A. Appendix

A.1. VidCab construction

We begin by sourcing video clips from EgoClip [2], excluding any videos associated with downstream tasks such as Egoschema [3]. Next, we clean the narrations by removing special tags like '#C C' and perform deduplication within each video, resulting in approximately 0.8M narrations. Using our Narration Pairing Encoding method, we generate a prefix set containing 0.6M entries and a postfix set with 5K entries, where the postfix is shared across all narrations and deduplicated. Finally, we create a training and evaluation split at a 10:1 ratio, referred to as VidCab-Train and VidCab-Eval, respectively.

A.2. Narration Pairing Encoding

In the above algorithm, we display the process of our Narration Pairing Encoding algorithm, which mainly includes two parts: **(i) Build Prefix Dictionary:** This step exhaustively enumerates all possible word combinations for each phrase to build a map between any prefix and the corresponding postfix narrations. **(ii) Extract Prefixes and Postfixes:** For each narration, we determine whether other narrations share its full prefix. If not, we add it to the prefix list. If they do, we extract and collect the differing postfixes from the narrations that share its prefix.

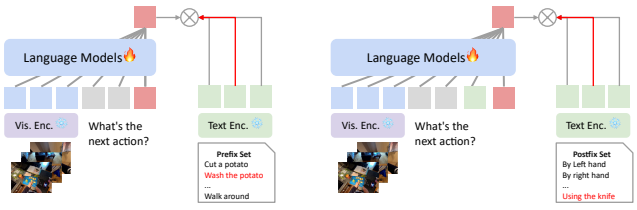


Figure 1. Illustration of VLog's progressively decode prefix and postfix vocabulary respectively.

In Fig. 1, we display how VLog progressively decode the prefix and postfix respectively. It first use the memory token to retrieve the prefix narration, and next it append the prefix narration and use the memory token to retrieve the postfix for a full narration.

A.3. Vocabulary Update Templates

Below, we attach the prompts for Qwen2.5 [4], which is used for produce narrations directly.

Algorithm 1 Narration Pairing Encoding

Require: List of narrations $\mathcal{N}$
Ensure: Prefix list $\mathcal{P}$, Postfix list $\mathcal{S}$

Step 1: Build Prefix Dictionary
Initialize prefix dictionary $\mathcal{D} \leftarrow$ empty
for all narration $n \in \mathcal{N}$ **do**
  Split n into words $[w_1, w_2, \ldots, w_k]$
  for $i = 1$ to k **do**
    Prefix $p \leftarrow w_1\, w_2\, \ldots\, w_i$
    Add n to $\mathcal{D}[p]$
  end for
end for
Step 2: Extract Prefixes and Postfixes
Initialize prefix list $\mathcal{P} \leftarrow \emptyset$, postfix list $\mathcal{S} \leftarrow \emptyset$
for all narration $n \in \mathcal{N}$ **do**
  if $\mathcal{D}[n]$ contains only n **then**
    Add n to $\mathcal{P}$
  else
    for all other narration $s \in \mathcal{D}[n]$ **do**
      if $s \neq n$ **then**
        Get suffix $t \leftarrow$ Remove prefix n from s
        Add t to $\mathcal{S}$
      end if
    end for
  end if
end for

List possible short actions that could take place in the scene. Write each action as a short narration (a verb with a noun). Separate by ';' The following is examples.
scene: In the heart of the kitchen, a man skillfully slices into a ripe mango, its golden flesh gleaming under the light.
narration: Slice mango; Hold knife; Cut mango; Place seed; Wipe counter; Drop pieces; Grip mango; Rest knife; Smell mango; Gather chunks.

scene: A woman sits by the fireplace, knitting a scarf as the flames crackle warmly in the background.
narration: Knit scarf; Hold needles; Loop yarn; Adjust thread; Pull stitch; Rest hands; Drop yarn; Smell smoke; Listen flames; Rub hands; Fold scarf; Gather wool; Stare fire; Sit still; Tap needle.

scene: {scene}
narration:

The {scene} is output by LLaVA-OV-0.5B [1] with prompt: "What is the overall activity in the scene? Answer briefly in one sentence."

A.4. Experimental Settings

Our SigLIP model [5] is based on `google/siglip-so400m-patch14-384`. During training, we fully fine-tune the GPT-2 model with a batch size of 32, a learning rate of 3e-4, and a sampling rate of 8 frames per short video clip. For long videos, such as those in the EgoSchema dataset [3], we do not

compress the entire video into a single embedding. Instead, we uniformly sample long videos into multiple fixed-length clips (1 second each) and process them in a streaming fashion.

A.5. Complexities Analysis

Let us clarify each term when generating N narrations: **(i) Encoding:** We embed the entire vocabulary once and then reuse it – $O(1)$. **(ii) Decoding:** This should be $O(\alpha N)$, where α is the speed per decoding step. **(iii) Upgrading** (optionally): $O(C)$, where C is the upgrade times ($C \ll N$). For a large N, the overall complexity $O(1)+O(\alpha N)+O(C) \rightarrow O(\alpha N)$ *remains efficient as the encoding and upgrading costs become negligible.* Below is the timing analysis on 4.6K VidCap-Eval:

Models	Vocab. size	**Encoding**(s)	**Decoding**(s)	**Upgrading**(s)	**R@1**	**Total (s)**
GPT-2	32K	–	207.8	–	7.9	207.8
VLog	4.6K	3.6	10.4	–	12.4	**14.0**
VLog	4.6K (+486)	3.6	10.5	38.4	**13.7**	52.4

A.6. Subwords *v.s.* Narration Vocabulary on Easy *v.s.* Complex tasks?

Our VLog is prioritize task-specific efficiency over generalist models. We compare the two in the below Table.

	Domain	Vocabulary	Backbone	Decoding	Highlights
VideoLLMs	General	Subwords	LLMs (2B+)	Token Gen.	Multi-Purpose
VLog	**Specific**	**Narrations**	GPT-2 (**345M**)	**Retrieval**	**Efficiency**

Whether 'Subwords-' or 'Narration-' Vocabulary is depends on how tasks define minimal semantic units for videos. Subwords capture every detail but may be redundant for long videos, while narrations offer event contexts quickly but may miss finer details. To balance expressive granularity and efficiency, an idea is to cooperate two fashions like our vocab. upgrading or retrieve narration first and then generate minimal subwords as needed. We are interested in further exploring the latter.

A.7. Improvment by Stronger LLM

We chose GPT-2 because its *simplicity and lightweight nature* make it a representative baseline. To demonstrate scaling, we upgraded GPT-2 to Qwen2-7B, resulting in significant performance gains, and beat its comparable baseline Qwen2-VL-7B.

Models	Size	EgoSchema QA val.	Decoding Time (s)
Qwen2-VL	7B	72.8	79.4
VLog (GPT-2)	345M	70.4	**2.3**
VLog (Qwen2)	7B	**74.8**	6.4

B. Qualitative Examples

B.1. VLog for Reasoning Retrieval

In Fig.3, we illustrate how VLog retrieves the vocabulary (blue indicating prefixes and green indicating postfixes) conditioned on different queries. For instance, in example (b), the query "What is the next activity in the video?" retrieves "Grab a bag of chips using the left hand" as a result, while the query "What is the previous activity in the video?" retrieves "Adjust the steering wheel using the hand" as a result, demonstrating VLog 's capability to infer relationships between sequential events.

B.2. How does Vocabulary Updating work?

In Fig.4, we demonstrate how VLog's vocabulary updating process effectively expands its descriptive range. Given the first frame of a video clip, LLaVA-OV [1] generates an initial brief description, which is then passed to Qwen2.5 [4] to imagine and expand possible vocabulary terms. For instance, in (a), LLaVA-OV identifies a simple construction project involving multiple yellow pencils, and Qwen2.5 extends this by generating potential actions such as "Arrange pencils" and "Hold pencils," which collectively capture most events in the video.

However, limitations still exist with the models. For example, in (c), while the activity "Make Pineapple Fritters" is identified, the model struggles to detect the specific ingredient "pineapple," making it challenging for the expanded vocabulary to recognize or describe the desired object accurately. These challenges highlight areas for improvement in object-specific vocabulary generation.

B.3. Limtation by VLog.

Broad Descriptive Range

A mother helping her child with homework.

Character Information

Hermione told Harry a piece of news that surprised Harry.

Figure 2. VLog still fail to capture the video with broad descriptive range or high-level information *e.g.* characters.

We acknowledge that VLog still has limitations, as illustrated in 2. For example, when videos have a broad expressive range, such as those involving multiple individuals or focusing on different aspects depending on subjective interpretation, it becomes challenging to rely on a narration-wise closed-set vocabulary. Additionally, in more complex scenarios, such as movies, where character information and di-

What is the activity in the video?
Making Bricks
What is the activity in the video?
(Hit the bricks with mallet) (using the right hand)
What is the next activity in the video?
(Arrange brick) (using the left hand)
What is the previous activity in the video?
(Pour out the soil) (with both hands)

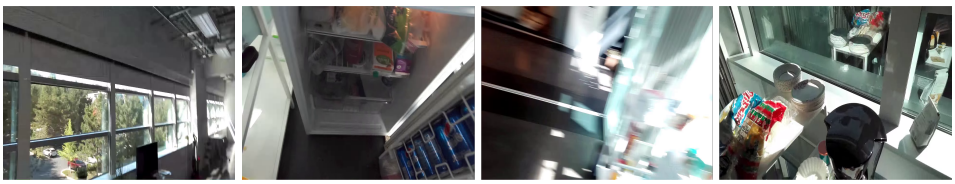
(a)

What is the activity in the video?
Grocery shopping indoors
What is the activity in the video?
(Look at snacks) (on a shelf)
What is the next activity in the video?
(Grab a bag of chips) (using the left hand)
What is the previous activity in the video?
(Adjust the steering wheel) (using the hand)

(b)

What is the activity in the video?
Bike mechanic
What is the activity in the video?
(Adjust a wheel) (on a table)
What is the next activity in the video?
(Place down a tire) (on a table)
What is the previous activity in the video?
(walks around) (a garage)

(c)



What is the activity in the video?
Indoor Navigation
What is the activity in the video?
(Adjust a wheel) (on a table)
What is the next activity in the video?
(Place down a tire) (on a table)
What is the previous activity in the video?
(walks around) (a garage)

(d)

Figure 3. **VLog enables retrieval through reasoning**, conditioned on different queries. Blue represents prefixes, while green represents postfixes.

alogues play a central role, the current approach struggles. These cases may require a return to a generalist model capable of handling subword tokens for richer representations.

References

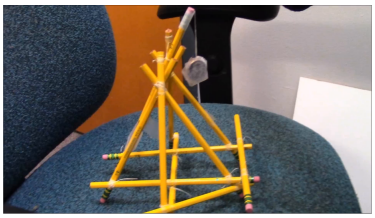
[1] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[2] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[3] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023.

[4] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[5] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.



(a): Build a Trebuchet

LLaVA-OV-0.5B:
The overall activity in the scene appears to be a simple construction project involving multiple yellow pencils. The image shows the pencils arranged in a triangular formation, with one pencil on top of another, suggesting a structure or object that might be crafted or assembled using these pencils.

Qwen2.5-0.5B:
'Arrange pencils; Hold pencils; Position pencils; Arrange vertically; Rotate pencils; Use pencil sharpener; Apply pressure; Press and hold pencil; Move pencil up and down; Press and hold pencil again; Push pencil back down; Check if pencils are aligned correctly; Continue until complete; Clean up workspace; Repeat process.'

(b): Fold an Origami Parrot

LLaVA-OV-0.5B:
The overall activity in the scene is a person holding and working on a paper flower from a paper bird origami kit

Qwen2.5-0.5B:
'Hold paper flower; Unfold paper; Shape flower petals; Arrange petals; Pinch stem; Secure flowers; Review design; Repeat assembly; Take final step; Gently unfold.'

(c): Make Pineapple Fritters

LLaVA-OV-0.5B:
The overall activity in the scene is cooking, specifically preparing batter for frying or baking.

Qwen2.5-0.5B:
'Mix dough; Pour flour into bowl; Stir until smooth; Add water; Knead dough; Place dough in bowl; Cover and knead; Roll out dough; Heat oil in pan; Place dough in hot oil; Fry or bake; Serve; Repeat process if needed; Cook thoroughly; Allow to cool; Serve immediately.'

(d): Cut a Doll's Hair

LLaVA-OV-0.5B:
The overall activity in the scene involves creating a little twiisesie texture on a doll by using bantu knots.

Qwen2.5-0.5B:
'Knit bantu knot; Hold doll; Wrap loop; Insert loop; Twist loops; Pinch fingers; Pull tight; Gently pull; Rest doll; Repeat; Create twiisesie pattern; Gently stretch; Stretch doll; Adjust tension; Press firmly; Rest doll.'

Figure 4. **Illustration of VLog's vocabulary updating process.** Given the first frame of a video clip, LLaVA-OV [1] provides a brief initial description, which is then passed to Qwen2.5 [4] to generate and expand the possible vocabulary.