A Video examples

We provide a video on the project's webpage: http://tobyperrett.github.io/its-just-another-day, with 6 examples of Captioning by Discriminative Prompting (CDP) on the timeloop movie and egocentric benchmarks. In each case, we note the matching caption in black and the conditioned caption (by the chosen discriminative prompt(s)) in blue.

B Additional models

In the main paper, we presented results using the SOTA baseline model for each benchmark (egocentric and timeloop), and demonstrated that when incorporating CDP, results improve on both. Here we show Average Recall@1 with other captioners and embedding spaces.

Table 5 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 for 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 from the main paper.

	LaViLa V/T space				EgoVLP V/T space			
Captioner	T=0	T=5	T=10	T=30	T=0	T=5	T=10	T=30
EILEV [57]	15	15	15	16	17	17	17	19
EILEV + CDP	**17**	**18**	**20**	**26**	**19**	**23**	**26**	**32**
TSF-B LaViLa VCLM [60]	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 [60]	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 egocentric benchmark.

Table 6 are results on the timeloop movie benchmark. CLIP averages features over all frames. Again, results in green indicate those from the main paper.

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.

C Additional Benchmark Statistics

Fig. 11 explores the distribution of scenarios which appear in the egocentric benchmark. These roughly match the scenarios present in Ego4D. We also show a

	CLIP V/T space				InternVideo V/T space			
Captioner	T=0	T=2	T=4	T=10	T=0	T=2	T=4	T=10
VideoBLIP [56]	25	24	30	33	33	32	35	36
VideoBLIP + CDP	25	24	**35**	**37**	33	**38**	**42**	**50**
VideoLlama [59]	**31**	35	36	33	35	43	43	38
VideoLlama + CDP	28	**36**	**41**	**42**	**42**	**48**	**53**	**63**

Table 6: Additional models and evaluation spaces on the timeloop movie benchmark.

wordle of the narrations used to generate the benchmark in Fig. 12. Interestingly, movement features a lot (*e.g.* "walks", "around"), and it is often difficult for models to distinguish between different parts of an environment.

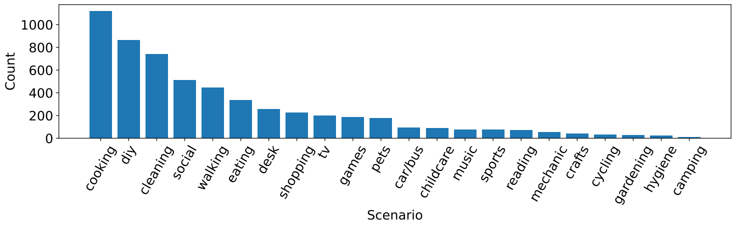

Fig. 11: Scenarios in the egocentric benchmark.

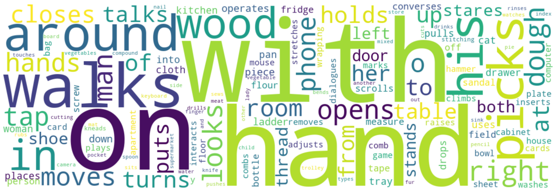

Fig. 12: Wordle of narrations used to create the egocentric benchmark.

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

In the paper, we evaluated unique captioning on sets of 10 identical narrated clips drawn from Ego4D, and showed that CDP is able to significantly improve the retrieval performance of the LaViLa VCLM on this task. The final ablation

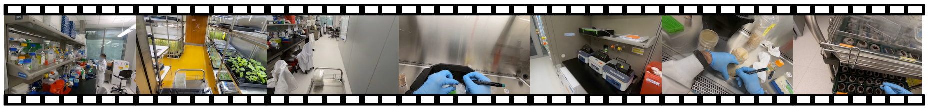

Fig. 13: Example of a long 36 minute egocentric video in a lab.

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

Table 7: Text → Video retrieval on long egocentric videos (average 40 minutes).

is expanded here, where we give an example retrieval use case on long egocentric videos.

We select 10 long videos from the Ego4D NLQ test set from different scenarios (lab work, cooking, sports, construction *etc.*). An example is shown in Figure 13. 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.

D.1 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).

D.2 Results

Table 7 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.

D.3 Complexity

In Section 3.2 of the main paper, 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.

E Accuracy of CDPNet

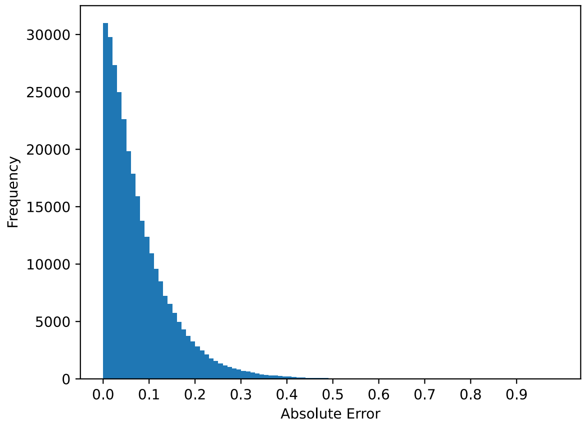

Fig. 14: Error of CDPNet on a held out validation set.

Figure 14 shows a histogram of the absolute errors of CDPNet, on a held out validation set of egocentric footage, containing 30,000 clip/caption pairs. The absolute error is the mean of the absolute difference between the ground-truth video/caption cosine similarity (Eq. 2 in main paper), and the predicted similarity by CDPNet (Eq. 6 in main paper):

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

The figure shows most errors to be small, and the error has mean = 0 and standard deviation = 0.11.

F Prompts

The 10 prompts used for the egocentric benchmark (ablated in Section 5.5) are:

- #C C picks
- #C C holds
- #C C looks at
- #C C moves the
- #C C walks towards the
- #C C walks around the
- #C C goes past the
- #C C is in the
- #O the man
- #O the woman

The 10 prompts used for timeloop movies are:

- Who is in the scene in this video?
- What is the man doing in this video?
- What is the woman doing in this video?
- Where are they in this video?
- What are they picking in this video?
- Who are they talking with in this video?
- What are they holding in this video?
- What are they looking at in this video?
- What are they moving in this video?
- Where are they going in this video?