The Devil is in the Distributions: Explicit Modeling of Scene Content is Key in Zero-Shot Video Captioning

Mingkai Tian[1] **Guorong Li**[1*] **Yuankai Qi**[2] **Anton van den Hengel**[3] **Qingming Huang**[1]

[1]School of Computer Science and Technology, University of Chinese Academy of Sciences
[2]Macquarie University
[3]Australian Institute for Machine Learning, The University of Adelaide

liguorong@ucas.ac.cn

Abstract

Zero-shot video captioning requires that a model generate high-quality captions without human-annotated video-text pairs for training. State-of-the-art approaches to the problem leverage CLIP to extract video-informed text prompts to guide language models in generating captions. However, by using representations at a single granularity (*e.g.,* noun phrases or full sentences), these methods tend to focus on one key aspect of the scene and build a caption that ignores the rest of the visual input. To address this issue, and generate more accurate and complete captions, we propose a novel progressive multi-granularity textual prompting strategy for zero-shot video captioning. Our approach constructs three distinct memory banks, encompassing noun phrases, scene graphs of noun phrases, and entire sentences. Moreover, we introduce a category-aware retrieval mechanism that models the distribution of natural language surrounding the specific topics, to promote prompt diversity while ensuring visual relevance. Extensive experiments on both in-domain and cross-domain settings demonstrate that the proposed method consistently outperforms state-of-the-art approaches.

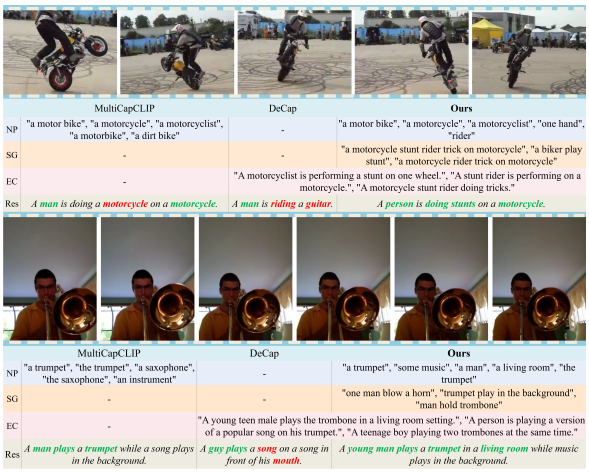

Figure 1: Our method generates more accurate and comprehensive descriptions compared to current zero-shot captioning methods. NP, SG, and EC denote noun phrase, scene graph (triplets are displayed as concatenated strings), and entire caption prompt, respectively. "Res" indicates the generated captions, with correct and incorrect words highlighted in green and red. In the top example, MultiCapCLIP (Yang et al., 2023) fails to capture the rider-motorcycle interaction. In the bottom example, MultiCapCLIP's top-K retrieval strategy produces repetitive similar noun phrases and lacks person and environment information. DeCap (Li et al., 2023a) struggles to fully understand the video details due to its coarse-grained prompt of global caption embedding.

1 Introduction

Video captioning is the task of generating text descriptions for video content, serving as a bridge between vision and language. Directly modeling associations between modalities is vulnerable to missing the structure in either, resulting in captions that focus on a single scene element. We address this by modeling both scene and language structure at multiple semantic levels, enabling more comprehensive and context-aware caption generation.

Traditional supervised video captioning methods (Zheng et al., 2020; Ye et al., 2022; Lin et al., 2022; Tian et al., 2024) utilize an encoder-decoder architecture trained on large-scale, manually annotated video-text pairs. The encoder leverages pre-trained convolutional neural networks, while the decoder uses LSTMs (Hochreiter and Schmidhuber, 1997) or Transformers (Vaswani et al., 2017). Despite achieving remarkable performance, their reliance on human-labeled data pairs constrains real-world scalability. To address these limitations, zero-shot video captioning has emerged to generate descriptions without relying on video-text pairs for training. Existing methods mainly include training-free approaches and text-only trained models.

Training-free methods leverage pre-trained

*Corresponding author

Findings of the Association for Computational Linguistics: EACL 2026, pages 1916–1929
March 24-29, 2026

vision-language models (*e.g.,* CLIP (Radford et al., 2021)) to guide pre-trained language models (*e.g.,* GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019)) during inference. For instance, Tewel et al. employ randomly initialized pseudo-tokens and prefix prompts ("Video of") to facilitate text generation by GPT-2, updating pseudo-tokens based on gradients from CLIP's cross-modal similarity. However, visual supervision after token generation can lead to the language model's priors dominating the captioning process, resulting in hallucinations unrelated to the video content.

The other line of work trains a text decoder exclusively on text corpora. Textual units (*e.g.,* nouns, noun phrases, complete sentences) are extracted from the training corpus to form various memory banks. During training, embeddings of textual units serve as prefix prompts for the text decoder to reconstruct the original caption. At inference, visual features are used to retrieve relevant textual semantic units from the memory bank via CLIP similarity, which are then fed into the text decoder. For example, MultiCapCLIP (Yang et al., 2023) stores noun phrases and retrieves top-16 elements as prompts, while DeCap (Li et al., 2023a) stores full captions and uses a global embedding as the prefix. Other text-only trained zero-shot image captioning methods, such as MeaCap (Zeng et al., 2024), retrieves top-5 relevant descriptions, and feeds parsed key entities to a pre-trained language model.

Despite recent progress, existing methods remain limited in prefix prompt construction and retrieval strategies. They typically use single-granularity textual units, such as nouns, noun phrases, or complete sentences, as prompts, failing to fully exploit multi-granularity textual units to provide rich information for the language model. While noun phrases provide more attributes than simple nouns, they lack inter-entity interactions. In addition, global sentence embeddings may dilute fine-grained details. As illustrated in Figure 1 (top), MultiCapCLIP misses the rider's stunt action providing only noun phrase prompts, whereas DeCap fails to identify the "motorcycle". Moreover, simple top-K most similar retrieval tend to yield semantically repetitive elements, reducing the diversity of the prompts and the accuracy of the generated captions. In Figure 1 (bottom), MultiCapCLIP repetitively prompts with musical instrument phrases but neglects information about the person and the environment.

To address these challenges, we propose a progressive multi-granularity textual prompting strategy. We construct three memory banks comprising noun phrases, scene graphs incorporating noun phrases, and entire sentences, ensuring the text decoder receives comprehensive semantic cues. Existing parsers (Li et al., 2023b) typically extract scene graphs with noun-only nodes, while we propose a text-similarity-based approach to enhance initial scene graphs by incorporating noun phrases with additional attributes wherever possible. We further develop category-aware retrieval mechanism with top-p filtering for noun phrases and scene graphs, ensuring both diversity and visual relevance. Figure 1 shows our generated captions.

Overall, our main contributions are as follows:

- We propose a progressive multi-granularity textual prompting strategy, providing the language model with comprehensive semantic information at varying levels of abstraction.
- We introduce a category-aware retrieval method with top-p post-processing for semantic units, enhancing the diversity and relevance of the prompts during inference.
- Extensive experiments on the MSR-VTT, MSVD, and VATEX benchmarks demonstrate the effectiveness of our method, achieving 5.7%, 16.2%, and 3.4% CIDEr improvements over state-of-the-art methods with the same pre-trained backbone.

2 Related Work

Frozen vs. Trainable Language Models Both training-free and text-only training methods for zero-shot video captioning rely on pre-trained language and vision-language models. Training-free methods keep them frozen. ZeroCap (Tewel et al., 2022) and its variant for zero-shot video captioning (Tewel et al., 2023) employ GPT-2 (Radford et al., 2019) to iteratively predict new tokens, with cross-modal similarity calculation via CLIP (Radford et al., 2021) after each token is generated. ConZIC (Zeng et al., 2023) utilizes a pre-trained BERT (Devlin et al., 2019) for Gibbs sampling. In contrast, text-only training approaches (Nukrai et al., 2022; Fei et al., 2023; Zeng et al., 2024; Yan et al., 2025) fine-tune the weights of GPT-2 or CBART (He, 2021), or train transformers from scratch (Li et al., 2023a; Yang et al., 2023).

Textual Memory Bank Text-only training methods often employ a textual memory bank for ef-

ficient storage and rich semantics. ViECap (Fei et al., 2023) and EntroCap (Yan et al., 2025) build a memory bank of object class names from Visual Genome (Krishna et al., 2017) and retrieve relevant nouns with CLIP during inference. DeCap (Li et al., 2023a) and MeaCap (Zeng et al., 2024) build memory banks with training captions, employing sentence-level and core noun embeddings as prefix prompts, respectively. MultiCapCLIP's (Yang et al., 2023) memory bank is composed of the 1000 most frequent noun phrases parsed from training captions, allowing the decoder to generate sentences from concept prompts. However, existing methods have not fully explored the potential of multi-granularity semantic guidance for caption generation. To address this, we construct three hierarchical memory banks to ensure comprehensive information for the language model, leading to excellent experimental performance.

3 Method

As shown in Figure 2, our approach includes three key processes: (1) Memory Bank Construction (top-left): constructing memory banks at three progressive granularities from captions; (2) Training Process (top-right): retrieving prompts from memory banks using perturbed text embeddings; (3) Inference Process (bottom): generating diverse, visually-relevant prompts via category-aware retrieval with top-p refinement during inference.

3.1 Multi-Granularity Memory Bank Construction

Our method constructs three distinct memory banks to capture progressive multi-granularity semantics, which are used to obtain prompts during the caption generation process: noun phrases, scene graphs incorporating noun phrases, and entire sentences.

Noun Phrase Memory Bank ($\mathcal{M}_{\text{NP}}$) Let $\mathcal{S}$ represent the set of all captions in the training split. For each caption $S \in \mathcal{S}$, we identify all the noun phrases from S using SpaCy[1], forming a set $\mathcal{P}(S)$. The complete set of noun phrases from the training corpus is denoted as:

$$\mathcal{P} = \bigcup_{S \in \mathcal{S}} \mathcal{P}(S). \tag{1}$$

We then rank the noun phrases in $\mathcal{P}$ based on their frequency of occurrence and retain the top-N_p most frequent noun phrases to construct the noun phrase memory bank:

$$\mathcal{M}_{\text{NP}} = \{p_1, p_2, \cdots, p_{N_p}\}, \tag{2}$$

where p_i denotes the i-th most frequent noun phrase. These noun phrases provide fundamental object-level semantics, serving as the basic building blocks of textual prompts.

Scene Graph Memory Bank ($\mathcal{M}_{\text{SG}}$) Scene graphs are crucial for capturing the relationships between entities within a video. To build the memory bank, we first utilize an off-the-shelf textual parser (Li et al., 2023b) to extract basic scene graphs from each caption. Initially, the results include triples of the $\langle \textit{subject}, \textit{predicate}, \textit{object} \rangle$ form, such as $\langle \textit{boy}, \textit{play}, \textit{basketball} \rangle$ for the sentence "A young boy is playing basketball". We enhance the scene graph by transforming it from being based solely on nouns to incorporating noun phrases wherever possible. For the objects at the beginning and end of the initial scene graph, if attribute information exists in the caption, we include it in. In the example above, we identify "young boy" as a noun phrase and transform the initial scene graph into $\langle \textit{young boy}, \textit{play}, \textit{basketball} \rangle$.

For the i-th basic scene graph $g_i = \langle \textit{sub}_i, \textit{pred}_i, \textit{obj}_i \rangle$ extracted from the caption S, we first identify all noun phrases in $\mathcal{P}(S)$ that contain $\textit{sub}_i$, and combine them with $\textit{sub}_i$ itself to form the set $\mathcal{A}_i$. Similarly, we form the set $\mathcal{B}_i$ for $\textit{obj}_i$. These sets are then used to create the enhanced scene graph set $\mathcal{X}_i$, where each enhanced graph is a triple of the form $\langle a, \textit{pred}_i, b \rangle$, with $a \in \mathcal{A}_i$ and $b \in \mathcal{B}_i$. Next, with the embeddings of S and all of the enhanced scene graphs in $\mathcal{X}_i$, denoted as $\mathbf{E}_S$ and $\mathbf{E}_{\mathcal{X}_i}$ respectively, all produced by BGE (Xiao et al., 2024) (we encode a scene graph by encoding the string formed by concatenating the subject, predicate, and object with a single space), we calculate the cosine similarity between them. The enhanced scene graph x_{best} with the highest cosine similarity to S is selected as the final improved representation of the original scene graph g_i. Finally, we collect the enhanced scene graphs from all captions and select the top-N_g most frequent to form the scene graph memory bank $\mathcal{M}_{\text{SG}}$. This is used to provide richer and more semantically informative prompts for caption generation.

$$\mathcal{M}_{\text{SG}} = \{x_1, x_2, \ldots, x_{N_g}\}, \tag{3}$$

where x_i denotes the i-th most frequent enhanced

[1] https://spacy.io

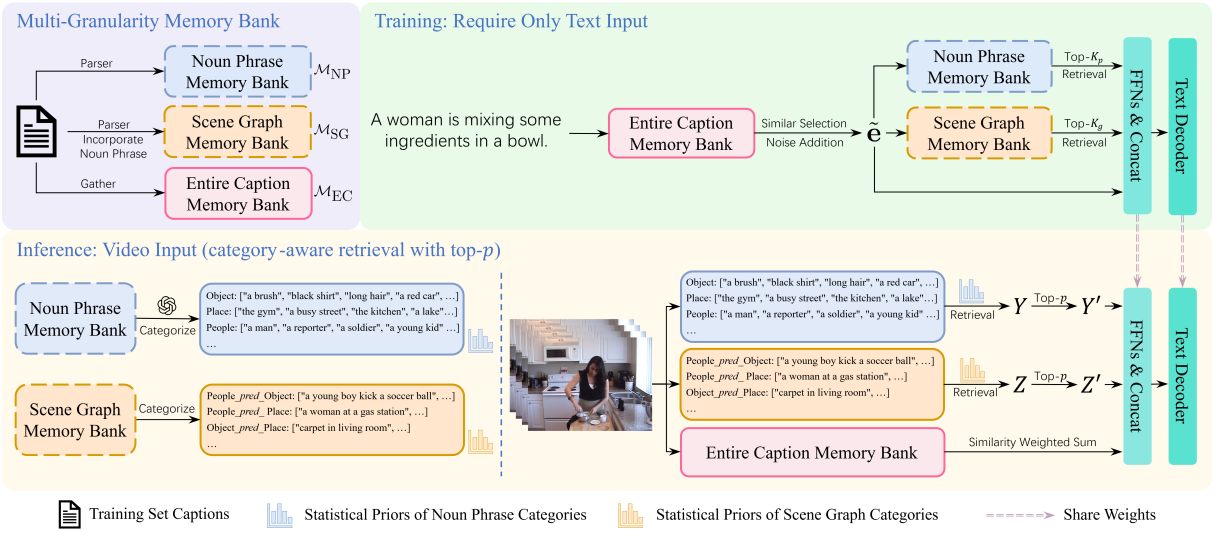

Figure 2: We construct noun phrase memory bank $\mathcal{M}_{\text{NP}}$ and scene graph memory bank $\mathcal{M}_{\text{SG}}$ by parsing training captions, selecting high-frequency elements, and enhancing scene graphs with noun phrases to include more attribute information. The entire caption memory bank $\mathcal{M}_{\text{EC}}$ contains all training captions. During training, following MultiCapCLIP (Yang et al., 2023), we retrieve top-K elements from memory banks using perturbed embedding $\tilde{\mathbf{e}}$ and train a text decoder to reconstruct the original text. During inference, we first classify $\mathcal{M}_{\text{NP}}$ with GPT-4 (OpenAI et al., 2024) and $\mathcal{M}_{\text{SG}}$ based on noun phrase categories, compute statistical priors, retrieve a diverse set of relevant noun phrase and scene graph elements using CLIP embeddings with top-p filtering and generate a weighted embedding from $\mathcal{M}_{\text{EC}}$ using softmax similarity scores between video and caption features. Three types of prompt are transformed by respective FFNs and concatenated to generate the final caption.

scene graph. Pseudocode for $\mathcal{M}_{\text{SG}}$ construction is provided in the Appendix B.

Entire Caption Memory Bank ($\mathcal{M}_{\text{EC}}$) This memory bank provides holistic textual descriptions that help maintain linguistic coherence in the generated results. It is simply the aforementioned $\mathcal{S}$.

These three memory banks—noun phrases, enhanced scene graphs, and entire sentences—serve as progressively richer textual representations of the visual content and are significant for guiding the caption generation process.

3.2 Training Procedure

Our training objective is to learn a text decoder that generates captions conditioned on multi-granularity prompts, while maintaining robustness to noise during cross-modal retrieval. The text decoder is constructed as a stack of Transformer (Vaswani et al., 2017) decoder blocks. The training process consists of the following steps.

Embedding Augmentation For each caption S_o, we first retrieve the most similar M captions from the training set, based on their cosine similarity to the CLIP sentence embedding $\mathbf{e}(S_o)$. One of the M captions is randomly selected as S_r. We use $\mathbf{e}(S_r)$ and add Gaussian noise $\epsilon \sim \mathcal{N}(0, \lambda^2)$ to obtain a perturbed embedding $\tilde{\mathbf{e}}$ following MultiCapCLIP (Yang et al., 2023):

$$\tilde{\mathbf{e}} = \mathbf{e}(S_r) + \epsilon. \tag{4}$$

Memory Bank Retrieval The perturbed embedding $\tilde{\mathbf{e}}$ is used to retrieve the top-K_p noun phrases and top-K_g scene graphs from $\mathcal{M}_{\text{NP}}$ and $\mathcal{M}_{\text{SG}}$, respectively, based on the cosine similarity of CLIP embeddings. The representations of these retrieved elements are denoted as $\mathbf{e}_{\text{NP}}$ and $\mathbf{e}_{\text{SG}}$.

Prompt Construction Passing $\mathbf{e}_{\text{NP}}$, $\mathbf{e}_{\text{SG}}$, and $\tilde{\mathbf{e}}$ through individual FFNs yields $\mathbf{e}'_{\text{NP}}$, $\mathbf{e}'_{\text{SG}}$, and $\mathbf{e}_{\text{EC}}$, which are concatenated as the final prefix prompt:

$$\mathbf{P} = \texttt{Concat}(\mathbf{e}'_{\text{NP}}, \mathbf{e}'_{\text{SG}}, \mathbf{e}_{\text{EC}}). \tag{5}$$

Finally, the model is optimized with the cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, \mathbf{P}; \theta), \tag{6}$$

where y_t is the target word of sentence S_o at time step t, and θ denotes the model parameters.

3.3 Inference with Category-Aware Retrieval

During inference, we utilize a specialized strategy for prompt generation, combining category-aware

Algorithm 1 Category-based Statistics Computation

Input: $\mathcal{V}$: Training videos; $\mathcal{P}_\kappa$: Noun phrases of category κ

Output: μ_κ, p_κ: Avg. noun phrases per video; category probability

$N_\kappa \leftarrow 0$, $N_\kappa^{\mathcal{P}} \leftarrow 0$.
for $V \in \mathcal{V}$ **do**
  Parse captions of V to extract noun phrases $\mathcal{P}_V$.
  $\mathcal{P}_V \leftarrow$ Remove duplicates from $\mathcal{P}_V$.
  if $\mathcal{P}_V \cap \mathcal{P}_\kappa \neq \emptyset$ **then**
    $N_\kappa \leftarrow N_\kappa + 1$.
    $N_\kappa^{\mathcal{P}} \leftarrow N_\kappa^{\mathcal{P}} + |\mathcal{P}_V \cap \mathcal{P}_\kappa|$.
  end if
end for
$p_\kappa \leftarrow \frac{N_\kappa}{|\mathcal{V}|}$, $\mu_\kappa \leftarrow \frac{N_\kappa^{\mathcal{P}}}{N_\kappa}$.
return p_κ, μ_κ

retrieval with top-p post-processing, to ensure diverse and relevant textual prompts for accurate and expressive video captions.

Noun Phrase Prompt Generation The inference begins with generating prompts from $\mathcal{M}_{\text{NP}}$, involving noun phrase classification, relevant candidates retrieval based on statistical priors, and refinement with a top-p mechanism. Specifically, the retrieval step employs statistical priors tailored to the in-domain and cross-domain settings, respectively, and is discussed separately below.

Classification with GPT-4: We leverage GPT-4 to autonomously determine categorization for $\mathcal{M}_{\text{NP}}$ without a predefined taxonomy. The model generates emergent categories (*e.g.,* singular people, object, place) and assigns all phrases accordingly. Distribution details are provided in the Appendix C.

In-Domain Retrieval with Statistical Priors: To adaptively determine hyperparameters (*e.g.,* the number of most relevant elements to select per-category) and obviate manual configuration during the categorized retrieval process, we initially compute in-domain statistical priors. For a video V from the training video set $\mathcal{V}$, unique noun phrases from its corresponding captions form a set $\mathcal{P}_V$. For each category κ with noun phrases $\mathcal{P}_\kappa \subseteq \mathcal{M}_{\text{NP}}$, we compute two statistics: a probability of occurrence, $p_\kappa = \frac{N_\kappa}{|\mathcal{V}|}$, where N_κ is the number of videos containing at least one noun phrase in $\mathcal{P}_\kappa$, and $|\mathcal{V}|$ is the total number of training set videos; and an average frequency, $\mu_\kappa = \frac{N_\kappa^{\mathcal{P}}}{N_\kappa}$, where $N_\kappa^{\mathcal{P}}$ is the total count of noun phrases parsed from the training corpus that overlap with $\mathcal{P}_\kappa$. This process is formalized in Algorithm 1.

Next, for a test video $V_{\text{test}} = \{f_t\}_{t=1}^{T}$, we compute frame-level cosine similarity between its CLIP visual features $\phi(f_t)$ and text embeddings $\mathbf{e}(n)$ of noun phrase $n \in \mathcal{P}_\kappa$. The video-phrase similarity s_n is derived by averaging frame-level scores. We retrieve the top-$\texttt{round}(\mu_\kappa)$ noun phrases from $\mathcal{P}_\kappa$ based on s_n, and retain all of them with probability p_κ. The retained noun phrases across all categories are aggregated into a set Y.

Cross-Domain Retrieval with Statistical Priors: In the cross-domain scenario, visual features of the target domain are leveraged to retrieve textual units from memory banks constructed from the source domain. These retrieved prompts are subsequently fed into the text decoder pre-trained on the source domain to generate captions. We compute category statistics using only the source domain training captions $\mathcal{S}$. For each category κ, we compute the total count of noun phrases parsed from the training corpus that belong to the category's noun phrase set $\mathcal{P}_\kappa$, denoted as N_κ. The category with the minimum count N_κ is designated as the base category, with its corresponding count b. Then, for each category κ, we retrieve $r_\kappa = \texttt{round}(N_\kappa / b \cdot B)$ noun phrases, where B is a pre-defined base retrieval number. We adopt the same cross-modal similarity computation method as in the in-domain setting and aggregate the top-r_κ most relevant elements from each category into a set Y.

Top-p Refinement: To balance relevance and diversity, we refine Y using a top-p strategy (Holtzman et al., 2020): we normalize the similarities between the video and all noun phrases in Y into a probability distribution $\{\hat{s}_n | n \in Y\}$ (each score divided by the sum), sort Y in descending order of $\hat{s}_n$, and select the smallest subset $Y' \subseteq Y$ whose cumulative probability $\sum_{n \in Y'} \hat{s}_n$ reaches a predefined threshold τ. We encode Y' using CLIP text encoder to obtain the noun phrase prompt $\mathbf{e}(Y')$ of V_{test}.

Scene Graph Prompt Generation For the scene graph, we follow a pipeline similar to the previous section: Elements in $\mathcal{M}_{\text{SG}}$ are classified by pairing the categories of their subject and object noun phrases (*e.g.,* "People_*pred*_Object" in Figure 2). Noun phrases not in $\mathcal{M}_{\text{NP}}$ are assigned to the category of their nearest neighbor in $\mathcal{M}_{\text{NP}}$ based on BGE embedding similarity. Then, we compute the statistical priors analogously to the noun phrase case and aggregate the retrieved items into a set Z.

Finally, top-p filtering is applied to Z, and the filtered result Z' is encoded with CLIP's text encoder to produce the scene graph prompt $\mathbf{e}(Z')$.

Entire Caption Prompt Generation For the entire caption prompt, we compute the similarity between each video and each global caption embedding in $\mathcal{M}_{\text{EC}}$ as in DeCap (Li et al., 2023a), then generate a single prompt token $\mathbf{e}_c$ via softmax-weighted summation.

Integrating Prompts and Generating Captions $\mathbf{e}(Y')$, $\mathbf{e}(Z')$, and $\mathbf{e}_c$ are processed by their respective FFNs and concatenated into a sequence, which is fed into our text decoder to generate the final caption for V_{test}.

4 Experiments

4.1 Experimental Setups

Datasets We evaluate our method on three datasets: MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), and VATEX (Wang et al., 2019b) under CC BY 4.0 licence. MSR-VTT includes 10000 videos, split into 6513 training, 497 validation, and 2990 testing videos. MSVD contains 1970 YouTube clips, divided into 1200 training, 100 validation, and 670 testing videos. VATEX comprises over 30000 videos, and we use 25006 clips for training, and 2893 and 5792 videos for validation and testing respectively following MultiCapCLIP (Yang et al., 2023). Experiments on VATEX focus exclusively on the English corpus.

Evaluation Metrics Following the common practice, we evaluate the caption quality with four metrics, including BLEU@4 (B@4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), ROUGE-L (R) (Lin and Och, 2004) and CIDEr (C) (Vedantam et al., 2015). Among them, the CIDEr is specifically designed to evaluate captioning systems and better captures human judgment of consensus better than the others. We also report Self-BLEU (Zhu et al., 2018), a measure of text diversity in our ablation study on retrieval strategies.

Implementation Details Table 1 lists key hyperparameters. We employ the frozen pre-trained CLIP (ViT/B-16) for feature extraction. The text decoder is a 6-layer Transformer (∼41M params) trained from scratch. Further details see Appendix A.

Datasets	N_p	N_g	N_c	K_p	K_g	M	λ^2
MSVD	1000	37711	48774	13	16	5	0.01
MSR-VTT	1000	100000	130260	14	19	5	0.01
VATEX	3000	400000	250060	10	13	5	0.01

Table 1: Hyperparameters of in-domain experiments. N_c represents the number of captions in training split. For MSVD, 37711 is the total number of enhanced scene graphs derived from all training captions.

4.2 In-domain Captioning

As shown in Table 2 and 3, our method establishes new state-of-the-art zero-shot performance across all benchmarks. Three key patterns emerge:

Vertical Dominance Our method significantly outperforms existing text-only training methods, achieving CIDEr scores of 39.3% and 92.9% on MSR-VTT and MSVD, respectively, surpassing MultiCapCLIP by 5.7% and 16.2%. We also observe superior performance on VATEX across all metrics, confirming the efficacy of our approach.

Horizontal Consistency The CIDEr disparity between MSVD (92.9%) and VATEX (41.4%) reflects the positive correlation between performance and dataset complexity. VATEX's longer videos and denser temporal relations present a greater challenge than MSVD's short clips.

Supervised Proximity On MSVD, our 92.9% CIDEr outperforms supervised methods like SAAT (Zheng et al., 2020), STR (Liu et al., 2023), and POS-CG (Wang et al., 2019a). While a gap remains on MSR-VTT and VATEX, our results highlight the potential of our approach in narrowing the gap with traditional supervised methods.

4.3 Cross-domain Captioning

We conduct cross-domain experiments on MSR-VTT and MSVD to evaluate the generalization ability (Table 4). Our method achieves 51.7% CIDEr (+11.8% over MultiCapCLIP, +23.5% over DeCap) and 28.9% B@4 (+4.1% over MultiCapCLIP) in MSR-VTT ⇒ MSVD task, and 28.1% CIDEr in MSVD ⇒ MSR-VTT task, consistently outperforming previous methods. These improvements underscore the superior generalization performance of our approach, attributed to the utilization of multi-granularity textual semantic units and category-aware retrieval with top-p filtering. This facilitates effective knowledge transfer from source to target domain without video-text pairs.

4.4 Alignment Quality Analysis

Beyond captioning metrics, we further evaluate the quality of video–text alignment using recall of video–text retrieval as well as CLIPScore (Hessel et al., 2021) and BERTScore (Zhang et al., 2020).

Settings	Method	Pre-trained Model	Training Data		MSR-VTT				MSVD			
			Video	Text	B@4	M	R	C	B@4	M	R	C
Supervised	SGN (2021)	ResNet-101 + C3D	✓	✓	40.8	28.3	60.8	49.5	52.8	35.5	72.9	94.3
	POS-CG (2019a)	InceptionResNetV2	✓	✓	42.0	28.2	61.6	48.7	52.5	34.1	71.3	88.7
	SAAT (2020)	InceptionResNetV2 + C3D	✓	✓	39.9	27.7	61.2	51.0	46.5	33.5	69.4	81.0
	STR (2023)	InceptionResNetV2 + I3D	✓	✓	-	25.8	54.8	47.6	-	34.2	68.6	86.5
	VPT (2022)	CLIP (ViT/B-16)	✓	✓	41.2	27.9	61.5	50.3	54.6	36.0	73.1	94.7
	CoCap (2023)	CLIP (ViT/B-16)	✓	✓	43.1	29.8	62.7	56.2	55.9	39.9	76.8	113.0
Zero-shot	ZeroCap (2022)	CLIP (ViT/B-32) + GPT-2_{Medium}	×	×	2.3	12.9	30.4	5.8	2.9	16.3	35.4	9.6
	ZeroCap-Video (2023)	CLIP (ViT/B-32) + GPT-2_{Medium}	×	×	3.0	14.6	27.7	11.3	3.0	17.8	31.4	17.4
	RETTA (2026)	CLIP (ViT/L-14) + GPT-2_{Medium}	×	×	14.0	19.3	42.2	24.3	23.3	28.5	56.4	49.8
	Knight (2023)	CLIP (ResNet50x64) + GPT-2_{Large}	×	✓	25.4	**28.0**	50.7	31.9	37.7	**36.1**	66.0	63.8
	IFCap (2024)	CLIP (ViT/B-32) + GPT-2_{Base}	×	✓	27.1	25.9	-	38.9	40.6	34.2	-	83.9
	ERFC (2025)	CLIP (ViT/B-32) + GPT-2_{Base}	×	✓	21.4	21.7	48.8	21.0	25.1	28.2	58.0	39.8
	DeCap‡ (2023a)	CLIP (ViT/B-16)	×	✓	26.6	23.5	53.2	29.7	35.2	29.0	65.2	41.3
	MultiCapCLIP† (2023)	CLIP (ViT/B-16)	×	✓	22.0	24.4	50.2	33.6	40.2	34.2	68.6	76.7
	Ours	CLIP (ViT/B-16)	×	✓	**31.4**	26.5	**55.1**	**39.3**	**45.7**	35.9	**71.5**	**92.9**

Table 2: In-domain captioning results on the MSR-VTT and MSVD test sets. ‡ indicates the reproduced results on both datasets using CLIP (ViT/B-16) for fair comparison. † denotes that the results on MSVD are from our implementation. The best scores are highlighted in **bold**, and the second highest scores are underlined.

Settings	Method	VATEX			
		B@4	M	R	C
Supervised	VATEX (2019b)	28.4	21.7	47.0	45.1
	HRNAT (2022)	32.5	22.3	49.0	50.7
	CoCap (2023)	31.4	23.2	49.4	52.7
Zero-shot	RETTA (2026)	11.4	16.3	32.6	23.8
	DeCap‡ (2023a)	19.2	19.3	42.8	27.5
	MultiCapCLIP† (2023)	21.7	20.1	43.3	38.0
	Ours	**23.8**	**21.0**	**44.5**	**41.4**

Table 3: In-domain performance on VATEX. ‡ indicates the reproduced results using CLIP (ViT/B-16). † marks the implementation with English annotations.

Method	MSR-VTT ⇒ MSVD	MSVD ⇒ MSR-VTT
	B@4 / M / R / C	B@4 / M / R / C
DeCap‡	23.1 / 25.4 / 56.9 / 28.2	16.4 / 18.6 / 50.3 / 8.8
MultiCapCLIP‡	24.8 / 28.9 / 57.7 / 39.9	20.4 / 22.2 / 50.9 / 22.4
Ours	**28.9 / 30.4 / 60.6 / 51.7**	**25.0 / 23.2 / 54.7 / 28.1**

Table 4: Performance on cross-domain captioning. ‡ denotes our reproduction with CLIP (ViT/B-16).

For video–text retrieval, we extract visual features from 8 uniformly sampled frames of each video using GME-Qwen2-VL-2B (Zhang et al., 2025), and compute the video–text similarity by averaging the cosine similarities between frame-level visual features and the text feature. As shown in Table 5 and Table 6, our method consistently outperforms prior zero-shot approaches, with particularly strong gains on VATEX (Wang et al., 2019b), a more carefully curated dataset with higher-quality annotations. These results suggest that our improvements arise from stronger video–text semantic alignment rather than task-specific decoding strategies.

4.5 Ablation Studies

Multi-granularity Textual Prompts We evaluate the effectiveness of our progressive multi-granularity prompts through in-domain ablation studies on MSR-VTT and MSVD, as shown in Table 7. Starting with "w/o Prompt", adding noun phrase prompts ("NP") boosts performance across all metrics—*e.g.,* B@4 increases from 25.9% to 28.7% on MSR-VTT and CIDEr increases from 60.8% to 79.7% on MSVD—highlighting their role in providing key entity information. Incorporating scene graphs enhanced with noun phrases ("NP+SG") yields further gains, notably a 9.5% CIDEr increase on MSVD, capturing relational and action details. The full model ("NP+SG+EC"), which integrates entire caption prompts, achieves peak performance, demonstrating that the progressive inclusion of multi-granularity prompts continually refines caption quality by capturing complementary semantic information at various abstraction levels.

Category-aware Retrieval with top-p We evaluate our retrieval strategy for cross-domain captioning with only noun phrase prompts in Table 8. Both "Direct Top-K" and "Category-aware" methods retrieve the same number of noun phrases: the former selects the most similar ones from the entire memory bank, while the latter draws from each category based on statistical priors. This category-aware strategy improves retrieval diversity, as indicated by a lower self-BLEU, though CIDEr slightly declines due to potential noise from irrelevant categories. Applying top-p filtering further enhances diversity and significantly boosts caption quality,

Dataset	Method	CLIPScore	RefCLIPScore	BERTScore (BERT-base)	BERTScore (RoBERTa-Large)
MSR-VTT	DeCap	0.7037	0.7789	0.6900	0.9106
	MultiCapCLIP	**0.7393**	0.7822	0.6963	0.9183
	Ours	0.7366	**0.7990**	**0.7167**	**0.9222**
MSVD	DeCap	0.6180	0.7099	0.7834	0.9399
	MultiCapCLIP	**0.7449**	0.8024	0.8023	0.9434
	Ours	0.7364	**0.8089**	**0.8238**	**0.9491**
VATEX	DeCap	0.7625	0.7912	0.6777	0.9012
	MultiCapCLIP	0.8155	0.8017	0.6924	0.9187
	Ours	**0.8307**	**0.8279**	**0.7014**	**0.9211**

Table 5: Additional unsupervised and model-based metrics on the in-domain test sets. CLIPScore and RefCLIPScore (Hessel et al., 2021) are computed with CLIP (ViT/B-16).

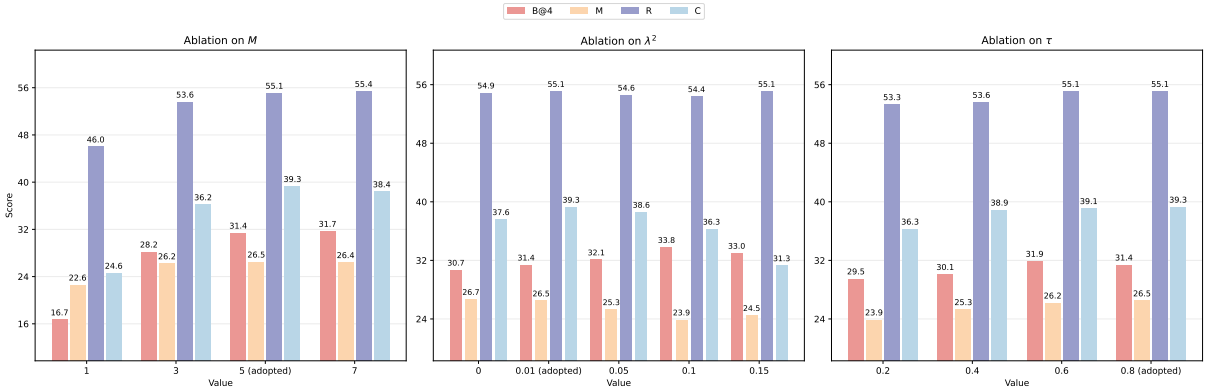

Figure 3: Ablation studies of the number of top-M retrieved captions, the Gaussian noise variance λ^2, and the top-p threshold τ on the MSR-VTT dataset.

Method	Text→Video			Video→Text		
	R@1	R@5	R@10	R@1	R@5	R@10
MSR-VTT						
DeCap	1.94	7.86	13.48	2.21	9.67	15.42
MultiCapCLIP	6.42	18.86	27.96	8.26	23.61	33.85
Ours	6.45	18.66	28.09	7.86	23.31	32.27
VATEX						
DeCap	3.88	13.81	21.44	5.84	17.82	27.69
MultiCapCLIP	11.62	30.23	42.25	14.45	36.84	49.84
Ours	14.52	35.51	47.86	17.08	40.75	54.09

Table 6: Video–text retrieval results on MSR-VTT and VATEX using GME-Qwen2VL-2B (Zhang et al., 2025).

Prompt Type	MSR-VTT	MSVD
	B@4 / M / R / C	B@4 / M / R / C
w/o prompt	25.9 / 25.8 / 52.6 / 29.8	33.4 / 31.4 / 64.9 / 60.8
NP	28.7 / 26.0 / 54.0 / 33.9	40.9 / 34.1 / 68.7 / 79.7
NP+SG	29.7 / 26.2 / 54.8 / 37.3	45.0 / 35.6 / 71.1 / 89.2
NP + SG + EC	**31.4 / 26.5 / 55.1 / 39.3**	**45.7 / 35.9 / 71.5 / 92.9**

Table 7: Ablation study on multi-granularity prompts. NP: Noun Phrase, SG: Scene Graph, EC: Entire Caption.

with CIDEr increasing from 31.0% to 41.4% on the MSR-VTT ⇒ MSVD task. These results highlight the effectiveness of combining category-aware retrieval with top-p filtering to balance diversity and relevance for higher-quality captions.

Beyond the core design choices, we further analyze the impact of several hyperparameters, including the number of retrieved captions, the Gaussian noise variance λ^2, and the top-p threshold τ, as illustrated in Figure 3. The smooth and consistent performance trends across different hyperparameter choices indicates that our method is robust rather than sensitive to specific parameter settings.

Scaling Up Pre-trained Multimodal Models Table 9 explores the impact of scaling pre-trained multimodal models on captioning ability using two relatively large datasets, MSR-VTT and VATEX. As model size increases, all metrics improve significantly. For example, on MSR-VTT, scaling from CLIP (ViT/B-16) to GME-Qwen2VL-7B (Zhang et al., 2025) boosts the CIDEr score from 39.3% to 48.2%, with an even greater improvement on VATEX (41.4% to 62.2%, +20.8%). These results demonstrate that larger multimodal models enhance our approach's ability to retrieve relevant

Retrieval Strategy	B@4	M	R	C	Self-BLEU (↓)
MSVD ⇒ MSR-VTT					
Direct Top-K	19.7	21.7	50.7	20.9	0.565
Category-aware	19.9	22.1	51.3	19.1	0.432
Category-aware w/ top-p	**20.9**	**22.8**	**51.9**	**23.2**	**0.419**
MSR-VTT ⇒ MSVD					
Direct Top-K	21.3	28.2	55.9	34.7	0.598
Category-aware	23.0	28.0	57.6	31.0	0.538
Category-aware w/ top-p	**25.2**	**29.1**	**58.7**	**41.4**	**0.475**

Table 8: Ablation study on retrieval strategies, with metrics for caption quality and prompt diversity.

VL Model	Size	VATEX	MSR-VTT
		B@4 / M / R / C	B@4 / M / R / C
CLIP (B-16)	0.15B	23.8 / 21.0 / 44.5 / 41.4	31.4 / 26.5 / 55.1 / 39.3
CLIP (L-14)	0.43B	24.2 / 21.4 / 44.4 / 46.9	31.5 / 26.4 / 55.7 / 40.1
G-Q-2B	2.2B	26.3 / 23.2 / 46.3 / 50.9	31.8 / 26.6 / 56.1 / 41.5
G-Q-7B	8.2B	32.7 / 24.5 / 49.9 / 62.2	34.8 / 28.5 / 57.7 / 48.2

Table 9: Impact of different pre-trained vision-language models on in-domain zero-shot video captioning performance. B-16: ViT/B-16, G-Q: GME-Qwen2VL.

Method	Encoding	Prompt Cons	Decoding	All	CIDEr (%)
DeCap		–/–/0.67	5.80	103.55	41.3
MultiCapCLIP	97.08	0.40/–/–	29.80	127.28	76.7
Ours		0.55 / 0.81 / 0.67	30.11	129.22	92.9

Table 10: Average per-video inference latency (ms). "Prompt Cons" represent NP / SG / EC construction time.

textual elements, leading to notable gains.

The advantage is particularly pronounced on the more complex VATEX dataset, underscoring the potential of our method with future, powerful multimodal models. See Appendix D.3 for qualitative comparisons of retrieved text units and generated captions across different pre-trained model scales.

4.6 Inference Efficiency

Table 10 shows per-video inference latency on test set of MSVD (batch size 1). Our prompt construction (category-aware retrieval with top-p post-processing) takes only 2.03ms (0.55 + 0.81 + 0.67) per video on average, negligible compared to video encoding (97.08ms) and text decoding (30.11ms). Although total latency is slightly higher than prior methods, this overhead brings substantial improvement in captioning quality.

4.7 Qualitative Analysis

Figure 4 presents a qualitative comparison between our method and other state-of-the-art approaches on three videos. In the first video, our model accurately identifies the key entity "birthday cake" from the noun phrase memory bank, and, by leveraging

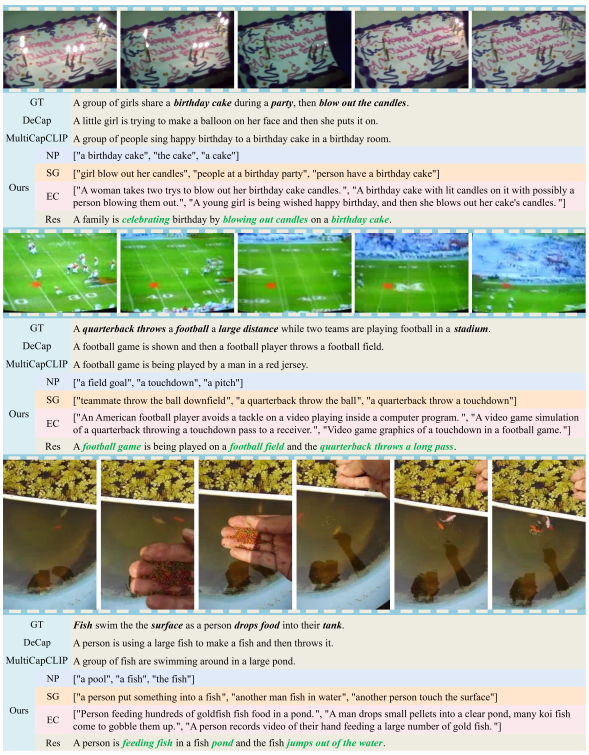

Figure 4: Qualitative comparison of our method and other state-of-the-art methods. Bold black: ground-truth keywords; bold green: accurate predictions.

both the scene graph and entire caption prompts, recognizes the action "blow out candles" which enables the model to generate a caption that covers all the essential information. For the remaining two videos, our three types of prompts continue to work synergistically. Our retrieval method allows noun phrases and scene graphs provide fundamental entities and action details, respectively, while entire captions capture the global context. As a result, our method produces more comprehensive captions.

5 Conclusion

We propose a zero-shot video captioning framework with two key innovations. First, a multi-granularity prompting strategy hierarchically integrates noun phrases (fine-grained entities), attribute-enriched scene graphs (structured object interactions), and entire captions (contextual coherence) to comprehensively represent visual semantics. Second, a category-aware retrieval with top-p filtering leverages training-data priors for adaptive and diverse prompt selection while preserving semantic relevance. Experiments on three datasets achieve favorable performance, and ablation studies confirm our design's effectiveness and potential.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62272438, in part by Beijing Natural Science Foundation L25700, and in part by the Fundamental Research Funds for Central Universities (E2ET1104).

Limitations

Our method outperforms training-free and text-only trained zero-shot video captioners across multiple video–text benchmarks. However, the construction of the noun phrase and enhanced scene graph memory banks uses frequency-based truncation, which can under-represent rare yet useful concepts and relations, limiting gains on videos with long-tail events or fine-grained attributes. As future work, we will construct the memory banks by jointly considering frequency and unit-level information gain. Concretely, a candidate noun phrase that exhibits high mutual information with phrases already in the bank may be admitted with lower probability, encouraging complementary rather than redundant entries. We will apply the same principle to the enhanced scene graph bank, which we expect to further improve overall performance.

Ethical Considerations

The datasets used in our study are publicly available video–text benchmarks consisting of natural videos and captions, which have been carefully pre-processed for academic research and therefore pose no obvious ethical concerns. However, when extending our approach to train language models on real-world text corpora and transferring them to broader video captioning scenarios, it is crucial to carefully examine the text data to avoid incorporating biased or discriminatory content into the model's parametric knowledge.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, pages 65–72.

Qianyue Bao, Fang Liu, Licheng Jiao, Yang Liu, Shuo Li, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. 2025. Erfc: Energy-aware reinforcement feedback calibration for zero-shot captioning. *IEEE Trans. Circuits Syst. Video Technol.*

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *ICCV*, pages 3113–3123.

Lianli Gao, Yu Lei, Pengpeng Zeng, Jingkuan Song, Meng Wang, and Heng Tao Shen. 2022. Hierarchical representation network with auxiliary tasks for video captioning and video question answering. *IEEE Trans. Image Process.*, 31:202–215.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with BART. In *EMNLP*, pages 8653–8666.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *ECCV*, pages 709–727.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. In *EMNLP*, pages 20715–20727.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023a. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *ICLR*.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023b. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *ACL (Findings)*, pages 6377–6390.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, pages 605–612.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, pages 17928–17937.

Zhu Liu, Teng Wang, Jinrui Zhang, Feng Zheng, Wenhao Jiang, and Ke Lu. 2023. Show, tell and rephrase: Diverse video captioning via two-stage progressive training. *IEEE Trans. Multim.*, 25:7894–7905.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Yunchuan Ma, Laiyun Qing, Guorong Li, Yuankai Qi, Amin Beheshti, Quan Z. Sheng, and Qingming Huang. 2026. RETTA: retrieval-enhanced test-time adaptation for zero-shot video captioning. *Pattern Recognit.*, 171:112170.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected CLIP. In *EMNLP (Findings)*, pages 4055–4063.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D. Yoo. 2021. Semantic grouping network for video captioning. In *AAAI*, pages 2514–2522.

Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. 2023. Accurate and fast compressed video captioning. In *ICCV*, pages 15558–15567.

Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. 2023. Zero-shot video captioning by evolving pseudo-tokens. In *BMVC*, pages 429–432.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, pages 17897–17907.

Mingkai Tian, Guorong Li, Yuankai Qi, Shuhui Wang, Quan Z. Sheng, and Qingming Huang. 2024. Rethink video retrieval representation for video captioning. *Pattern Recognit.*, 156:110744.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. 2019a. Controllable video captioning with POS sequence guidance based on gated fusion network. In *ICCV*, pages 2641–2650.

Junyang Wang, Ming Yan, Yi Zhang, and Jitao Sang. 2023. From association to generation: Text-only captioning by unsupervised cross-modal mapping. In *IJCAI*, pages 4326–4334.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4580–4590.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *SIGIR*, pages 641—-649.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296.

Jie Yan, Yuxiang Xie, Shiwei Zou, Yingmei Wei, and Xidao Luan. 2025. Entrocap: Zero-shot image captioning with entropy-based retrieval. *Neurocomputing*, 611:128666.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. Multicapclip: Auto-encoding prompts for zero-shot multilingual visual captioning. In *ACL*, pages 11908–11922.

Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. 2022. Hierarchical modular network for video captioning. In *CVPR*, pages 17918–17927.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. Meacap: Memory-augmented zero-shot image captioning. In *CVPR*, pages 14100–14110.

Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *CVPR*, pages 23465–23476.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *CVPR*, pages 9274–9285.

Qi Zheng, Chaoyue Wang, and Dacheng Tao. 2020. Syntax-aware action targeting for video captioning. In *CVPR*, pages 13093–13102.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *SIGIR*, pages 1097–1100.

A Implementation Details

This section provides additional details on our implementation. We adopt the same model architecture as MultiCapCLIP (Yang et al., 2023), where the text decoder consists of a 6-layer Transformer (Vaswani et al., 2017) with 8 attention heads and a hidden size of 512. The CLIP (ViT/B-16) (Radford et al., 2021) model encodes textual units, which are subsequently processed by a feed-forward network (FFN) with both input and output dimensions set to 512 before being fed into the text decoder. During training, we apply label smoothing with a value of 0.1, while for inference, we employ beam search with a beam size of 3 to generate text tokens. All experiments are conducted over 10 epochs using the AdamW (Loshchilov and Hutter, 2019) optimizer, incorporating a linear warm-up phase over the first 10% of the training steps.

For in-domain tasks, the MSR-VTT (Xu et al., 2016) and MSVD (Chen and Dolan, 2011) datasets utilize a peak learning rate of 1×10^{-4}, which remains fixed after the warm-up phase. In contrast, the VATEX (Wang et al., 2019b) dataset employs a peak learning rate of 5×10^{-4}, followed by a linear decay to 0 after the warm-up period. The threshold τ for top-p post-processing remains consistent across noun phrases and scene graphs. The value of τ is set to 0.6 for both MSVD and VATEX, and to 0.8 for MSR-VTT.

Algorithm 2 Enhanced Scene Graph Memory Bank Construction

Input: $\mathcal{S}$: Training captions; $\{\mathcal{P}(S)\}_{S\in\mathcal{S}}$: Noun phrase sets; N_g: Frequency threshold
Output: $\mathcal{M}_{\text{SG}}$: Enhanced SG memory bank

$\mathcal{X}_{\text{all}} \leftarrow \varnothing$
for caption $S \in \mathcal{S}$ **do**
  $\mathcal{G}_S \leftarrow \text{TextualSceneGraphParser}(S)$
  for $g_i = \langle sub_i, pred_i, obj_i \rangle \in \mathcal{G}_S$ **do**
    $\mathcal{A}_i \leftarrow \{p \mid p \in \mathcal{P}(S) \wedge sub_i \text{ is substring of } p\} \cup \{sub_i\}$
    $\mathcal{B}_i \leftarrow \{p \mid p \in \mathcal{P}(S) \wedge obj_i \text{ is substring of } p\} \cup \{obj_i\}$
    $\mathcal{X}_i \leftarrow \{\langle a, pred_i, b\rangle \mid a \in \mathcal{A}_i, b \in \mathcal{B}_i\}$
    $\mathbf{E}_S \leftarrow \text{BGE}(S)$
    $\mathbf{E}_{\mathcal{X}_i} \leftarrow \text{BGE}(\mathcal{X}_i)$
    $x_{\text{best}} \leftarrow \arg\max_{x_i^j \in \mathcal{X}_i} \cos(\mathbf{E}_S, \mathbf{E}_{\mathcal{X}_i}[x_i^j])$
    $\mathcal{X}_{\text{all}} \leftarrow \mathcal{X}_{\text{all}} \cup \{x_{\text{best}}\}$
  end for
end for
$\mathcal{F} \leftarrow \{(x, \text{count}(x \in \mathcal{X}_{\text{all}})) \mid x \in \mathcal{X}_{\text{all}}\}$
$\mathcal{M}_{\text{SG}} \leftarrow \text{Sort}(\mathcal{F}, \text{count descending})[0 : N_g]$
return $\mathcal{M}_{\text{SG}}$

In the cross-domain setting, for the MSR-VTT $\Rightarrow$ MSVD task, the parameters K_p and K_g are configured to 12 and 34, respectively. For the MSVD $\Rightarrow$ MSR-VTT task, these parameters are set to 14 and 25, respectively. The learning rate and scheduler configurations mirror those of the in-domain tasks, with τ fixed at 0.5 for both cross-domain tasks.

We conduct a thorough anonymization procedure by manually inspecting the data from widely used benchmark datasets. Whenever information identifying specific individuals is found, it is replaced with neutral expressions such as “he”, “she”, or “a person”. All experiments are implemented using PyTorch (Paszke et al., 2019) and conducted on a single NVIDIA RTX A6000 GPU. Training on the text data of MSR-VTT, MSVD, and VATEX datasets takes approximately 40, 15, and 77 minutes, respectively.

Category Name	Category Description	Example Noun Phrases
Video Overall Description	Video types and content descriptions	“a talk show", “sports highlights", “a music video"
Abstract Noun Phrases	Noun phrases representing abstract concepts	“different types", “the features", “beauty"
Plural People	Nouns representing plural people	“some kids", “two men", “a band"
Personal Pronouns	Pronouns referring to people	“he", “us", “they"
Object Noun Phrases	Noun phrases representing physical objects	“black t-shirt", “a frying pan", “a race car"
Place Noun Phrases	Noun phrases representing different environments	“a busy street", “a basketball court", “a restaurant"
Singular People	Noun phrases representing single people	“the man", “a police officer", “Captain America"
Quantifiers & Others	Quantifiers and other general terms	“one", “which", “another"

Table 11: GPT-4-generated Categories for Noun Phrases Memory Bank $\mathcal{M}_{\text{NP}}$ with Example Phrases.

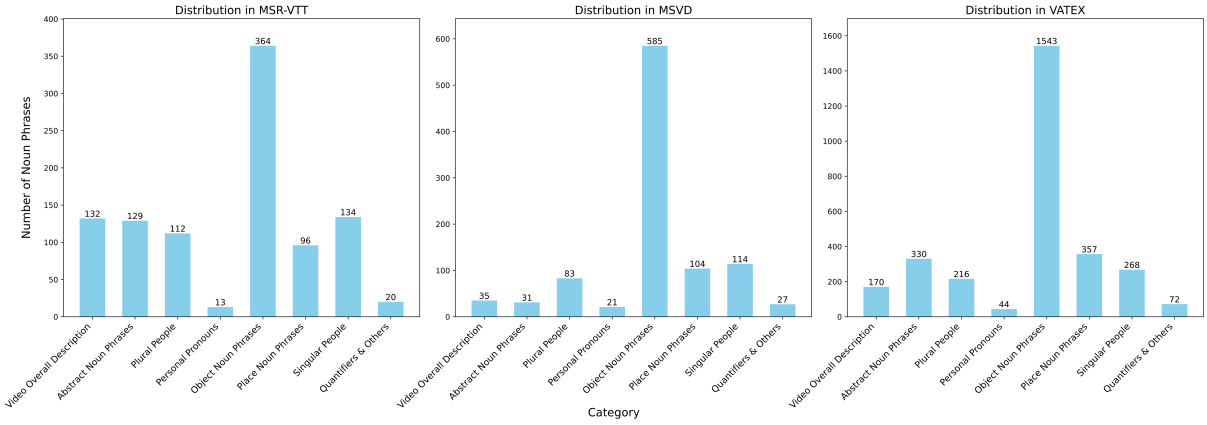

Figure 5: Distribution of Noun Phrases across Different Categories in the MSR-VTT, MSVD, and VATEX Datasets.

B Scene Graph Memory Bank Construction

Algorithm 2 details the process of constructing the enhanced scene graph memory bank. It takes the training captions as input and generates a memory bank consisting of scene graphs enriched with noun phrases.

C Classification of Noun Phrase Memory Bank

We first performed unsupervised classification using GPT-4 (OpenAI et al., 2024) on the noun phrase memory bank $\mathcal{M}_{\text{NP}}$ of MSR-VTT. Subsequently, the same categories are applied to the classification process for the MSVD and VATEX datasets. Table 11 presents the eight categories identified by GPT-4, together with their interpretations. Examples of noun phrases belonging to each category are provided in the last column. Figure 5 illustrates the distribution of noun phrases across these categories in the MSR-VTT, MSVD, and VATEX datasets. As can be observed, object noun phrases consistently dominate across all datasets, followed by singular people. For the complex VATEX dataset, which contains more diverse and intricate scenes, place noun phrases also exhibit a significant presence.

Size of Memory Bank		VATEX			
NP	SG	B@4	M	R	C
1000	200,000	23.1	20.6	44.1	37.4
1000	400,000	23.3	21.0	44.2	39.1
3000	400,000	23.8	21.0	44.5	41.4

Table 12: Impact of memory bank size on in-domain captioning, evaluated on VATEX test set. NP: Noun Phrase, SG: Scene Graph.

D Other Ablation Studies

D.1 Impact of Memory Bank Size

As illustrated in Table 12, we evaluate the performance of in-domain zero-shot video captioning with varying sizes of memory bank of noun phrase and scene graph containing noun phrases. Following DeCap (Li et al., 2023a), where the prompt at the entire caption granularity occupies only one token, computations during retrieval are considerably simplified. Therefore, we fix the size of the entire caption memory bank to the number of captions in the training set. Upon increasing the sizes of both the noun phrase and scene graph memory bank, we observe improvements across all metrics, with the enhancement in CIDEr (Vedantam et al., 2015) being the most notable. A larger memory

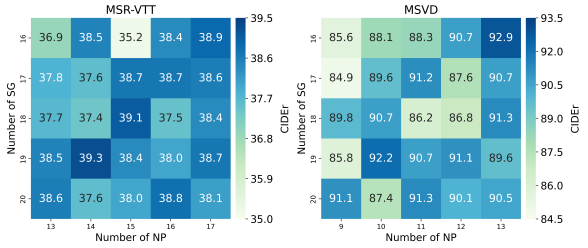

Figure 6: Impact of number of selected elements from noun phrase and scene graph memory banks on in-domain CIDEr scores. NP: Noun Phrase, SG: Scene Graph.

Figure 7: Comparison of the three granularities of text prompts retrieved using different pre-trained multimodal models and the generated captions, denoted as "Res". We emphasize ground-truth important words and accurate words of our generated descriptions in bolded black and green respectively.

bank suggests a more comprehensive and enriched knowledge base from the training set, thereby enhancing the generalization capability from training to inference phases.

D.2 Impact of top-*K* Selection from Memory

During training, we retrieve fixed numbers of elements from both the noun phrase (NP) and scene graph (SG) memory banks for each caption, which are then fed into the language decoder as prefix prompts for reconstruction. As visualized in Figure 6, our ablation study on MSR-VTT and MSVD systematically investigates how varying selection quantities of NP and SG affects the CIDEr metric. Notably, the model demonstrates well robustness across different parameter combinations, maintaining consistently high performance levels. Through grid search optimization, we ultimately identify top-14 NP with top-19 SG as the optimal configuration for MSR-VTT, while top-13 NP paired with top-16 SG achieves peak performance on MSVD.

D.3 Qualitative Gains from Scaling Multimodal Models

Building on the quantitative analysis of scaling up pre-trained multimodal models of the main paper, we present a qualitative analysis in Figure 7, visually illustrating the benefits. Compared to CLIP (ViT/B-16), larger models like GME-Qwen2VL-2B (Zhang et al., 2025) and GME-Qwen2VL-7B (Zhang et al., 2025) retrieve more video-relevant textual units from the same memory bank, leading to more semantically accurate and detail-rich captions.