LLM-based Description Enrichment for Short Video Clustering

Juliano Yugoshi[1], Ricardo Marcacini[1]

Institute of Mathematics and Computer Science, University of São Paulo, Brazil
{juliano.yugoshi, ricardo.marcacini}@usp.br

Abstract.

The rapid growth of short-video platforms has created large collections of visual content that are difficult to organize manually. Video clustering offers a practical way to explore these collections, but its quality depends on representations that capture the meaning of each video, not only its visual appearance. Compact vision-language models can generate useful descriptions, although these descriptions are often too generic to support fine-grained semantic grouping. We introduce a method for improving unsupervised video clustering by building richer semantic representations from short visual descriptions before embedding generation. Instead of relying directly on the initial captions produced by a compact vision-language model, the method expands each description with information about actions, objects, context, intent, scene type, and possible high-level categories. Multiple Large Language Models are used to generate complementary semantic views of the same video, which are then converted into embeddings and clustered without using labels. Experiments on the MSR-VTT benchmark show that enriched descriptions produce more coherent clusters than the original descriptions: the best overall configuration (Claude-enriched descriptions with OpenAI embeddings) reaches an NMI of 0.3740, a 5.71% relative gain over its caption-only baseline, while the largest relative improvement (13.18%) is obtained with Qwen embeddings. The results indicate that LLM-based enrichment helps reduce the semantic gap between visual content and high-level meaning, offering a practical path to organize unlabeled video collections under a known taxonomy.

CCS Concepts: • **Computing methodologies** → **Unsupervised learning**; *Cluster analysis*; Neural networks.

Keywords: unsupervised clustering, semantic enrichment, video clustering, vision-language models

1. INTRODUCTION

Short-video platforms such as TikTok, YouTube Shorts, and Instagram Reels have generated large, heterogeneous collections of visual content relevant to researchers, educators, and content analysts [Van Daele et al. 2024; Molem et al. 2024; Zannettou et al. 2024]. Organizing these collections presents significant challenges because newly collected videos often lack reliable instance-level annotations. In many practical applications, however, the high-level taxonomy is already defined by the platform or domain specialists. The challenge is therefore to organize unlabeled videos under a known semantic vocabulary without requiring manual video-level annotation.

Current video representation methods typically extract visual features from pre-trained models such as CLIP [Radford et al. 2021] or generate textual descriptions using vision-language models [Zhang et al. 2024; Liu et al. 2023]. Large multimodal models offer richer interpretations but limited scalability, while compact VLMs such as SmolVLM2 reduce this cost [Hugging Face 2024] at the expense of short, generic descriptions that emphasize visible elements only. This limitation reflects the semantic gap between low-level visual similarity and higher-level meaning [Smeulders et al. 2000]: a soccer goal and a basketball dunk are visually distinct yet share the broader category of sports. A hybrid strategy that shifts semantic reasoning to the textual domain, via LLM-based enrichment of compact VLM captions,

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

avoids repeated large multimodal inference on raw video, though it does not eliminate computational cost[1].

This work addresses the following question: *Can LLM-based semantic enrichment of compact visual descriptions improve unsupervised short-video clustering quality?* We propose a method that transforms compact visual descriptions into richer semantic representations before embedding generation. A compact VLM produces initial captions, which are then expanded using multiple heterogeneous LLMs to incorporate information about actions, objects, context, intent, and high-level categories. The enriched descriptions are embedded and clustered using K-Means without labels. Experiments on the MSR-VTT dataset show that enrichment improves clustering quality. The best overall configuration, combining Claude-enriched descriptions with OpenAI embeddings, achieves an NMI of 0.3740 ± 0.0023, a 5.71% relative gain over the corresponding caption-only baseline. With Qwen embeddings, Claude enrichment yields the largest relative improvement across all configurations, increasing NMI by 13.18% over its baseline. Paired Wilcoxon tests ($p < 10^{-3}$) indicate consistent directional differences across clustering initializations. Because between-run variance is small, standardized effect sizes can become unusually large; we therefore interpret Cohen's d together with the absolute NMI differences rather than as stand-alone evidence of practical relevance.

The contributions of this work are: (i) a taxonomy-guided semantic enrichment method for clustering unlabeled videos without instance-level supervision or fine-tuning; (ii) a modular formulation that separates visual captioning, taxonomy-guided textual enrichment, embedding projection, and clustering; (iii) a systematic empirical analysis across heterogeneous LLM and embedding configurations; and (iv) statistical and sensitivity analyses that quantify the robustness of the observed gains.

2. RELATED WORK

Unsupervised organization of video collections is commonly addressed by extracting representations and grouping instances according to similarity criteria [Smeulders et al. 2000]. In videos, this is difficult because each instance combines appearance, motion, objects, actions, scenes, and context. Multimodal models improved this setting. CLIP aligns images and text in a shared embedding space and supports strong zero shot transfer [Radford et al. 2021]. CLIP4Clip extends this idea to video retrieval by aggregating frame level information and linking it to textual representations [Luo et al. 2021].

Recent work has explored unsupervised and multimodal video clustering, including UMC for semantic discovery [Zhang et al. 2024], VTC for retrieval-oriented clustering [Liu et al. 2025], and DVC2 using intrinsic temporal structures [Wang et al. 2025], reinforcing that clustering quality depends on both algorithm and representation. Compact VLMs such as SmolVLM2 reduce computational cost [Hugging Face 2024] but produce short, visually-focused captions that miss intent, context, and broader categories. Prior work has enriched such captions textually, e.g., LaCLIP [Fan et al. 2023], CuPL [Pratt et al. 2023], and LLM2CLIP [Huang et al. 2024], or improved interpretability via generated descriptions [Menon and Vondrick 2023], while others adapt multimodal models through learned conditioning (CoOp, CoCoOp, Visual Prompt Tuning, SoftCPT, MVLPT) [Zhou et al. 2022b; 2022a; Jia et al. 2022; Ding et al. 2024; Shen et al. 2024], typically at the cost of task-specific supervision or fine-tuning.

Our work focuses on a taxonomy-guided clustering setting in which the category vocabulary is known, but video-level assignments are not. Unlike fully taxonomy-agnostic clustering, the method deliberately uses the known category vocabulary as semantic context during textual enrichment. Unlike zero-shot classification, however, the category predicted inside the prompt is not used as the final assignment: only the enriched textual description is embedded, and the final partition is produced by K-Means. This decomposition allows us to study how taxonomy-guided textual enrichment changes the geometry of the representation space without training or fine-tuning the multimodal encoder.

[1]A detailed cost breakdown is provided in the project's public GitHub repository.

Table I: Positioning of the proposed method in relation to prior work.

Approach	References	Main representation	Video	LLM enrich.	No tuning	Inst. superv.	Known taxon./semantic prior
Classical clustering (survey)	[Smeulders et al. 2000]	Generic data features	Partial	No	Yes	Mixed*	Mixed*
UMC	[Zhang et al. 2024]	Multimodal representations	Yes	No	Yes	No	No
DVC2	[Wang et al. 2025]	Temporal video representations	Yes	No	Yes	No	No
VTC	[Liu et al. 2025]	Text-video contrastive embeddings	Yes	No	No	Yes	No
Compact VLM captioning	[Hugging Face 2024]	Automatically generated captions	Yes	No	Yes	N/A	N/A
Textual enrichment methods	[Fan et al. 2023; Pratt et al. 2023; Menon and Vondrick 2023; Huang et al. 2024]	Rewritten or generated text	Partial	Yes	Partial	No	Partial
Learned conditioning methods	[Zhou et al. 2022b; 2022a; Jia et al. 2022; Ding et al. 2024; Shen et al. 2024]	Learned textual or visual conditioning	Partial	No	No	Yes	Yes
Our work	—	**Compact VLM captions enriched by LLMs**	**Yes**	**Yes**	**Yes**	**No**	**Yes**

*This entry surveys multiple heterogeneous CBIR approaches; individual methods within it vary between fully unsupervised and taxonomy-guided settings.

3. METHODOLOGY

3.1 Problem Definition

Let $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ be a collection of unlabeled short videos, and let $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ denote a known high-level category vocabulary, defined a priori by the platform or domain specialists, with $K = |\mathcal{Y}|$. The goal is to learn a clustering assignment $g : \mathcal{V} \to \{1, 2, \ldots, K\}$ without using class labels during the grouping process. Since the videos are unlabeled, the quality of the clustering depends mainly on the representation function $f : \mathcal{V} \to \mathbb{R}^d$ used before clustering. Given the representations $\{f(v_i)\}_{i=1}^n$, the method produces a partition $\mathcal{C} = \text{Cluster}_K(\{f(v_i)\}_{i=1}^n)$. Although $\mathcal{Y}$ is known and may be used to guide the semantic representation, the ground-truth mapping between individual videos and categories in $\mathcal{Y}$ is never exposed to the clustering process itself.

The central methodological question is how to define f so that the resulting vector space reflects semantic relations between videos rather than only low level visual similarity. In short video collections, two videos may be visually different but semantically close. Conversely, visually similar videos may belong to different semantic contexts. This motivates a representation strategy that introduces an intermediate textual semantic layer, guided by $\mathcal{Y}$, before embedding projection.

3.2 Semantic Representation through Textual Enrichment

The proposed method represents each video through a composition of visual description, textual enrichment, and embedding projection. Let ϕ denote a compact visual description function. For each video v_i, the initial caption is given by $c_i = \phi(v_i)$, a short textual description of the visible content acting as an efficient abstraction of the video.

This first caption also creates a semantic bottleneck: compact VLM descriptions tend to emphasize

visible objects, scenes, and actions, while omitting broader context, intent, scene function, and high level categories. To address this limitation, we introduce a semantic enrichment function T_m, where m denotes a language model used as an enrichment component. Given the initial caption c_i, a fixed instruction schema q, and the known category vocabulary $\mathcal{Y}$, the enriched description is $s_i^{(m)} = T_m(c_i, q, \mathcal{Y})$.

The role of T_m is not to assign final labels, but to expand the initial caption into a richer semantic description by making explicit elements such as actions, objects, context, intent, scene type, and possible high level categories drawn from $\mathcal{Y}$. When multiple language models are used, the method produces a set of enriched descriptions $\mathcal{S}_i = \{s_i^{(m_1)}, s_i^{(m_2)}, \ldots, s_i^{(m_r)}\}$ for each video, each treated as a model specific semantic view of the same visual content.

3.3 Embedding Projection and Clustering

After semantic enrichment, each enriched description is mapped to a vector space using a text embedding model. Let ψ_e denote an embedding function, where e identifies the embedding model, so that $x_i^{(m,e)} = \psi_e(s_i^{(m)})$. The complete representation induced by enrichment model m and embedding model e is defined in Equation (1).

$$f_{m,e}(v_i \mid \mathcal{Y}) = \psi_e(T_m(\phi(v_i), q, \mathcal{Y})). \tag{1}$$

Here, ϕ converts video into text, T_m increases semantic density using the known vocabulary $\mathcal{Y}$, and ψ_e maps the enriched text into a numerical space suitable for clustering. The baseline removes the enrichment stage, embedding the original caption directly as $f_{0,e}(v_i) = \psi_e(\phi(v_i))$, which allows us to isolate the effect of enrichment on the embedding space.

Before clustering, vectors are normalized dimension wise using $z = (x - \mu)/\sigma$, where μ and σ are estimated from the embedding matrix, and clustered with K-Means. For a given representation $f_{m,e}$, the clustering objective is shown in Equation (2).

$$\min_{\{\boldsymbol{\eta}_k\}_{k=1}^{K}} \sum_{i=1}^{n} \min_{k \in \{1,\ldots,K\}} \left\| z_i^{(m,e)} - \boldsymbol{\eta}_k \right\|_2^2. \tag{2}$$

Here, $\boldsymbol{\eta}_k$ is the centroid of cluster k and $z_i^{(m,e)}$ is the normalized embedding of video v_i. Although $\mathcal{Y}$ is available to the enrichment stage, the ground-truth video-to-category mapping is never exposed to captioning, enrichment, embedding, or clustering, and is used only afterward to compute NMI and ARI. In summary, the method treats compact VLM captions as semantic seeds expanded in the textual domain before embedding, keeping the approach practical for large collections while letting LLMs add higher-level semantic information prior to clustering.

4. EXPERIMENTAL EVALUATION

We investigate whether taxonomy-guided enrichment improves the clustering of unlabeled videos across two embedding models and multiple enrichment strategies, complemented by statistical and sensitivity analyses. The source code is available on GitHub[2].

4.1 Dataset

Our experiments use MSR-VTT [Xu et al. 2016], a benchmark comprising 10,000 short videos organized into a 20-category taxonomy. In our problem setting, the category names are treated as known a priori task information. However, the video-to-category ground-truth assignments remain hidden from the representation and clustering pipeline and are used exclusively for external evaluation.

[2] `https://github.com/julianoyugoshi/video-clustering-enrichment.git`

Initial descriptions are generated using SmolVLM2, a 2.7B parameter vision-language model, with frames sampled at 1 FPS. Descriptions are enriched using Gemini Flash 2.5, Claude 3 Haiku, and Llama 3.3 70B through the OpenRouter API [OpenRouter 2024], all configured with identical zero-shot instructions and structured JSON output. Only the enriched description field is used for embedding generation. Table II summarizes the experimental components.

Table II: Experimental components

Component	Model/Method	Details
Description generation	SmolVLM2	2.7B parameters, 1 FPS
Semantic enrichment	Gemini, Claude, Llama	Zero-shot, JSON output
Embedding (model 1)	Qwen Embedding 8B	4,096 dimensions
Embedding (model 2)	OpenAI Text Embedding 3 Large	3,072 dimensions
Preprocessing	StandardScaler	Applied before clustering
Clustering	K-Means	$K = 20$, k-means++

4.2 Evaluation Criteria

Clustering quality is assessed using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) [Strehl and Ghosh 2002; Hubert and Arabie 1985], which compare predicted clusters against reference categories without requiring label supervision during training. Mean scores and 95% confidence intervals are estimated via nonparametric bootstrap with $B = 1,000$ resamples [Efron and Tibshirani 1993]. Statistical significance is evaluated using paired Wilcoxon Signed Rank tests [Wilcoxon 1945], with effect size quantified by Cohen's d [Cohen 1988].

K-Means is executed with k-means++ initialization, $n_{\text{init}} = 10$, over 15 independent runs with controlled variation in clustering initialization. Because the taxonomy is assumed to be known a priori, its cardinality is also known and the main evaluation uses $K = |\mathcal{Y}| = 20$. This choice does not use the hidden video-to-category assignments. We additionally vary K from 2 to 30 to assess whether the relative behavior of the representations depends on the nominal taxonomy size.

4.3 Experimental Setup

The semantic enrichment stage uses the same instruction schema and the same known category vocabulary for all LLMs. Access to the taxonomy is intentional and follows the problem definition: in the target scenario, the semantic categories are known before processing, but the category of each incoming video is not. The prompt uses this vocabulary as contextual guidance for enrichment rather than as instance-level supervision. The prompt used is shown in Fig 1.

```
You are a semantic enrichment assistant for a video collection organized under a known
high-level taxonomy.
Analyze the visual description using the following category vocabulary as
semantic context: {MSRVTT_CATS}
1. Indicate the single known category that is most semantically compatible with the
description.
2. Propose a complementary free-form category or concept.
3. Provide a brief semantic justification in English.
4. Provide an Enriched Description in English, making actions, objects, context, intent,
scene type and high-level semantics explicit.
Return ONLY valid JSON using the existing schema.
```

Fig. 1: Prompt used for semantic enrichment.

Only the `enriched_description` field is used to construct the embedding. The `predicted_category`, `suggested_category`, and `justification` fields are not concatenated to the embedded text and are not used as cluster assignments or pseudo-labels. This design keeps the downstream clustering stage independent from the category selected by the LLM.

4.4 Baselines

The evaluation comprises two caption-only baselines and six enriched configurations obtained from three LLMs and two embedding models. All configurations use $K = |Y| = 20$.

The baselines embed the original SmolVLM2 captions, whereas the proposed variants embed descriptions expanded by an LLM conditioned on Y. Thus, the comparison measures the combined effect of taxonomy-guided textual enrichment relative to caption-only representations; it does not isolate taxonomy access from textual expansion.

Table III: Experimental hyperparameter configuration

Parameter	**Value**
Number of clusters (K)	20
K-Means initialization	`k-means++`
Initialization trials (n_{init})	10
Runs per configuration	15
Total configurations	8
Bootstrap resamples (B)	1,000
Confidence level	95%
Random seed	42

4.5 Results

Table IV compares direct clustering of SmolVLM2 captions with the variants of our method. The baseline corresponds to embedding the original captions without semantic enrichment. The proposed variants apply LLM based enrichment before embedding generation. This comparison directly evaluates whether the semantic layer introduced by our method improves the clustering space relative to a caption-only representation under the same taxonomy-guided setting (Section 4.4). Because both the baseline and the enriched variants are evaluated with the same known cardinality $K = |\mathcal{Y}|$, the reported gains reflect the effect of LLM-based textual enrichment specifically, rather than the mere availability of the number of target clusters.

Table IV: Clustering results: baseline and enriched variants

Representation	**Embedding**	**NMI ↑**	**ARI ↑**
Baseline	Qwen	0.2943 ± 0.0074	0.1757 ± 0.0129
Baseline	OpenAI	0.3538 ± 0.0106	0.2429 ± 0.0138
Gemini enriched	Qwen	0.3082 ± 0.0054	0.1853 ± 0.0100
Gemini enriched	OpenAI	0.3548 ± 0.0051	0.2356 ± 0.0124
Claude enriched	Qwen	0.3331 ± 0.0045	0.2183 ± 0.0119
Claude enriched	OpenAI	$\mathbf{0.3740 \pm 0.0023}$	$\mathbf{0.2564 \pm 0.0088}$
Llama enriched	Qwen	0.3246 ± 0.0040	0.2027 ± 0.0064
Llama enriched	OpenAI	0.3549 ± 0.0054	0.2389 ± 0.0142

With Qwen embeddings, all enrichment variants outperform the baseline. Claude enrichment achieves the maximum gain: NMI increases from 0.2943 to 0.3331 (13.18% improvement). Llama and Gemini also improve performance, with smaller margins. With OpenAI embeddings, improvements are selective. Claude enriched descriptions reach the best overall result: NMI of 0.3740 ± 0.0023 and ARI of 0.2564 ± 0.0088 (5.71% gain over baseline). Gemini and Llama produce marginal NMI gains with no ARI improvement, indicating that enrichment benefit depends on both the LLM used and the embedding model's representation space. Table V reports statistical validation of enriched variants against baselines.

The repeated p-value 6.1×10^{-5} is the minimum exact Wilcoxon value for n=15 paired runs with consistent-direction differences, reflecting directional consistency rather than varying evidence strength. The large Cohen's d for Qwen reflects low variance across K-Means initializations; we thus report absolute NMI differences alongside effect sizes. We also analyze whether the observed improvements depend on fixing the number of clusters to the MSR-VTT taxonomy. Figure 2 reports NMI for K ranging from 2 to 30.

Table V: Statistical validation: Wilcoxon tests and effect size

Embedding	Enrichment	p-value	Sig.	$\Delta\mu_{\text{NMI}}$	Cohen's d
Qwen	Gemini	6.1×10^{-5}	Yes	+0.0140	9.85
Qwen	Claude	6.1×10^{-5}	Yes	+0.0391	18.96
Qwen	Llama	6.1×10^{-5}	Yes	+0.0306	12.49
OpenAI	Gemini	0.1688	No	+0.0010	0.39
OpenAI	Claude	6.1×10^{-5}	Yes	+0.0209	3.51
OpenAI	Llama	0.1514	No	+0.0011	0.31

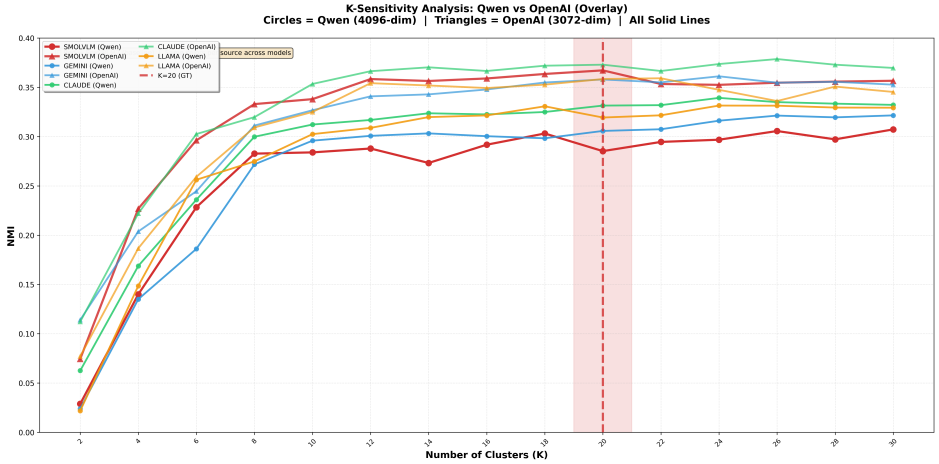

Fig. 2: NMI sensitivity to cluster count for $K \in [2, 30]$. Vertical line marks $K = 20$ (MSR-VTT taxonomy).

Because the taxonomy and its cardinality are known a priori in our setting, performance near $K = 20$ should not be interpreted as evidence of taxonomy discovery or as evidence against the use of taxonomy information. Instead, the sensitivity analysis evaluates robustness to the clustering cardinality. The relative advantage of enriched representations across a broader range of K values indicates that the observed behavior is not restricted to a single K-Means configuration.

5. CONCLUDING REMARKS

This work investigated LLM-based semantic enrichment as an intermediate representation step for taxonomy-guided video clustering without instance-level supervision. Instead of clustering short captions produced by a compact Vision-Language Model directly, the proposed method enriches them before embedding generation, making actions, context, intent, scene type, and high-level categories more explicit.

Experiments on MSR-VTT show that this strategy improves clustering quality over the SmolVLM2 baseline. The best overall result was obtained with Claude-enriched descriptions and OpenAI embeddings, reaching an NMI of 0.3740 ± 0.0023, a 5.71% relative gain over the caption-only baseline. With Qwen embeddings, Claude enrichment produced the largest relative improvement, increasing NMI by 13.18% over its own baseline. Four of the six enriched configurations significantly outperformed their corresponding baselines.

The proposed method requires no fine-tuning or annotation during clustering and offers a practical way to improve the organization of unlabeled video collections. Our evaluation assumes a known high-level taxonomy and therefore does not address fully open-world taxonomy discovery. This assumption is deliberate and reflects applications in which platforms or domain specialists define the categories of interest before unlabeled content arrives, so the method should be interpreted as taxonomy-guided clustering without instance-level supervision, not as fully taxonomy- gnostic clustering. Future work will investigate taxonomy-free clustering, automatic selection of K, larger datasets, taxonomy-conditioned

controls, and comparisons with purely visual and multimodal representations.

REFERENCES

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, USA, 1988.

Ding, K., Wang, Y., Liu, P., Yu, Q., Zhang, H., Xiang, S., and Pan, C. Prompt tuning with soft context sharing for vision-language models, 2024.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, London, UK, 1993.

Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. Improving clip training with language rewrites. In *Advances in Neural Information Processing Systems*. Vol. 36. New Orleans, USA, 2023.

Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., Dai, Q., Dai, X., Chen, D., Luo, C., and Qiu, L. Llm2clip: Powerful language model unlocks richer cross-modality representation, 2024.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification* 2 (1): 193–218, 1985.

Hugging Face. Smolvlm2: A small vision language model. https://huggingface.co/HuggingFaceTB/SmolVLM2, 2024.

Jia, M., Tang, L., Cheng, B., Xu, S., Ding, S., and Dai, J. Visual prompt tuning. In *European Conference on Computer Vision*. Tel Aviv, Israel, 2022.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023.

Liu, P., Wang, X., Cui, Z., and Ye, W. Queries are not alone: Clustering text embeddings for video search, 2025.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada, pp. 11451–11461, 2021.

Menon, S. and Vondrick, C. Visual classification via description from large language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. Kigali, Rwanda, 2023.

Molem, A., Makri, S., and Mckay, D. Keepin' it reel: Investigating how short videos on tiktok and instagram reels influence view change. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery, New York, NY, USA, pp. 317–327, 2024.

OpenRouter. Openrouter: Unified api for language models. https://openrouter.ai, 2024.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 15691–15701, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. Virtual, pp. 8748–8763, 2021.

Shen, Y., Mao, Z., and Xu, H. Multitask vision-language prompt tuning, 2024.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (12): 1349–1380, 2000.

Strehl, A. and Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* vol. 3, pp. 583–617, 2002.

Van Daele, T., Iyer, A., Zhang, Y., Derry, J. C., Huh, M., and Pavel, A. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2024.

Wang, Z., Mi, S., and Zhang, Y. Dvc2: Deep video cascade clustering from video structures. *Neurocomputing* vol. 657, 2025.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (6): 80–83, 1945.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 5288–5296, 2016.

Zannettou, S., Nemes-Nemeth, O., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., and Roesner, F. Analyzing user engagement with tiktok's short format video recommendations using data donations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2024.

Zhang, H. et al. Unsupervised multimodal clustering for semantics discovery in multimodal utterances, 2024.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (3): 1449–1469, 2024.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision* vol. 130, pp. 2337–2348, 2022b.