracy as a common practice to allow retrieving similar instances in the corpus, they do not explicitly enforce uniqueness. In this paper, we build our egocentric unique captioning benchmark using footage from the massive-scale Ego4D dataset (Grauman et al., 2022).

Timeloop movies. Despite the challenges they pose to plot understanding, timeloop movies rarely feature in the computer vision literature, where the emphasis for movies/TV is on scale (Lei et al., 2018; Bain et al., 2020; Huang et al., 2020; Yue et al., 2023). Whilst they are scarce compared to standard movies (with only 71 dating back to 1947 listed on Wikipedia (List, 2024)) and thus not suitable for large-scale training, timeloop movies can provide insightful diagnostic and qualitative results. For example, the movies *Groundhog Day* and *Run Lola Run* were used for location retrieval assessment in Sivic and Zisserman (2003), as tests of identical and near-duplicate shot retrieval in Chum et al. (2007), and the movie's repeating structure was determined in Schaffalitzky and Zisserman (2003) by matching shots of the same location. In this work, we revisit timeloop movies as an evaluation-only benchmark for unique video captioning, where understanding the temporal context of repetitive clips is essential.

3 Method: Captioning by Discriminative Prompting

Captioning by Discriminative Prompting (CDP), generates *unique captions* for a set of visually similar clips, so they can be discriminated and teased apart in the visual-text embedding space. CDP is built around three key ideas:

1. A set of *discriminative prompts*, in order to direct a captioner to focus on properties of one clip which distinguish it from others. These properties are chosen by contrasting all similar clips, and thus provide our mechanism for conditioning a single-clip captioner on multiple clips.
2. A *combinatorial search* over all prompts and clips, to find the exact combination of prompts that will generate the most unique set of captions.
3. A network, *CDPNet*, which approximates the most computationally expensive part of the search - autoregressively captioning each clip with all prompts and then computing video/text embedding similarities.

Note that we can, with significant computational cost, generate unique captions by performing the combinatorial search for discriminative prompts without training an additional network. This would be possible but inefficient. We introduce this process first in Section 3.2 as our method builds on the search, and it provides a good insight into the constraints and comparisons necessary for caption uniqueness. CDPNet is a required approximation to make our proposed approach feasible for inference, described in Section 3.3.

3.1 Problem statement and uniqueness definition

Given a set of N video clips $\mathcal{V} = \{v_1, v_2, ..., v_N\}$, we aim to output a corresponding set of unique captions $\mathcal{C} = \{c_1, c_2, ..., c_N\}$. We assume access to two trained foundation models:

- A video captioner $\Theta(v, p)$, which takes in a video clip v and optional prompt p, and produces the caption text c.
- A dual-encoder video/text model, with encoders $f(v)$ and $g(c)$ for projecting video clips and caption text into a joint embedding space. The video-text similarity is measured by cosine similarity between their embeddings.

These trained models are general and do not need to have been trained jointly. They remain frozen throughout.

A correct output of our method is one where captions are distinct, i.e. $c_i \neq c_j; \forall i \neq j$, but also can be correctly matched to the clip it was generated from. Formally, v_i is captioned uniquely by c_i with respect to all other $v_j \in \mathcal{V}$ and $c_j \in \mathcal{C}$ if the following condition is satisfied, where $\langle \cdot, \cdot \rangle$

denotes cosine similarity:

$$\langle f(v_i), g(c_i)\rangle > \max\Big(\max_{j\neq i}\langle f(v_i), g(c_j)\rangle, \max_{j\neq i}\langle f(v_j), g(c_i)\rangle\Big). \tag{1}$$

3.2 Combinatorial search for unique captions

Discriminative prompts. For each clip v_i, our goal is to select one or more discriminative prompt(s) that will induce a unique caption with respect to all other clips in $\mathcal{V}$. While there are many ways in which clips can differ, we propose to use a given set of general prompts. We define a bank of P prompts $\mathcal{B} = \{p_1, ..., p_P\}$. Fixed prompts are more suited to this task than learned, as they are interpretable, and can be designed to increase diversity and reflect known model capabilities. We typically select the most frequent N-grams from the training set.

Selecting a single discriminative prompt. We define the similarity function, s, between a video v_i and a caption generated from another video v_j using the prompt $p_k \in \mathcal{B}$ as

$$s(v_i, v_j, p_k) = \langle f(v_i), g(\Theta(v_j, p_k))\rangle, \tag{2}$$

which measures clip/caption similarity in the shared embedding space. We next define the uniqueness margin, $\mathcal{M}$, for clip v_i, with respect to the other clips in $\mathcal{V}$, using prompt p_k, as

$$\mathcal{M}(v_i, p_k) = s(v_i, v_i, p_k) - \max\Big(\max_{j\neq i}\big(s(v_j, v_i, p_k)\big), \max_{j\neq i}\big(s(v_i, v_j, p_k)\big)\Big). \tag{3}$$

For the clip v_i, we denote the chosen discriminative prompt as $p_{\hat{k}}$, which is the prompt with index $\hat{k}$ that maximises $\mathcal{M}$, *i.e.* $\hat{k} = \arg\max_k \mathcal{M}(v_i, p_k)$. If $\mathcal{M}(v_i, p_{\hat{k}}) > \lambda$, where λ is the margin of confidence, then the caption $c_i = \Theta(v_i, p_{\hat{k}})$ uniquely describes v_i, as defined by Eq. 1. That is, v_i is closer to its caption c_i than to any other caption, and c_i is closer to v_i than to any other video. If $\mathcal{M}(v_i, p_{\hat{k}}) \leq \lambda$, then we have determined that it is not possible to identify v_i uniquely, given the other video clips in the set and the capabilities of the captioner and embedding space. In such cases, multiple discriminative prompts are required to find unique captions, which we describe next.

Selecting multiple discriminative prompts. Given P prompts in our bank and N clips in $\mathcal{V}$, the exact search for the best combination of *multiple* prompts would have complexity $\mathcal{O}(NP^P)$ if every prompt could be chosen for every clip. Instead, we constrain the maximum number of prompts to be selected for each clip as $\alpha << P$, which reduces the complexity to $\mathcal{O}(NP^\alpha)$. This not only constrains the search space, but also keeps our captions concise and focused on the most discriminative aspects of each clip.

Our problem is now to define the margin for combinations of prompts, and to find the combination which obtains the maximum margin. The margin for a single prompt reflects that a clip is distinct from all others with respect to that single prompt. We extend to multiple prompts by adjusting the similarity measure (Eq. 2) used to compute the margin. Instead, we take the mean of the similarities of each individual prompt in the combination.

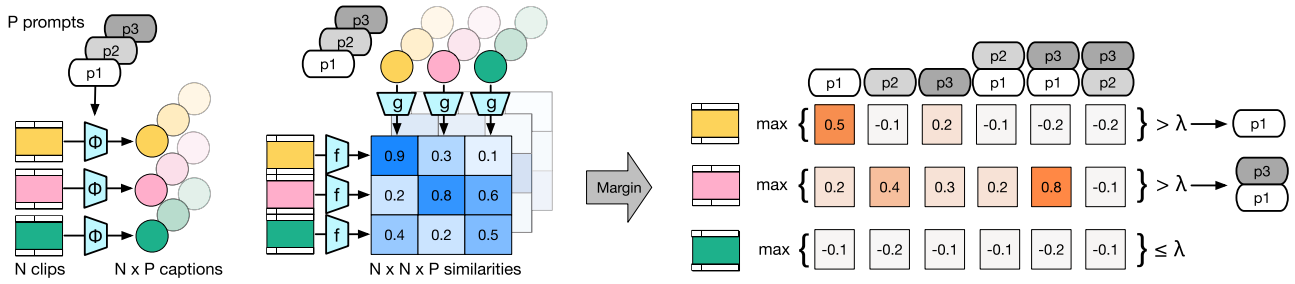
We first define all combinations of prompts in $\mathcal{B}$ up to order α as $\mathcal{B}^\alpha$. For example, $\mathcal{B}^2$ contains all the combinations of $\{p_i\}$ and $\{p_i, p_j\}$. We denote the k-th combination in $\mathcal{B}^\alpha$ as B_k^α. We then adjust the similarity to:

$$s^\alpha(v_i, v_j, B_k^\alpha) = \frac{1}{|B_k^\alpha|}\sum_{p\in B_k^\alpha}\langle f(v_i), g(\Theta(v_j, p))\rangle \tag{4}$$

We use s^α instead of s in calculating the uniqueness margin (in Eq. 3) to choose the best prompt combination $B_{\hat{k}}^\alpha$.

We now have a complete pipeline for computing margins, and thus creating unique captions, for a set of clips with no learning necessary, which is illustrated in Fig. 2. In 2a, captions are extracted with every prompt for every clip using the frozen single-clip captioner. In 2b, clips and captions are projected into the embedding space by f and g, where their similarities are computed. In 2c, for each video, margins are computed from the similarities for every prompt combination (here, $\alpha = 2$, so all individual and pairs of prompts are tried). If the maximum margin for a clip is $> \lambda$, *i.e.* $\mathcal{M}(v_i, B_{\hat{k}}^\alpha) > \lambda$, then the prompt combination with the maximum margin generates a unique caption for that clip. However, it is possible that no combination of prompts is able to uniquely identify a clip, i.e. $\mathcal{M}(v_i, B_{\hat{k}}^\alpha) \leq \lambda$. In such cases, we allow the temporal duration of video clips to be extended in time, which we describe next.

Temporal extension As clips are taken from a longer video, we explore advancing to a subsequent clip so as to distinguish identical clips. As a result of the expansion, we can caption the two clips into "X then Y" vs "X then Z", where Y and Z are distinct follow-up events in the longer video. We denote a video v_i advanced by time t as $v_i(\blacktriangleright t)$. We define the set of all (prompt, time) combinations up to time τ as $\mathcal{B}^\tau$ and define the k-th element as B_k^τ. Our problem is now to adjust the similarity to account for prompt/time combinations. We can again achieve this by adjusting the similarity for clips up to time τ as follows:



(a) Each clip is captioned using every prompt.

(b) Clips and captions are projected by f and g into a shared space, where similarities (Eq. 4) are computed.

(c) The margins (Eq. 3) are computed using the similarities for every clip/prompt combination. We search for the max margin per clip, which generates a unique caption if $> \lambda$.

Fig. 2 Pipeline for computing the margin at a single timestep for three clips. This example uses $\alpha = 2$, so margins are computed for all single prompts and pairs of prompts. (a) and (b) are replaced by a learned network in Sec. 3.3

$$s^{\tau}(v_i, v_j, B_k^{\tau}) = \frac{1}{|B_k^{\tau}|} \sum_{p,t \in B_k^{\tau}} \Big\langle f\big(v_i(\blacktriangleright\, t)\big), g\Big(\Theta\big(v_j(\blacktriangleright\, t), p\big)\Big)\Big\rangle \tag{5}$$

Similar to the case of multiple prompts at a single timestep above, we replace s in Eq. 3 with s^{τ}. The process for generating unique captions is then the same as Fig. 2. The only differences would be the tensor in 2b having an additional dimension for the number of timestep advances (*i.e.* $N \times N \times P \times \tau$), and additional margin computations for prompt combinations at different timesteps.

3.3 Predicting Discriminative Prompts with CDPNet

In Sec. 3.1 and 3.2, we predict the best prompts exhaustively and exactly. That is, we have an exact process for generating unique captions in a given video/text space, with no training or fine-tuning. It uses exact embeddings of captioner outputs, computes margins, and searches over the set of prompt combinations to find the optimal combination. However, this process is not scalable. We recognise that the main bottleneck is attempting to generate a caption for every one of the $N \times P \times \tau$ video/prompt/time combinations.

We approximate this by training a Captioning by Discriminative Prompting Network, CDPNet, denoted Ψ. It takes in two clips v_i and v_j and a prompt p_k. It predicts the video-text similarity between the visual embedding of v_i, and the text embedding of v_j when captioned using prompt p_k. We denote this predicted similarity as:

$$\hat{s}_{ijk} = \Psi(v_i, v_j, p_k). \tag{6}$$

Importantly, this allows us to replace Fig. 2a and 2b with the direct prediction of similarities from video clips only, *without having to calculate a forward pass through the captioner or the embedding networks*.

We train Ψ by minimising the MSE between its output $\hat{s}_{ijk}$, and the computed cosine similarity of the embeddings $s(v_i, v_j, b_k)$ from Eq. 2. To allow scalability, CDPNet Ψ only operates on one prompt, and the search for combinations of prompts is handled by the averaging of similarities in the combinatorial search (Eq. 5).

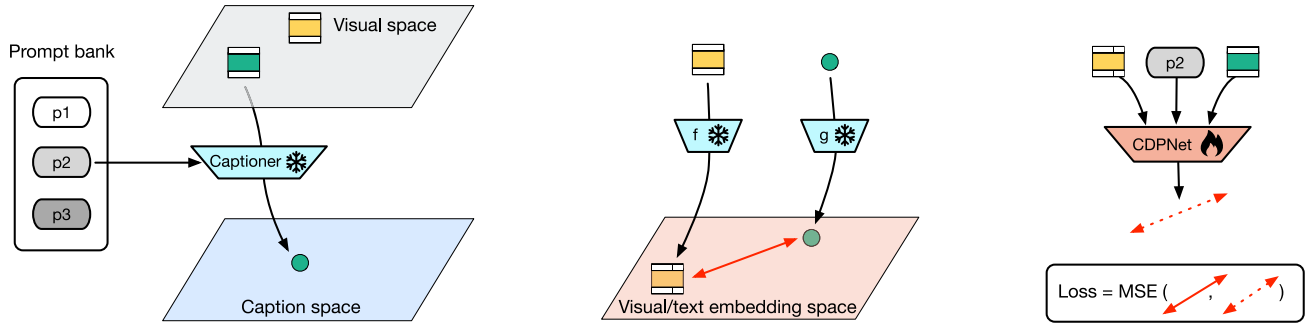
Fig. 3 shows the training process for CDPNet: 3a shows a green clip, and its generated caption (green circle), when captioned using a prompt from the bank; 3b shows the yellow clip and green caption being embedded by f and g in the shared visual/text space, where their similarity is computed; and 3c shows how this similarity is used as the training signal for Ψ, which takes just the clips and prompt (*i.e.* no caption).

We implement Ψ as a transformer encoder, and its prompt bank as a collection of learnable tokens, one per prompt. Ψ takes as input representations for the two clips as well as the prompt token selected from the bank. We use learned positional encodings to indicate the different inputs. We apply a linear layer to the output of the prompt token to regress the similarity $\hat{s}$.

In summary, CDPNet operates as follows. It is initialised with a bank of prompts. For a given set of clips, CDPNet directly predicts the visual-text similarity between each clip and every other clip with respect to every prompt in the bank, using visual input only (*i.e.* no captioning). The combinatorial search is run over CDPNet similarity predictions to find the combination of (max α) prompts for each clip which produces the maximum margin. If this margin is $> \lambda$, the clip and prompt combination are passed to the captioner to expecting unique captions. If the margin is $\leq \lambda$, the clip is advanced ($\blacktriangleright$) and the process is repeated until a unique prompt combination is found.

4 Unique Captioning Benchmarks

The ability to uniquely caption video clips can be assessed by checking if there is a one-to-one correspondence between



(a) A frozen captioner is run on a video input with a discriminative prompt from the bank.

(b) f and g project clips and captions, where their similarities are computed.

(c) CDPNet estimates embedding similarities.

Fig. 3 Training CDPNet, which aims to predict the similarity between a clip (yellow) and the caption from another clip (green), when conditioning the captioner with a prompt. (a) and (b) show how to compute the similarity between the clip and caption in a shared embedding space, which is used as the training signal. (c) shows CDPNet predicting the similarity only using the video clips and prompt in one forward pass

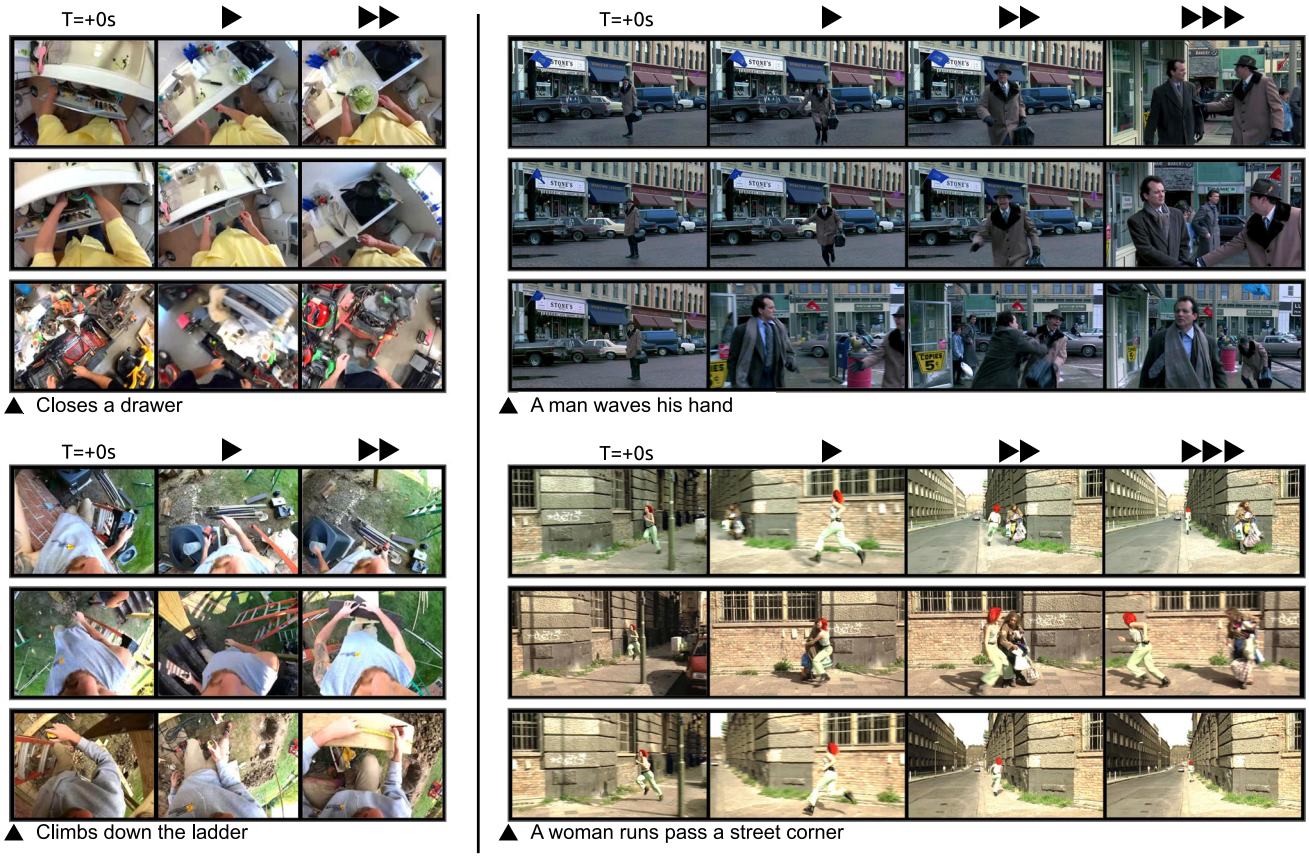
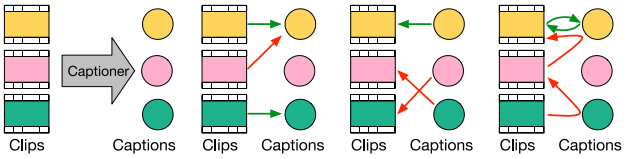


Fig. 4 Examples of the Unique Captioning Benchmarks, from Egocentric videos (left) and timeloop movies (right). We show 3 sequences from each set of clips – i.e. video clips with the same caption at T=+0. Subsequent clips are indicated by ▶. We note the common caption in each case



(a) First, clips are captioned. (b) Video→text R@1. (c) Text→video R@1 (d) Cycle@1.

Fig. 5 Unique captioning evaluation. First, in (a), clips are captioned. Video→text R@1 is 66% in (b), as 2/3 clips retrieve their captions. Text→video R@1 is 33% in (c), as only 1/3 captions retrieves the clip it was generated from. Average Recall@1 is 50% – the average of text→video and video→text R@1. (d) shows the stricter Cycle@1, which is 33%, as only the yellow clip has Video→text→video cycle consistency

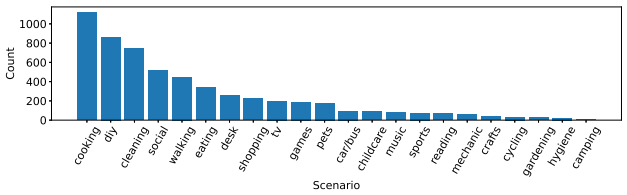


Fig. 6 Scenarios in the egocentric benchmark

Fig. 7 Wordle of narrations used to create the egocentric benchmark

the set of video clips and the predicted set of captions. We frame this as a retrieval problem and use standard retrieval metrics accordingly.

e4.pdf. Given a set of video clips $\mathcal{V}$ and generated captions $\mathcal{C}$, we measure how well each caption c_i can retrieve the video v_i and vice versa. We use both text-to-video and video-to-text recall @1, @2 and @3. Text-to-video R@K indicates that v_i is one of the K-closest videos to c_i in the shared embedding space, and vice versa. We present Avg R@1 as a standard combined metric. However, Avg R@1 does not check for one-to-one assignment. Accordingly, we introduce the more stringent Cycle@1 combined metric. This checks that video clip v_i retrieves the corresponding c_i *and* c_i then retrieves v_i. Formally, this means v_i and c_i must satisfy the uniqueness condition in Eq. 1. Fig. 5 illustrates these unique captioning metrics.

Egocentric benchmark. Videos of daily living are an ideal test of unique captioning, as we perform the same activities and routines over both short and long timescales. The egocentric viewpoint naturally captures these repetitions as we move around, so we use videos and narrations from Ego4D to construct this benchmark. For training, We sample 30K ground truth narrations from the Natural Language Query (NLQ) training split that are repeated 10 or more times, i.e. 10 or more clips share the same narrations. For each, we randomly sample 10 clips from the same or a different video, producing a benchmark of 300K clips. Our balanced sampling, per narration, helps alleviate the long-tail property where some narrations such as "look around" are common.

For evaluation, we select 300 sets of 10 clips from the NLQ validation set, following the same process. Sets remain fixed for all evaluations. To select prompts for the bank $\mathcal{B}$, we take N-grams from narrations on the training set. We sort by N-gram frequency, selecting the 10 largest, whilst manually removing semantically similar prompts to ensure diversity. Prompts are ablated in Sec. 5.6.

Fig. 6 explores the distribution of scenarios which appear in the egocentric benchmark. These roughly match the scenarios present in Ego4D. We also show a wordle of the narrations used to generate the benchmark in Fig. 7. Interestingly, movement features a lot (*e.g.* "walks", "around"), and it is often difficult for models to distinguish between different parts of an environment.

Timeloop movies benchmark. A timeloop movie is one where the plot centres around the characters reliving the same sequence of events over and over. For example, in the 1998 movie 'Run Lola Run', the character repeatedly starts running from the same premise in her repeated attempts to save her boyfriend. At each repetition, a difference occurs, which causes the story to diverge. This makes these timeloop movies an ideal test for unique captioning.

For the test set, we start with the Wikipedia timeloop list (List, 2024), and manually annotate the timestamps of these repeating moments. We only select sets when there are at least 3 repeating moments per movie, and where these moments are near-identical, making this a very challenging evaluation. For example, in the movie *Groundhog Day*, the main character is clearly filmed waking up 9 times, from which the story diverges. These "waking-up" clips form a set of 9 clips. This process gives a total of 63 clips across 10 timeloop movies, ranging from 1993 to 2021. Sets range in size from 3 to 10.

Due to the scarcity of timeloop movies, they can only be used for a test set. For training , we use non-timeloop data as a proxy for timeloop clips. We curate our training set from the Condensed Movie Dataset (Bain et al., 2020), which consists of sequences of 2-minute movie truncations, from which we *exclude timeloop movies*. We construct sets of 10 video clips with CLIP visual similarity > 0.92 (chosen to match the egocentric training size), giving 30K sets of 10 clips. Examples from both benchmarks are shown in Fig. 4.

Our benchmark and training clips are available from the project's website.

Table 1 Egocentric benchmark using the LaViLa VCLM as the base captioner, and combined withCaptioning by Discriminative Prompting (CDP). At every T, improves by a significant margin on every metric. Improvements are shown in ***green*** for the combined metrics Avg R@1 and Cycle@1

T	#Clips	Method *Chance*	Text→Video R@1 *10*	 R@2 *20*	 R@3 *30*	Video→Text R@1 *10*	 R@2 *20*	 R@3 *30*	Avg R@1 *10*	Cycle@1 *1.0*
+0s	1	LaViLa VCLM	40	58	70	33	49	61	37	22.0
		LaViLa VCLM + CDP	**55**	**71**	**80**	**34**	**50**	**61**	**45** ***+8***	**26.0** ***+4.0***
+5s	2	LaViLa VCLM	42	61	72	34	51	62	38	23.0
		LaViLa VCLM + CDP	**69**	**81**	**88**	**44**	**60**	**71**	**57** ***+19***	**38.6** ***+15.6***
+10s	3	LaViLa VCLM	45	63	74	36	52	64	41	25.3
		LaViLa VCLM + CDP	**77**	**87**	**92**	**53**	**68**	**77**	**65** ***+24***	**47.1** ***+21.8***
+30s	7	LaViLa VCLM	47	66	76	38	55	67	43	27.2
		LaViLa VCLM + CDP	**86**	**92**	**95**	**66**	**80**	**85**	**76** ***+33***	**62.3** ***+35.1***

5 Experiments

5.1 Baselines

Our proposed method, Captioning by Discriminative Prompting (CDP), is applicable to any captioning model and does not fine-tune the captioner. We first focus our experiments on the SOTA baseline model for each benchmark (Zhao et al., 2023; Wang, 2024), giving results and examples. We then present results for other models 59 (Yu et al., 2023; Zhang et al., 2023; Zhao et al., 2023). For egocentric, we use the Visually-Conditioned-Language-Model (VCLM) from LaViLa (Zhao et al., 2023), which is specifically trained to caption egocentric clips. For the timeloop movie benchmark, we use Video-LLaMA (Zhang et al., 2023), which has recently been successfully used for describing movies (Song, 2023; Han et al., 2024). We use publicly available captioner checkpoints. Combined with our CDP, we demonstrate improved performance in every case.

CDP advances to the next clip in the continuous video when a unique caption cannot be predicted. For direct comparison, we evaluate all models on the same number of clips. $T = +0s$ indicates the models can only see one clip, which is 5s for egocentric and 2s for timeloop following model defaults. For the egocentric benchmark, $T = +5s$ indicates the methods are allowed access to the next $5s$ clip, $T = +10$ indicates access to the next two clips and so on. The timeloop movie benchmark, the equivalents are $T = +2s$ and $T = +4s$.

5.2 Implementation Details

We implement CDPNet (Ψ in Eq. 6) as a transformer encoder, with 2 layers, 4 heads and feedforward dimension of 1024. It is trained for 25 epochs using Adam with lr 0.0001, decaying by a factor of 10 at epochs 15 and 20, and batch size of 64. Default hyperparameter values are $\alpha = 3$ (Eq. 4) and $\lambda = 0.1$ (Sec. 3.3).

For the egocentric benchmark, we use EgoVLP 336px (Lin et al., 2022) as the evaluation network. For the captioner, we take the LaViLa VCLM-HR (Zhao et al., 2023), which is a GPT-2XL with trained cross attention layers and bridge to the visual features, and the Timesformer-L/Distilbert-Base 256d visual/text embedding network for training. Video clips are 5s long, from which 4 frames are uniformly sampled (the default with best performance). At maximum temporal extension, the inclusion of CDPNet and the search only increases captioning time from 4.5 to 5.8s, compared to 300s when we do not use CDPNet searching exhaustively. Note that these runtimes have the same memory budget – CDPNet has 1.6M trainable parameters vs 1.7B in LaViLa VCLM.

For timeloop movies, we use InternVideo (Wang, 2024) as the video/text evaluation space with 8×224px frames projected to a 768d feature. We use Video-LLaMA (Zhang et al., 2023) as the captioner, operating on 8×224px uniformly sampled frames (the default with best performance), and EVA-CLIP (Sun et al., 2023) video/text features as the embedding space during training.

5.3 Results

Table 1 shows results on the egocentric benchmark. At every timestep in every metric, CDP outperforms the captioner, LaViLa VCLM. The improvement is 8% Avg R@1 and 4% Cycle@1 acting on just the 5s clips without any advancement (*i.e.* $T = +0$). Whilst the LaViLa VCLM obtains small improvements with access to subsequent clips (*i.e.* as T increases), as expected, it is not able to pick out the specific aspects which make each clip unique. In contrast, the mechanism in CDP to find uniqueness provides greater improvements as more information becomes available over time. With additional access to the next clip (*i.e.* $+5s$),

Table 2 Timeloop movie benchmark using Video-LLaMA as the base captioner. is able to pick out more differences as the storylines diverge

T	#Clips	Method	Text→Video			Video→Text				
			R@1	R@2	R@3	R@1	R@2	R@3	Avg R@1	Cycle@1
		Chance	*16*	*32*	*48*	*16*	*32*	*48*	*16*	*3.0*
		Video-LLaMA	37	65	84	34	54	59	35	18.3
0s	1	Video-LLaMA + CDP	**47**	**71**	**86**	**36**	**63**	**78**	**42** ***+7***	**25.0** ***+6.7***
		Video-LLaMA	55	70	82	32	64	75	43	25.4
2s	2	Video-LLaMA + CDP	**51**	**70**	**81**	**45**	**64**	**75**	**48** ***+5***	**32.0** ***+6.6***
		Video-LLaMA	53	70	83	33	56	75	43	18.4
4s	3	Video-LLaMA + CDP	**62**	**77**	**85**	**44**	**68**	**84**	**53** ***+10***	**37.4** ***+19.0***
		Video-LLaMA	44	70	83	32	56	74	38	18.2
10s	5	Video-LLaMA + CDP	**73**	**87**	**95**	**53**	**75**	**84**	**63** ***+25***	**44.5** ***+26.3***

we have larger improvements of 19% Avg R@1 and 16% Cycle@1, with further gains as CDP is able to find uniqueness over more subsequent clips.

Table 2 shows results on timeloop movies with 2s clips. CDP is able to give a larger number of unique captions when allowed to advance through the story of each repetition. The base model Video-LLaMA struggles to generate unique captions at longer timescales because it is not conditioned on other clips, and thus has no mechanism to identify uniqueness. Because there is a large amount of information in longer clips, it ends up producing captions based on their most obvious properties, which tend to also be common between clips. With access to the next 2 clips (*i.e.* $+4s$), CDP improves Avg R@1 by 10% and Cycle@1 by 19%, with further gains over more clips.

5.4 Examples

Fig. 8 shows qualitative examples on egocentric footage. The unique captions generated by CDP are displayed to the right of each clip. 8a shows three clips captioned as "climbs the stairs". CDP is able to predict that all clips can be distinguished by the item the person is holding, and that there is no need to advance through the video. 8b shows a more challenging example, where the original query is "looks around the shelves". CDP captions the first two clips uniquely. For clip 1, CDP predicts the discriminative prompt "the other man...", which conditions the captioner to describe what the other person in the scene is doing. Clip 2 uses the discriminative prompt "looks at...", indicating the shopping list being read as unique for this clip. For clip 3, CDP cannot find a discriminative prompt at +0s or +5s, as captions would also apply to clip 2, and thus would not be unique. At +10s, it predicts uniqueness with the prompt "picks up". Fig. 8c shows an example where multiple prompts are required. Clip 1 and 2 can be distinguished with just one prompt, but clip 3 requires two prompts. Recall that CDP predicts the combination of prompts that will lead to the most unique caption (i.e. highest margin $\mathcal{M}$) in the visual/text space it was trained with, for a given number of prompts and temporal extent. If a unique-enough prompt combination is not found, increasing the number of prompts is attempted. In this case, CDP did not predict that the captioner, given the prompt "Picks up", would would give a high enough margin compared to clip 1. We can interpret this as "Picks up a plank of wood" on its own not being sufficient to distinguish clip 3, with confusion caused by there being planks of wood visible in clip 1. "Holds a nail" also describes clip 2, and thus the $\mathcal{M}$ for this caption is also low. CDP opts to combine both prompts and their combination is unique to clip 3.

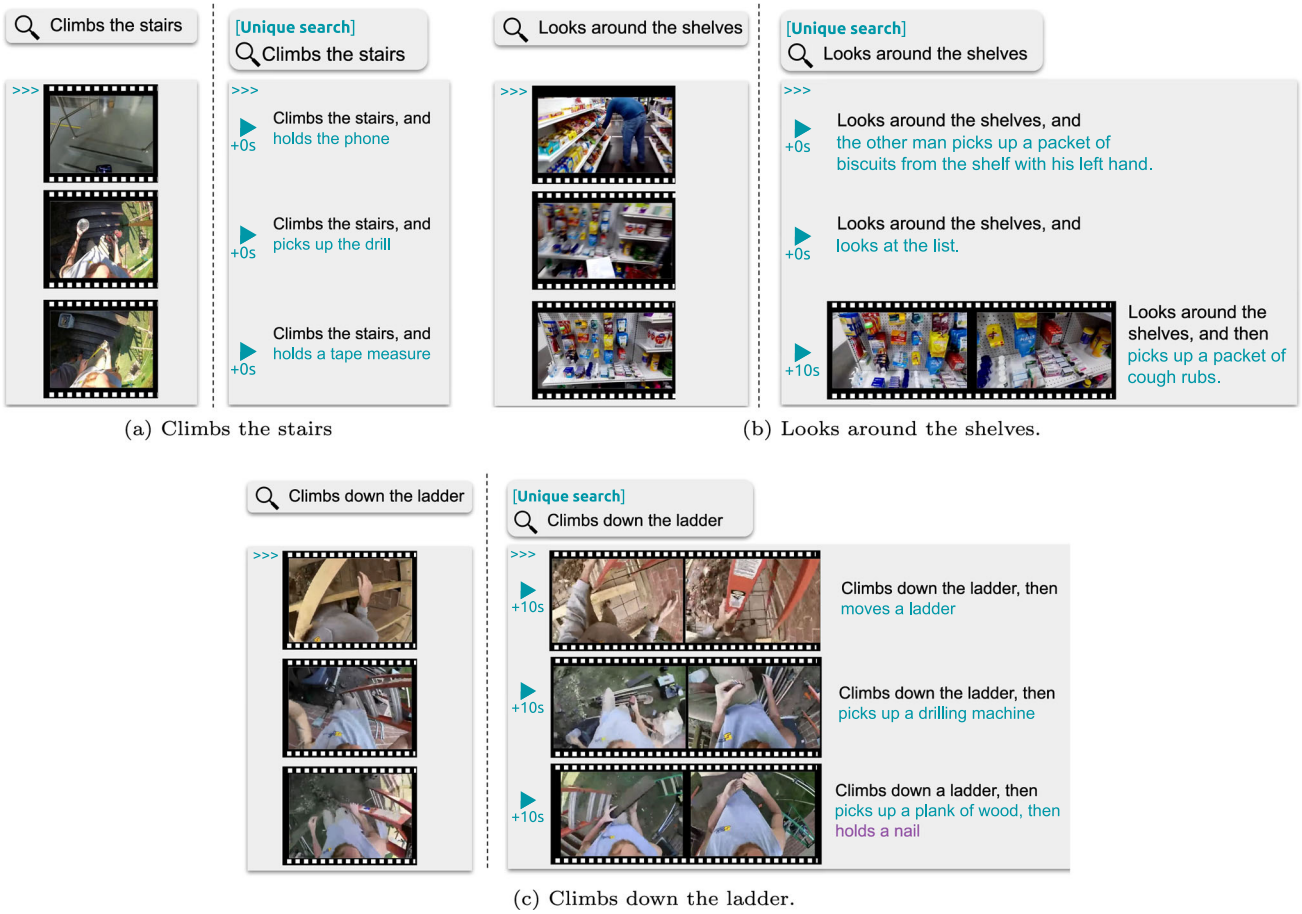
Fig. 9 shows unique captioning examples on three out of the eight instances of Tom Cruise sitting up in *Edge of Tomorrow*, when a timeloop begins. None of the clips can initially be distinguished. CDP finds uniqueness using the bus in clip 1 at 6s. In clip 2, the other soldiers in the scene at 4s. In clip 3, at 8s the woman with the backpack is not present in the other clips, resulting in a unique caption.

Fig. 10 shows examples on the seminal timeloop movie *Groundhog Day*. We show unique captions generated for three out of the nine loops which start with "a man wakes up". At 2s, clip 1 is the only one where the man is still lying down. After 6s, clip 2 is uniquely identified by the objects in the scene (windows). After 10s, CDP identifies that the other characters and location make it unique.

Fig. 11 shows examples of the protagonist waking up repeatedly in *The Map of Tiny Perfect Things*. In clip 1, CDP identifies them using their laptop in bed. In clip 2, their use of a phone is unique. In clip 3, it is the protagonist's father working on their laptop that is unique.

5.5 Additional Models

Having demonstrated that CDP improves the SOTA model for each benchmark, we now explore other captioners and



(a) Climbs the stairs

(b) Looks around the shelves.

(c) Climbs down the ladder.

Fig. 8 Qualitative egocentric examples. is able to caption the set in (a) uniquely. The third clip in (b) advances 10s to generate a unique caption. In (c), the first and second clips can be captioned uniquely with one prompt. The third clip requires two prompts to distinguish it from the first two, with the clauses shown in blue and purple

Fig. 9 Unique captioning example on Edge of Tomorrow (2014)

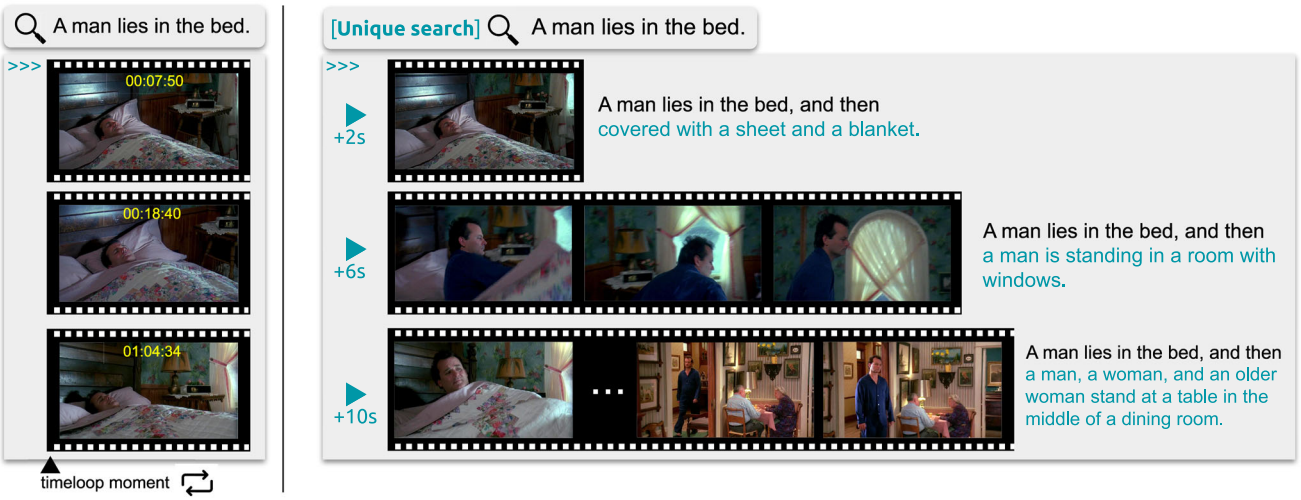


Fig. 10 Unique captioning example on Groundhog Day (1993)

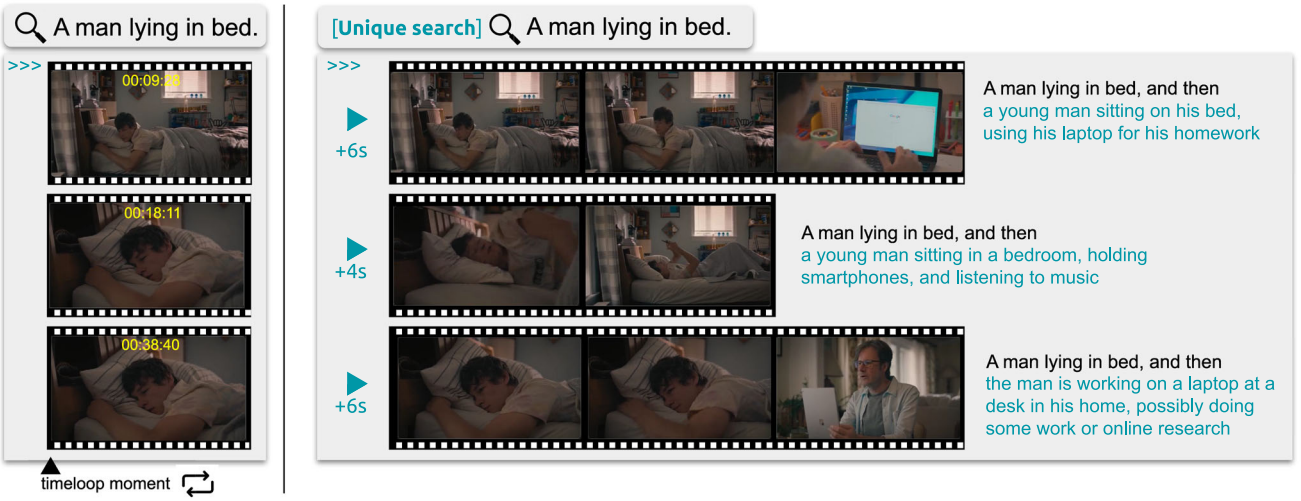


Fig. 11 Unique captioning example on The Map of Tiny Perfect Things (2021)

Table 3 Ablation on α, the maximum number of prompts. The LaViLa VCLM baseline is shown for comparison

T	# clips	Method	Prompts Max —	Chosen	V→T R@1	T→V R@1	Avg R@1	C@1
+0s	1	*LaViLa VCLM*	-	-	*45*	*54*	*50*	*34.3*
			1	1	45	63	54	37.2
		LaViLa VCLM + CDP	2	1.4	47	65	56	39.6
			3	1.6	**49**	**69**	**59**	**40.2**
+10s	3	*LaViLa VCLM*	-	-	*51*	*59*	*55*	*39.6*
			1	1	59	74	67	54.0
		LaViLa VCLM + CDP	2	1.5	**66**	**83**	**75**	**60.6**
			3	1.9	**66**	82	74	60.4

Table 4 Additional models and evaluation spaces on the egocentric benchmark. Avg. R@1 is shown.

Captioner	LaViLa V/T space T=0	T=5	T=10	T=30	EgoVLP V/T space T=0	T=5	T=10	T=30
EILEV Yu et al. (2023)	15	15	15	16	17	17	17	19
EILEV + CDP	**17**	**18**	**20**	**26**	**19**	**23**	**26**	**32**
TSF-B LaViLa VCLM Zhao et al. (2023)	30	34	34	37	32	34	37	38
TSF-B LaViLa VCLM + CDP	**36**	**48**	**54**	**67**	**37**	**50**	**56**	**67**
TSF-L LaViLa VCLM Zhao et al. (2023)	31	36	36	38	*37*	*38*	*41*	*43*
TSF-L LaViLa VCLM + CDP	**37**	**48**	**54**	**67**	***45***	***57***	***65***	***76***

Table 5 Additional models and evaluation spaces on the timeloop movie benchmark. Avg. R@1 is shown.

Captioner	CLIP V/T space T=0	T=2	T=4	T=10	InternVideo V/T space T=0	T=2	T=4	T=10
VideoBLIP [6]	25	24	30	33	33	32	35	36
VideoBLIP + CDP	25	24	**35**	**37**	33	**38**	**42**	**50**
VideoLlama Zhang et al. (2023)	**31**	35	36	33	*35*	*43*	*43*	*38*
VideoLlama + CDP	28	**36**	**41**	**42**	***42***	***48***	***53***	***63***

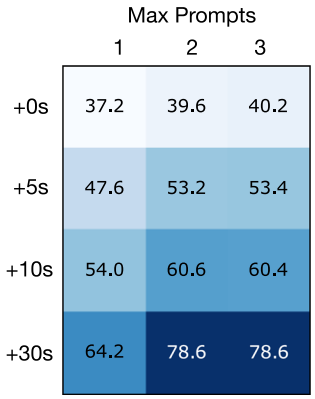


Fig. 12 Cycle@1 visualisation

embedding spaces, to show that CDP is not limited to specific models. Here we show Average Recall@1 with other captioners and embedding spaces. Table 4 shows results on the egocentric benchmark. Note that when evaluating a LaViLa VCLM variant in a LaViLa V/T space, we ensure they are not based on the same model. *i.e.* the default LaViLa V/T space is the Large variant, apart from the TFS-L LaViLa VCLM, which is evaluated in the Base space. This ensures a model is not evaluated with it's own features for fair comparison. Results in green indicate those corresponding to the Avg. R@1 column in Tab. 1.

Table 5 are results on the timeloop movie benchmark. CLIP averages features over all frames. Again, results in green indicate those corresponding to the Avg. R@1 column in Tab. 2.

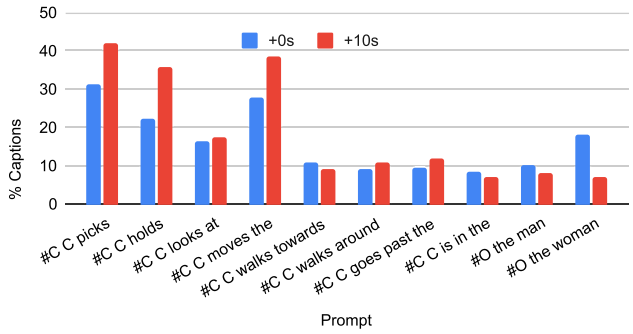
CDP delivers larger improvements on better base models. Better base models are more likely to give a correct caption grounded on the visual input when prompted. This is encouraging, as baseline models will improve over time, and indicates CDP will likely continue to be relevant.

Table 5 shows a slight drop in performance in the CLIP video-text evaluation space when CDP is applied to VideoLlama at $T = 0$. CLIP is not as well-suited as InternVideo for evaluating video-text similarities. It relies on appearance only, and our prompted captions contain actions which are likely to be confused in this space.

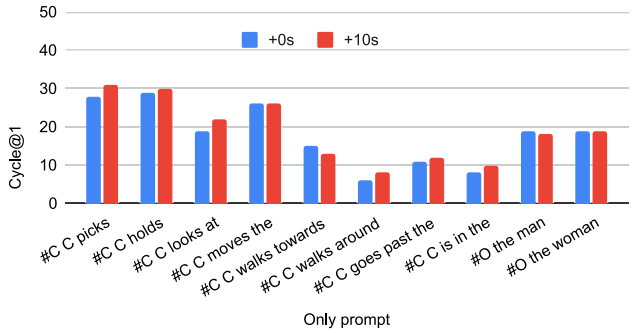
For all models without CDP, limited improvement in the unique captioning metrics is observed (and in some cases performance drops), when extending to larger T's. This is because just extending solely can produce a caption common to other clips. Whilst not necessarily incorrect captions, these generic captions are not unique to their clips, dropping performance in our unique captioning task.

5.6 Ablations

Ablations are performed on 50 egocentric sets of 10 clips.
Max allowed prompts α. We chose the maximum number of prompts per clip as $\alpha = 3$ for the main experiments as a reasonable trade-off between caption conciseness and retrieval accuracy. Table 3 shows results with T = +0s (one clip) and +10s (access to two subsequent clips) as we reduce α. We also record the average number of prompts chosen, as not every clip will require as many as α prompts. On just the first clip (T = +0s), $\alpha = 3$ provides the best results as expected, but with with access to more clips, performance saturates at

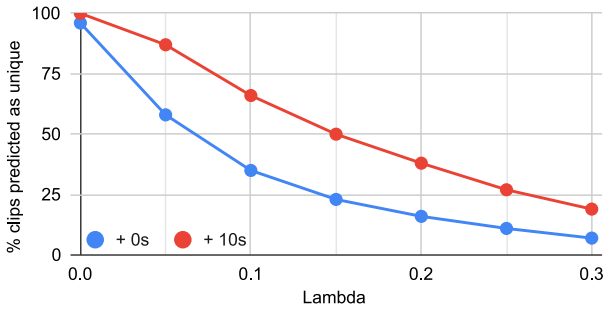


(a) % of captions each prompt is chosen for.

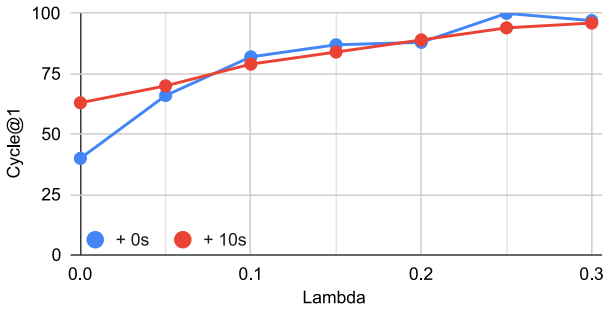


(b) Performance of each prompt individually.

Fig. 13 Prompt ablations, given at T= +0s and +10s



(a) Effect of λ on % clips predicted as unique.



(b) Effect of λ on Cycle@1 of clips predicted as unique.

Fig. 14 Ablation on the margin threshold λ, on T= +0s and T= +10s

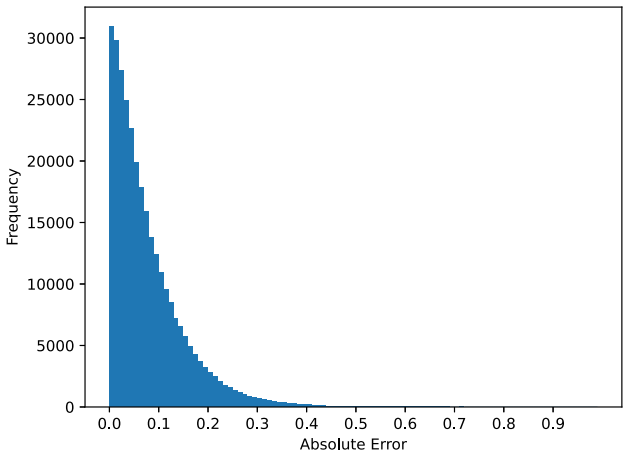


Fig. 15 Error of CDPNet on a held out validation set

$\alpha = 2$, as visualised in Fig. 12. With more time, it is more likely that unique properties will become available. Notably, CDP with one prompt outperforms the baseline.

Prompt ablation. To investigate the impact of each prompt, Fig. 13a shows when each prompt was chosen (adding up to > 100% due to combinations being chosen). Fig. 13b shows the performance of CDP with each prompt individually. Prompts relating to active object are chosen the most, and do best individually, as they are frequently the focus of egocentric videos.

Margin threshold λ. in Fig. 14a, we vary λ, and record the percentage of clips which have a margin $> \lambda$. As expected, a higher λ means fewer clips are predicted as unique. Naturally, the number of unique predictions for a fixed λ is greater for +10s than +0s, as CDP has more footage to identify uniqueness. We measure the Cycle@1 of clips with margin $> \lambda$ in Fig. 14b, where higher λ gives a higher Cycle@1. These results demonstrate that λ is a useful parameter to control the decision to advance time, and a proxy for prediction confidence.

Accuracy of CDPNet.

We use CDPNet in Section 3.3 to approximate the similarity between visual and text embeddings using only visual inputs and a prompt. We measure its accuracy by taking the absolute error between ground-truth video/caption cosine similarity (Eq. 2) and the predicted similarity by CDPNet (Eq. 6).

$$\text{absolute error} = |\hat{s} - s| \tag{7}$$

We evaluate on a held out validation set of egocentric footage, containing 30,000 clip/caption pairs, and plot a histogram of absolute errors in Figure 15. Most errors are small, and the error has mean = 0 and standard deviation = 0.11.

Table 6 Text → Video retrieval on long egocentric videos (average 40 minutes)

Method	R@1	R@2	R@3	R@5
LaViLa VCLM	12	20	26	33
LaViLa VCLM + CDP	**32**	**42**	**48**	**56**

5.7 Case Study: Long Egocentric Text-to-Video Retrieval

The main experiments were on sets of 10 identically narrated clips drawn from Ego4D, where we showed that CDP is able to significantly improve the retrieval performance of the LaViLa VCLM on this task. In this section, instead we experiment on one continuous long video, for the task of uniquely captioning every clip in that single video. We demonstrate the effectiveness of the captions produced by CDP by captioning every 5s clip in 10 long egocentric videos. We select 10 long videos from the Ego4D NLQ test set from different scenarios (lab work, cooking, sports, construction *etc.*). We break each video into consecutive 5s clips (*i.e.* clips are 0-5s, 5-10s, 10-15s...). The videos have an average length of 40.3 minutes, containing 483 clips each on average.

When attempting to caption, some clips will be similar producing identical captions. Temporally consecutive clips are especially challenging. We demonstrate how CDP can be used to improve retrieval with better captions, resulting in more effective text-to-video retrieval.

Experiment. We assess unique caption quality on the long video with Text→Video retrieval. We perform Text→Video retrieval in the joint video/text embedding space, where a text embedding is used as a query, and the result is the video with the closest embedding. For each clip, we generate its caption using either (i) LaViLa VCLM alone, or (ii) LaViLa VCLM with CDP. These captions are the text queries. We then attempt to retrieve each video clip by its generated caption in the shared video/text space, and measure Text→Video R@1, R@2, R@3 and R@5 retrieval. This is a good test of unique captioning, as better captions will obtain higher retrieval scores due to less confusion with clips they are not generated from. If a clip is not uniquely captioned, then multiple captions could refer to a single clip, giving lower retrieval scores. Both methods have access to T = +5s (*i.e.* the clip plus one subsequent clip). Note that we allow LaViLa VCLM to view both clips at once, as in the main experiments (as this performs better than just one clip).

Results. Table 6 shows Text → Video R@1, R@2 and R@3. CDP obtains an R@1 improvement of 21.5% compared to the LaViLa VCLM (36.0% compared to 14.5%), with larger gains for R@2 (+28.1%) and R@3 (+28.5%). Interestingly, CDP R@1 is higher than LaViLa R@3. Fig. 16 shows some of the example captions generated by CDP on a 36 minute egocentric video. Notice how it is able to distinguish between two clips at the sink, two clips pushing a trolley (one walking towards the exit, and one walking round a warehouse), and two clips using a pen.

Complexity. In Section 3.2, we discussed the complexity of the search. For a 40 minute video, using $\alpha = 3$ and 5s clips, the exact combinatorial search requires $< 1s$ on one CPU core. Even for a video 10x this length (6 hours), the search would take $< 30s$ on one CPU core, and is embarrassingly parallel. Our code is publicly available from the project's webpage.

6 Conclusion

In this paper, we introduced the problem of unique video captioning, to reflect the repetitive nature of daily life, the way repetitions are depicted in film, and the shortcomings of current methods to distinguish between these repetitive events.

We developed a framework, Captioning by Discriminative Prompting (CDP), based around observing all clips to be captioned. We introduced two benchmarks for unique captioning, based on egocentric footage and the repetitive moments in timeloop movies, and found CDP provides significant improvements on both.

There are a number of possible directions for future work. One would be learning prompts vs the fixed set used here, or a dynamic prompt bank where CDP could predict expanding the number of prompts to trial. Another would be to explore unique captioning across whole datasets. A third would be to incorporate multiple captioners with different specialisms.

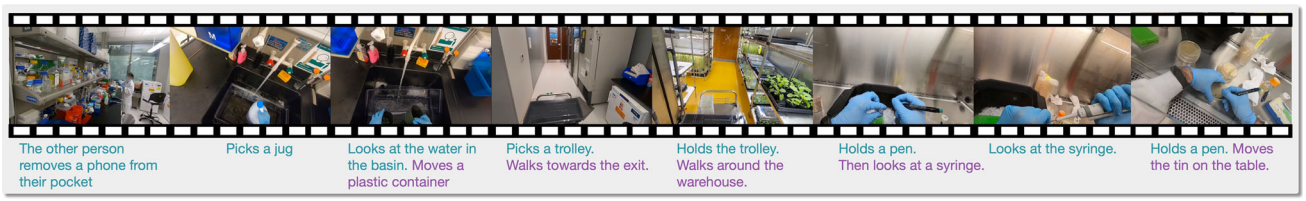


Fig. 16 Example captions generated by on a long 36 minute egocentric video in a lab

Acknowledgements Research is supported by EPSRC Programme Grant Visual AI (EP/T028572/1) and EPSRC UMPIRE (EP/T004991/1). This project acknowledges the use of the EPSRC funded Tier 2 facility, JADE-II.

Data Availability Code, benchmarks and data are publicly available at https://tobyperrett.github.io/its-just-another-day

References

Bagad, P., Tapaswi, M., & Snoek, C.G. (2023). *Test of time: Instilling video-language models with a sense of time*, 2503–2516.

Bain, M., Nagrani, A., Brown, A., & Zisserman, A. (2020). Condensed movies: Story based retrieval with contextual embeddings.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., & Guo, Y., et al. (2023). Improving image generation with better captions. *OpenAI dall-e-3*.

Carreira, J., & Zisserman, A. (2017). *Quo vadis, action recognition? a new model and the kinetics dataset*, 6299–6308.

Chum, O., Philbin, J., Isard, M., & Zisserman, A. (2007).*Scalable near identical image and shot detection*, 549–556 .

Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., & Price, W. et al. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision* 1–23.

Ding, N., et al. (2023). Image captioning with controllable and adaptive length levels. *IEEE Transactions on Pattern Analysis and Machine Intelligence.,46*(2), 764–779.

Duan, L.-Y., Wang, J., Zheng, Y., Jin, J. S., Lu, H., & Xu, C. (2006). *Segmentation, categorization, and identification of commercial clips from tv streams using multimodal analysis*, 201–210.

Dwibedi, D., Jain, V., Tompson, J.J., Zisserman, A., & Aytar, Y. (2024). *Flexcap: Describe anything in images in controllable detail*.

Goyal, R., Ebrahimi , K. S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., & Mueller-Freitag, M., et al. (2017). *The" something something" video database for learning and evaluating visual common sense*, 5842–5850.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., & Liu, X., et al. (2022). *Ego4d: Around the world in 3,000 hours of egocentric video*, 18995–19012.

Hampapur, A., et al. (2005). Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. *IEEE signal processing magazine,22*, 38–51.

Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., & Zisserman, A. (2023). *Autoad: Movie description in context*, 18930–18940 .

Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., & Zisserman, A. (2024). *Autoad iii: The prequel-back to the pixels*.

Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., & Russell, B. (2018). *Localizing moments in video with temporal language*, 1380–1390.

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). *Movienet: A holistic dataset for movie understanding*, 709–727 (Springer)

Islam, M. M., Ho, N., Yang, X., Nagarajan, T., Torresani, L., & Bertasius, G. (2024). Video recap: Recursive captioning of hour-long videos.

Lee, Y.J., Ghosh, J., & Grauman, K. (2012).*Discovering important people and objects for egocentric video summarization*, 1346–1353 (IEEE).

Lei, J., Yu, L., Bansal, M., & Berg, T.L. (2018). Tvqa: Localized, compositional video question answering. arXiv:1809.01696 .

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 12888–12900 (PMLR).

Lin, W., et al. (2024). Comparison visual instruction tuning. arXiv:2406.09240.

Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., & Wang, L. (2022). *Swinbert: End-to-end transformers with sparse attention for video captioning*, 17949–17958.

Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, EZ., Gao, D., Tu, R-C., Zhao, W., & Kong, W. (2022). *Egocentric video-language pretraining*

List of films featuring time loops. (2024). https://en.wikipedia.org/wiki/List_of_films_featuring_time_loops .

Liu, H., Zaharia, M., & Abbeel, P. (2023). Ring attention with blockwise transformers for near-infinite context. arXiv:2310.01889

Liu, Z., et al. (2022). Show, tell and rephrase: Diverse video captioning via two-stage progressive training. *IEEE Transactions on Multimedia.,25*, 7894–7905.

Long, Y., Wen, Y., Han, J., Xu, H., Ren, P., Zhang, W., Zhao, S., & Liang, X. (2023). *Capdet: Unifying dense captioning and open-world detection pretraining*, 15233–15243.

Lu, Y., et al. (2023). *Set prediction guided by semantic concepts for diverse video captioning*.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., & Zhou, M. (2020). Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv:2002.06353.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). *Howto100m: Learning a text-video embedding by watching hundred million narrated video clips*, 2630–2640 .

Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. arXiv:2111.09734.

Nagarajan, T., & Torresani, L. (2024). *Step differences in instructional video*, 18740–18750.

Ramakrishnan, S. K., Al-Halah, Z., & Grauman, K. (2023). Spotem: Efficient video search for episodic memory.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). *A dataset for movie description* .

Schaffalitzky, F., & Zisserman, A. (2003). Automated location matching in movies. *Computer Vision and Image Understanding,92*, 236–264.

Seo, P. H., Nagrani, A., Arnab, A., & Schmid, C. (2022). *End-to-end generative pretraining for multimodal video captioning*, 17959–17968.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). *Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning*, 2556–2565.

Sivic & Zisserman. (2003). *Video google: A text retrieval approach to object matching in videos*, 1470–1477 (IEEE).

Song, E. et al. (2023). Moviechat: From dense token to sparse memory for long video understanding. arXiv:2307.16449 .
Sun, Z., et al. (2023). Aligning large multimodal models with factually augmented rlhf. arXiv:2309.14525.
Sun, Q., Fang, Y., Wu, L., Wang, X., & Cao, Y. (2023). Eva-clip: Improved training techniques for clip at scale. arXiv:2303.15389
Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). *Movieqa: Understanding stories in movies through question-answering*, 4631–4640 .
Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). *Winoground: Probing vision and language models for visio-linguistic compositionality*, 5238–5248.
Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv:2401.06209.
Twenty questions. (2024). https://en.wikipedia.org/wiki/Twenty_questions.
Ventura, L., Yang, A., & Schmid, C. (2024). *& Varol, G*. CoVR: Learning composed video retrieval from web video captions.
Wang, Y., et al. (2024).*Internvid: A large-scale video-text dataset for multimodal understanding and generation* .
Wang, Z., Feng, B., Narasimhan, K., & Russakovsky, O. (2020). *Towards unique and informative captioning of images*.
Wang, Z., et al. (2022). Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems.,35*, 8483–8497.
Wu, C.-Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). *Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition*, 13587–13597.
Wu, W., Luo, H., Fang, B., Wang, J., & Ouyang, W. (2023). *Cap4video: What can auxiliary captions do for text-video retrieval?*, 10704–10713.
Xu, L.-Q., & Li, Y. (2003). *Video classification using spatial-temporal features and pca*, Vol. 3, III–485 (IEEE).
Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). *MSR-VTT: A large video description dataset for bridging video and language*, 5288–5296 .
Xue, Z., An, J., Yang, X., & Grauman, K. (2025). Progress-aware video frame captioning.
Yang, A., Nagrani, A., Seo, P. H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., & Schmid, C. (2023).*Vid2seq: Large-scale pretraining of a visual language model for dense video captioning*.
Yang, X., Zhang, H., & Cai, J. (2020). Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence,44*, 2313–2327.
Yu, K. P., Zhang, Z., Hu, F., & Chai, J. (2023). Efficient in-context learning in vision-language models for egocentric videos. arXiv:2311.17041 .
Yu, K.P. VideoBLIP. https://github.com/yukw777/VideoBLIP.
Yue, Z., Zhang, Q., Hu, A., Zhang, L., Wang, Z., & Jin, Q. (2023). Movie101: A new movie understanding benchmark. arXiv:2305.12140 .
Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv:2306.02858.
Zhao, Y., Misra, I., Krähenbühl, P., & Girdhar, R. (2023). *Learning video representations from large language models*.
Zhou, L., Xu, C., & Corso, J. (2018). *Towards automatic learning of procedures from web instructional videos*, Vol. 32 .