	EgoSchema Subset	EgoSchema Full Set
All tasks	40.4	41.2
without Multi-choice Tideo QA	24.3	24.7
without Tideo Summarization	35.6	38.0

Table 1: Ablations on proposed learning tasks.

Inference Strategy	TOPA-LLama2-7B	TOPA-LLama2-13B
visual features	30.6	38.3
visual-textual projected features	41.2	51.0
textual features (frame-level captions)	41.4	51.1
textual features (clip-level captions)	42.9	52.6

Table 2: Study on zero-shot inference strategies. For textual feature inference, we use LaViLa-xl as captioner.

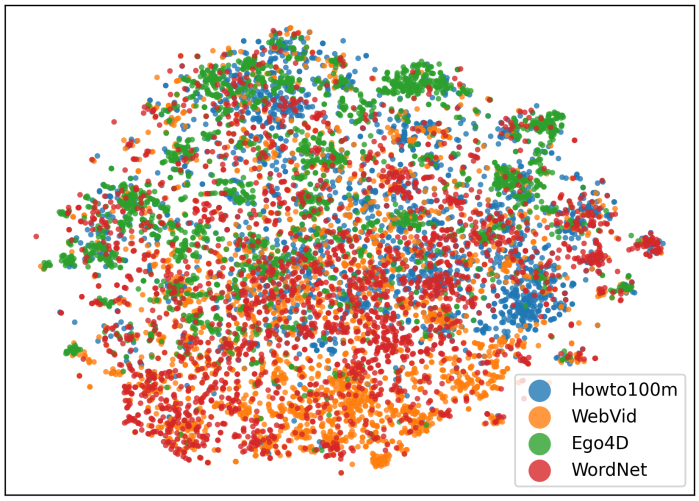

Figure 1: Visualization of Tideo features generated from different type of prompts.

	Howto100m	Ego4D	WebVid	WordNet
Vocab Size	17492	7320	15095	26486

Table 3: Vocabulary size of Tideos generated under different prompts. We randomly sampled 20,000 global captions from each type of Tideos for comparison.

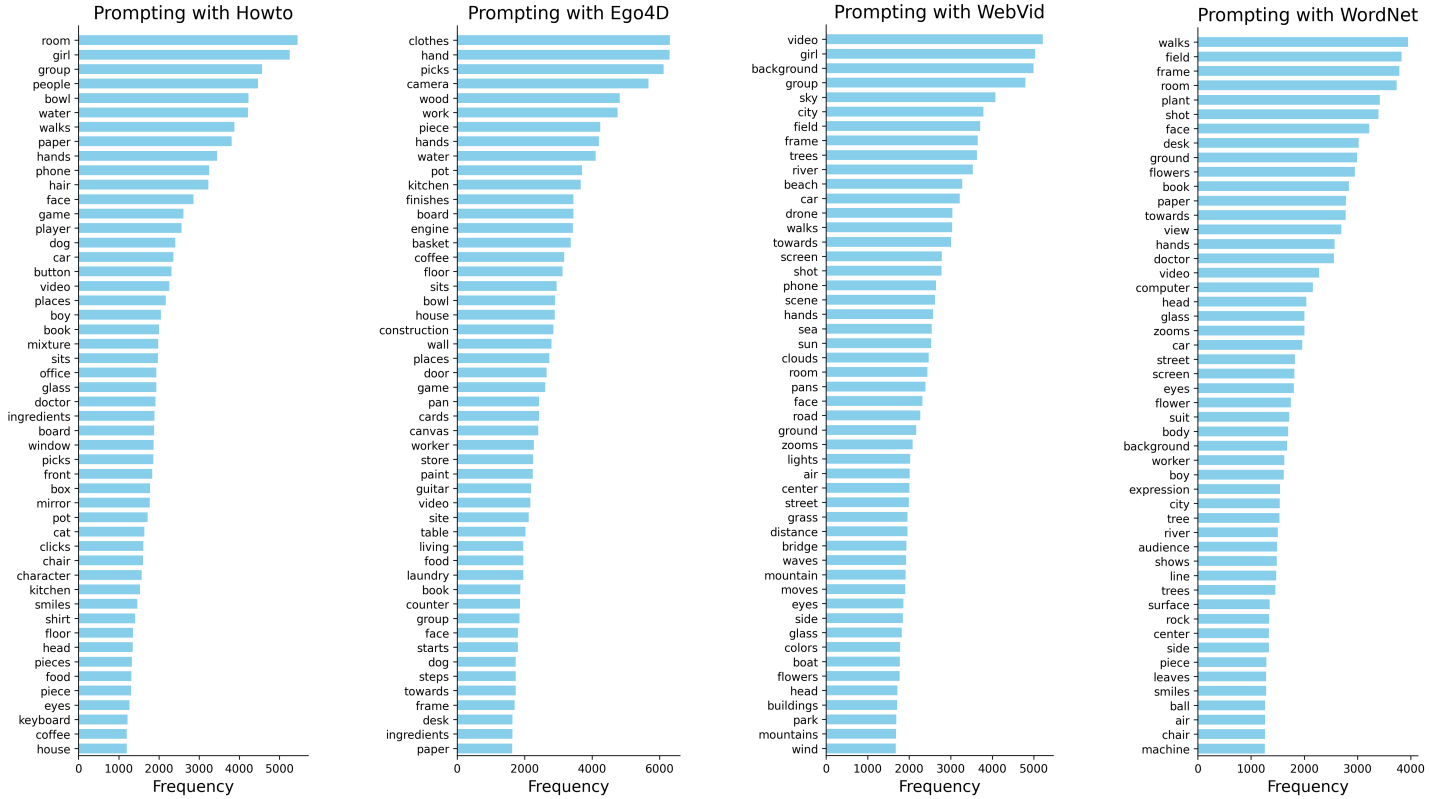

Figure 2: Word frequency of Tideos generated with different type of prompts.