Language-Driven Graphs for Short Video Similarity

Juliano Koji Yugoshi[1,2] **, Ricardo Marcondes Marcacini**[1]

[1] Institute of Mathematical and Computer Sciences (ICMC)
University of São Paulo (USP) – São Carlos, SP – Brazil

[2]Campus de Três Lagoas (CPTL)
Federal University of Mato Grosso do Sul (UFMS) – Três Lagoas, MS – Brazil

juliano.yugoshi@usp.br, ricardo.marcacini@usp.br

Abstract. *Comparing videos by similarity is a central task in modern video analysis, but its effectiveness depends critically on the chosen representation. While visual embeddings effectively capture appearance, they often lack semantic abstraction. Recently, Large Language Models (LLMs) have emerged as a promising way to generate rich textual descriptions, yet the structural properties they induce in video similarity spaces remain underexplored. We introduce a graph-based methodology to investigate these properties, systematically comparing the structure of similarity graphs derived from visual features, human-written text, and LLM-generated text. Our framework evaluates how well each graph preserves semantic consistency, both in immediate neighborhoods (local cohesion) and across longer paths (global organization). Our analysis reveals a fundamental trade-off: visual graphs exhibit high local purity but decay rapidly, whereas LLM-based graphs preserve superior global semantic coherence. We demonstrate that LLMs build an abstract space that prioritizes deep thematic links—such as grouping videos by concepts like 'stage performance' across formal categories—over superficial purity. This structure offers a powerful, semantically organized alternative to the local cohesion of visual models.*

1. Introduction

The proliferation of short-form video platforms has led to an unprecedented volume of user-generated content, creating significant challenges for content retrieval, organization, and recommendation systems [Nguyen and Veer 2024, Covington et al. 2016, Davidson et al. 2010]. Measuring video similarity is fundamental to these systems. However, the very definition of "similarity" is multifaceted, ranging from low-level visual patterns to high-level thematic concepts. The choice of how to represent a video dictates the nature of the similarity space and, consequently, the performance of any downstream application.

Conventional approaches have predominantly relied on visual features [Wray et al. 2021]. These methods often involve extracting frame-level embeddings using pre-trained models, such as those trained on large-scale datasets like HowTo100M [Miech et al. 2019] or general-purpose vision-language models like CLIP [Radford et al. 2021]. These frame-level features are then typically aggregated into a single vector to represent the entire video. These representations are effective at capturing fine-grained visual details but are limited by the "semantic gap" [Smeulders et al. 2000], where visual proximity fails to capture abstract conceptual

relationships. For example, a video of a soccer goal and a video of a basketball dunk are visually distinct but belong to the same high-level "sports" category.

An emerging alternative is to use multimodal Large Language Models (LLMs) to generate descriptive text that summarizes the video's content, as noted by [Marafioti et al. 2025] and [Zohar et al. 2025]. By encoding these descriptions using sentence embedding models [Reimers and Gurevych 2019], it is possible to measure similarity based on conceptual meaning rather than visual appearance. However, the implications of choosing language over vision extend beyond retrieval accuracy or alignment with human judgment. Thus, we raise the following research question: *Do visual and language-based representations induce similarity spaces with fundamentally different topological properties?* In particular, how do these modalities differ in terms of local neighborhood cohesion and semantic organization of the video similarity space? This leads to our central hypothesis: that language-based representations sacrifice strict local purity, not as a flaw, but to build a more meaningful semantic space connected by narrative bridges—thematic links that transcend rigid categories. Our work aims to structurally characterize this trade-off.

To answer these questions, this paper introduces a graph-based framework to analyze and compare the structure of video similarity spaces. We construct and compare three approaches:

- A graph from visual embeddings (CLIP), representing the state-of-the-art visual approach.
- A graph from LLM-generated text embeddings, representing a scalable language-based approch.
- A graph from human-written text embeddings, serving as a gold-standard semantic topline approch.

In the experimental analysis, we demonstrate that graphs built from visual features exhibit high local neighborhood cohesion. In contrast, graphs derived from language show superior global semantic organization, maintaining categorical integrity over longer, more exploratory paths in the graph. This finding suggests that visual similarity is an inherently "local" phenomenon, while language semantic similarity operates more "globally" within the video semantic space. Furthermore, our results empirically validate that LLM-generated descriptions serve as a highly effective proxy for human annotation, producing a similarity space with a structural integrity that closely mirrors our human-annotated topline. In summary, the contributions of this paper are threefold:

- We propose a graph-based framework for the structural analysis and comparison of video similarity spaces derived from different modalities.
- We identify and empirically characterize a fundamental topological trade-off: visual embeddings create locally cohesive but globally disorganized graphs, while language-based embeddings build globally coherent semantic spaces at the cost of strict local purity.
- We empirically validate that LLMs are a viable and scalable method for creating semantically robust video representations, capable of replicating the structural quality of human-level annotation.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details our methodology for feature extraction, graph construction, and

evaluation. Section 4 presents our experimental results, followed by a discussion. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

This section reviews prior work in three key areas relevant to our study: visual-based methods for video similarity, the use of language to provide semantic context for videos, and the application of graph-based analysis to understand representation spaces.

2.1. Visual-based Video Similarity

Early video similarity methods relied on the extraction and comparison of visual features. Early approaches often relied on hand-crafted features like SIFT or SURF aggregated over frames [Sivic and Zisserman 2003]. The advent of deep learning introduced new techniques with Convolutional Neural Networks (CNNs) pre-trained on large image datasets like ImageNet [Deng et al. 2009], thereby becoming the standard for frame-level feature extraction. These frame embeddings were then typically aggregated through pooling strategies (e.g., mean or max pooling) or sequence models like LSTMs to produce a single video-level representation [Yue-Hei Ng et al. 2015].

More recently, contrastively trained multimodal models have set a new state of the art. CLIP [Radford et al. 2021], trained on a massive dataset of image-text pairs, produces highly robust visual embeddings that align well with natural language concepts. Its success has spurred the development of video-specific adaptations, which aim to better capture temporal dynamics [Xu et al. 2021]. Despite their power in matching visual patterns, these methods inherently focus on appearance, objects, and actions. As noted by Smeulders et al. [Smeulders et al. 2000], they often fall short of capturing the high-level narrative or abstract theme of a video, a limitation known as the "semantic gap". Our work uses CLIP as a strong visual baseline precisely to probe the structural nature of this gap.

2.2. Language as a Semantic Bridge for Video

Language offers a natural way to bridge the semantic gap by providing high-level and interpretable summaries of video content. The task of video captioning, which aims to automatically generate textual descriptions for videos, has been a central focus of multimodal research. Early models often followed an encoder-decoder architecture, using CNNs to encode visual features and RNNs to decode them into text [Venugopalan et al. 2015].

The advent of Large Language Models (LLMs) has shifted this paradigm. Modern multimodal LLMs, such as LLaVA [Liu et al. 2023] and MiniGPT-4 [Zhu et al. 2023], combine a pre-trained visual encoder with a powerful LLM. By fine-tuning on instruction-following datasets, these models can produce detailed, narrative descriptions of visual content. However, many foundational models carry a significant computational cost. A recent and highly relevant trend focuses on developing smaller, more efficient models that retain high performance. An example is the approach proposed in SmolVLM [Marafioti et al. 2025] and in Smol-VLM-2 [Zohar et al. 2025], which demonstrates how to effectively pair a strong vision encoder (like SigLIP) with a compact language model (such as Phi-2 or Gemma). This approach yields models that are practical for large-scale applications while remaining highly capable. In this paper, our choice of

Smol-VLM-2 is motivated by this balance of cost and quality, making it a representative state-of-the-art model for scalable, semantic video analysis. While these models provide the raw material for semantic representation, understanding the structure of the resulting embedding space requires dedicated analytical tools, which leads to graph-based methods.

2.3. Graph-based Representation Analysis

Once data is represented in a high-dimensional space, analyzing its high-dimensional structure presents significant challenges. Graph-based methods provide a powerful framework for this purpose. By representing data points as nodes and their proximity as edges, one can transform a set of vectors into a relational structure [Von Luxburg 2007]. K-nearest neighbor (k-NN) graphs are a common and effective way to construct such structures from vector embeddings.

Graphs have been widely used for tasks like semi-supervised learning [Zhu et al. 2003] and clustering. In the context of representation learning, graph structure analysis can reveal important properties of an embedding space. For instance, Goldberg and Levy [Goldberg and Levy 2014] provided a deep analysis of the structure of word embeddings. Our work adopts a similar philosophy, but applies it to the domain of video similarity. We use random walks (a fundamental process for studying the properties of graphs [Lovász 1993]) as a probe to measure the semantic coherence of the similarity spaces induced by different modalities. In this sense, we can track the category of nodes visited during these walks and quantify abstract properties like "local cohesion" and "semantic organization". In our scenario, Local cohesion refers to the extent to which the immediate neighborhood of a node (i.e., nearby videos in the graph) shares the same semantic property, such as category. On other hand, semantic organization captures whether longer paths across the graph preserve such properties, i.e., whether a sequence of videos visited during a random walk remains semantically consistent. While prior work has used graphs to analyze embedding spaces, the specific topological nature of LLM-derived video spaces remains under-explored. In particular, no work has systematically analyzed their tendency to form thematic clusters that cross categorical boundaries, a structural property we investigate here.

3. Methodology

Our methodology explores the structural properties of video similarity spaces induced by different feature modalities. We construct k-nearest neighbor (k-NN) graphs using visual representations, LLM-generated textual descriptions, and human-annotated text representations. We then analyze random walks on these graphs to quantify and compare their ability to preserve semantic coherence at both local and global scales. The entire process is divided into three main stages: (1) data and feature extraction, (2) graph construction, and (3) structural evaluation via random walks.

3.1. Data and Feature Extraction

3.1.1. Dataset

We conducted our experiments on the MSR-VTT (Microsoft Research Video to Text) data set [Xu et al. 2016], a large-scale benchmark for video understanding. For our analysis,

we used the training and validation split ("TrainVal"), which comprises a diverse collection of 7,010 unique short video clips. Each video in this dataset is assigned to one of 20 distinct categories, such as "music", "people" ,"sports," and various other categories as shown in Table 1. We use these categories as the ground truth for our analysis.

Table 1. Videos Category

Category	Quantity	Category	Quantity
Music	507	Animation	221
People	173	Vehicles/Autos	571
Gaming	332	Howto	403
Sports/Actions	784	Travel	188
News/Events/Politics	355	Science/Technology	246
Education	200	Animals/Pets	443
Tv Shows	222	Kids/Family	387
Movie/Comedy	718	Documentary	112
Food/Drink	458	Cooking	232
Beauty/Fashion	341	Advertisement	117
Total			7010

3.1.2. Feature Modalities

For each of the 7,010 videos, we generated or extracted feature embeddings from three distinct modalities. The specific models used and the resulting vector dimensions are detailed below and summarized in Table 2.

Table 2. Summary of Feature Modalities and Embedding Specifications.

Modality	Source/Generation Model	Embedding Model	Dimension
Visual	ViT-L/14 (CLIP)	(Direct Output)	768
Human-Text	Human Annotators	`distiluse-base-multilingual`	512
LLM-Text	Smol-VLM-2	`distiluse-base-multilingual`	512

- **Visual Representation (Visual-CLIP):** We used the **ViT-L/14 CLIP image encoder** [Radford et al. 2021] to generate visual features. For each video, features were extracted from 8 uniformly sampled frames and then aggregated via mean-pooling to produce a single **768-dimensional** vector representing the entire clip.
- **Human-Annotated Text (Human-Text):** We concatenated the 20 ground truth captions available for each video into a single paragraph. This text was then encoded using the `distiluse-base-multilingual-cased-v2` sentence embedding model [Reimers and Gurevych 2019], generating a dense **512-dimensional** vector.
- **LLM-Generated Text (LLM-Text):** We leveraged the Smol-VLM-2 model [Zohar et al. 2025] to generate a synthetic, narrative description for each video in our dataset. The model was prompted with a simple instruction: "Describe in detail what is happening in this video." The resulting description was subsequently encoded using the same 512-dimensional sentence embedding model as the human-annotated text.

To ensure full reproducibility, the complete source code for this study—including scripts for feature extraction, graph construction, and random walk analysis—is publicly available on GitHub[1].

3.2. Graph Construction

From each of the three feature sets, we constructed a separate similarity graph. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of $n = 7,010$ videos. For each modality, the feature vectors are represented by a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the dimensionality of the embedding space. Let $\mathbf{x}_i$ denote the i-th row of $\mathbf{X}$, corresponding to the feature vector for video v_i.

We define three such matrices: $\mathbf{X}_{\text{visual}}$, $\mathbf{X}_{\text{human}}$, and $\mathbf{X}_{\text{llm}}$. Based on these, we construct the three graphs:

$$G_{\text{visual}} = (V, E_{\text{visual}}) \tag{1}$$

$$G_{\text{human}} = (V, E_{\text{human}}) \tag{2}$$

$$G_{\text{llm}} = (V, E_{\text{llm}}) \tag{3}$$

where the edge set E for each graph is constructed using a k-nearest neighbor (k-NN) model. Specifically, for each video $v_i \in V$, an edge is created to its k most similar videos, where similarity is determined by the cosine similarity between their corresponding feature vectors (e.g., between $\mathbf{x}_i$ and $\mathbf{x}_j$ for videos v_i and v_j).

Instead of using an arbitrary k, we determined this parameter dynamically for each modality. We selected the smallest integer k that ensures the resulting graph is fully connected, guaranteeing that our random walk analysis can explore the entire space without being trapped in disconnected components. The resulting values of k for each modality are presented in Table 3. The resulting k-NN graph is made undirected by creating an edge (v_i, v_j) if v_i is in the k-neighborhood of v_j or vice-versa. All edges are unweighted.

Table 3. Dynamically Determined k for Each Graph Modality.

Graph Modality	Resulting *k* for Connectivity
G_{visual}	14
G_{human}	4
G_{llm}	4

3.3. Structural Evaluation via Random Walks

To explore the structural properties of the graphs, we adopt a random walk-based evaluation [Lovász 1993]. A random walk is a stochastic process that traverses the graph by moving from a vertex to one of its neighbors, chosen uniformly at random. This process allows us to simulate an exploratory journey through the similarity space.

We measure the global semantic coherence of the random walking using a metric we call **Walk Purity**, adapted from the classical purity metric described in [Manning et al. 2008], applied here to the set of nodes visited during a random walk. For a given random walk of length l, represented by the sequence of nodes $W =$

[1] https://github.com/juliano-yugoshi/languagedrivengraphs.git

$(v_0, v_1, \ldots, v_{l-1})$, its purity is defined as the fraction of nodes in the walk that belong to the single most frequent category within that same walk. Formally, for a single walk W, the purity is:

$$\text{Purity}(W) = \frac{1}{l} \max_{c \in \mathcal{C}} \sum_{i=0}^{l-1} \mathbb{I}(C(v_i) = c) \tag{4}$$

where $\mathcal{C}$ is the set of all 20 ground-truth categories, $C(v_i)$ is the category of node v_i, and $\mathbb{I}(\cdot)$ is the indicator function.

The final purity for a given walk length l is the average Purity(W) over a large number of simulations ($N = 1000$ in our experiments). For each simulation, a starting node v_0 is chosen uniformly at random from the entire set of vertices V. It allows us to differentiate between local cohesion (high purity for short walks, e.g., $l = 2$) and global semantic organization (slower decay of purity as l increases). We compare our results against a Random Chance baseline, where the purity would be approximately $1/|\mathcal{C}| = 1/20 = 0.05$.

4. Results

A random walk analysis (Figure 1) reveals the distinct topological signature of each similarity graph. The Human-Text graph (G_{human}) defines the benchmark for semantic coherence. The Visual (CLIP) graph (G_{visual}), despite its strong local purity at $l = 2$, shows a sharp decay, confirming its reliance on non-semantic visual cues. The LLM-Text graph (G_{llm}), while exhibiting an intermediate purity score, consistently outperforms the random baseline, indicating a non-random, semantically coherent structure.

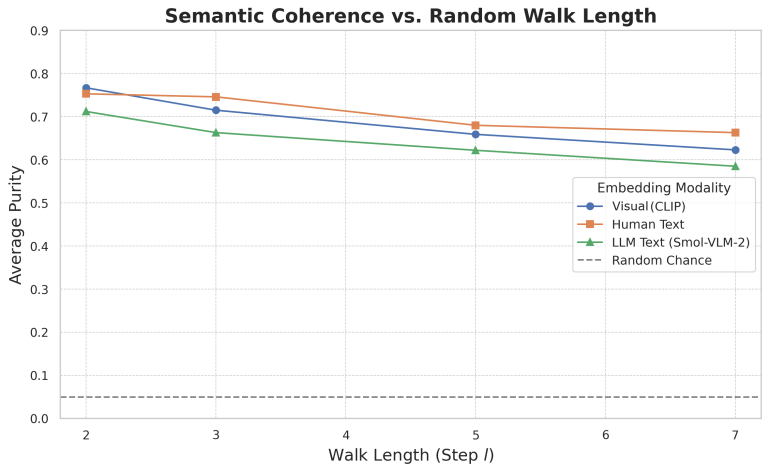
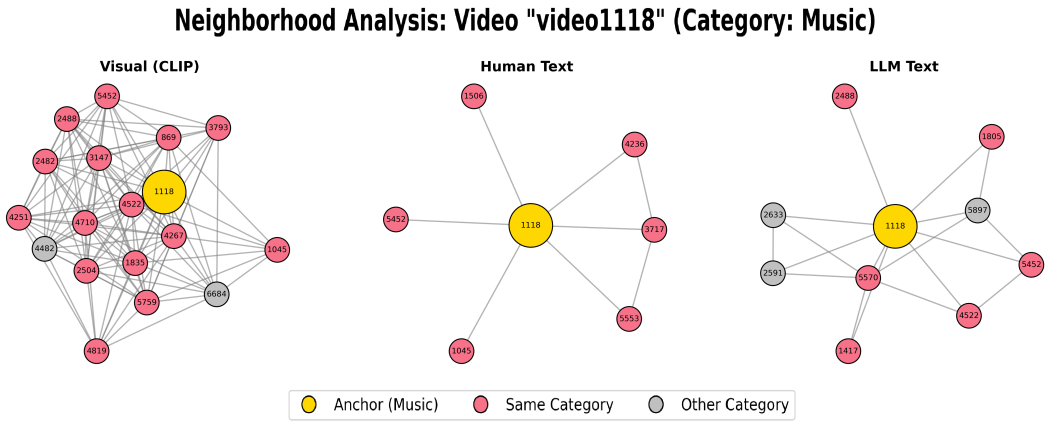


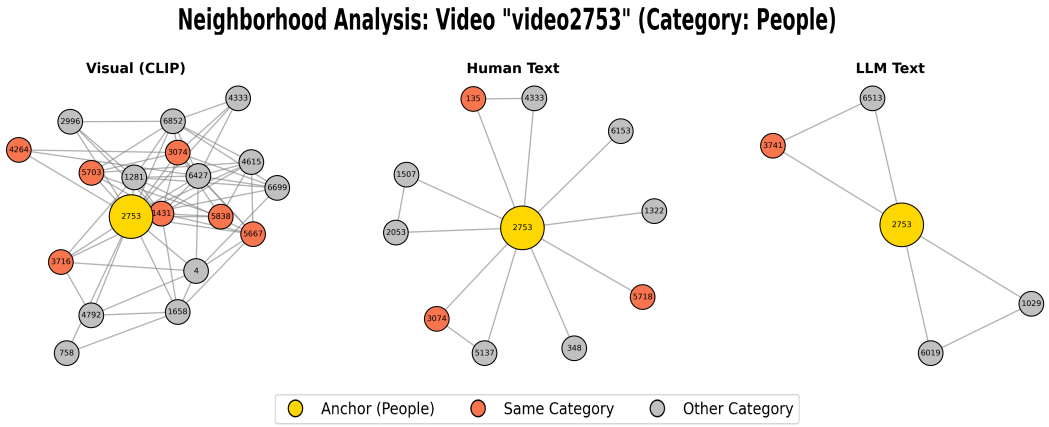
Figure 1. Categorical purity as a function of random walk length (l). The curves illustrate the distinct topological signatures, with the dashed line (0.05) representing random chance.

The LLM-Text graph's intermediate purity score initially seems paradoxical. Although derived from rich descriptive text, its space appears less categorically pure than the visual one at the most local level ($l = 2$). We hypothesize this is not a deficit, but rather evidence of a **functional trade-off**: the LLM-based representation prioritizes deep thematic connections over strict, and sometimes ambiguous, categorical labels.

A qualitative inspection of local neighborhoods (Figure 2) provides powerful evidence for this hypothesis. The Visual (CLIP) graphs (left column) consistently exhibit a dense, convoluted structure—a ”hairball effect”. In contrast, both Human-Text (center) and LLM-Text (right) graphs are significantly sparser. The critical distinction lies in *how* the LLM graph achieves this sparsity. While the Human-Text graph serves as a benchmark for categorical purity, the LLM graph selectively builds what we term **narrative bridges**, effectively refining or even correcting the dataset's ground-truth labels.



(a) Neighborhood of a ”Music” video (anchor: video1118).



(b) Neighborhood of a ”People” video (anchor: video2753).

Figure 2. Qualitative comparison of 1-hop neighborhoods. The visualizations reveal the core topological trade-off: Visual (CLIP) optimizes for local visual patterns, Human-Text for categorical purity, and LLM-Text for thematic, narrative connections.

The LLM-generated descriptions themselves reveal the logic behind these connections. For the ”Music” anchor in Figure 2a, which describes a ”live performance on stage”, its neighbors from other categories are not random. They are videos described as a ”live music performance” (node `5897`, cat: Commercials) and a ”live concert performance” (node `2633`, cat: People). The LLM ignores the formal labels and creates a highly coherent thematic cluster based on the core concept of a **stage performance**. The effect is even more nuanced for the ”People” anchor in Figure 2b, whose description contains the line ”YOU CAN'T HIDE YOUR FEELINGS.” Its neighbors are all videos

depicting "individuals seated at a table...engaged in a conversation" (e.g., node `1029`, cat: Travel; node `6019`, cat: Fashion). Here, the LLM links scenes based on a shared context of social interaction and intimacy, a theme primed by the anchor's explicit mention of emotion. This purposeful, cross-categorical linking explains the LLM's unique performance: it sacrifices superficial purity for a deeper, more associative semantic organization, yielding representations that are both robust and inherently interpretable.

Finally, the LLM-generated descriptions themselves reveal the narrative reasoning producing these connections. The following examples demonstrate how the LLM forges thematic links across different formal categories.

Thematic Link: "Stage Performance" (ref. Figure 2a)

`Anchor 1118` (Music): *"The video depicts a live performance on stage, featuring a male singer...holding a microphone. He is accompanied by a band..."*

`Neighbor 5897` (Commercials): *"The video captures a live music performance featuring a band on stage. The lead singer...holding a microphone, is the focal point."*

Thematic Link: "Social Interaction" (ref. Figure 2b)

`Anchor 2753` (People): *"...a group of three individuals seated in a dimly lit room...The scene transitions to a black screen with white text that reads, 'YOU CAN'T HIDE YOUR FEELINGS.'"*

`Neighbor 1029` (Travel): *"The video features a series of interactions between individuals seated at a round table...engaged in a conversation."*

5. Conclusions

In this work, we demonstrated that Large Language Models construct video similarity graphs not on principles of categorical purity, but through the creation of **narrative bridges**. Our analysis reveals that LLMs build a topology that prioritizes shared thematic concepts—such as grouping visually distinct videos under the common theme of a "stage performance"—over the dataset's formal labels. This structure, previously interpreted as a lack of precision, is instead a functional trade-off.

We establish that this abstract semantic space is particularly valuable for applications where content discovery and serendipity are critical, such as in recommendation engines that aim to broaden user taste beyond their immediate interaction patterns. Our analysis, therefore, offers a clear heuristic for practitioners: deploy the local precision of visual models for tasks requiring fine-grained visual matching, but leverage the associative power of LLMs for semantic exploration and recommendation.

Future work should explore hybrid methods that fuse visual and linguistic features to potentially achieve both local and global coherence. Applying this graph-based analysis to other modalities, including audio and event transcripts, also offers a path toward modeling multimodal semantic similarity with even greater fidelity.

References

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 191–198. ACM.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Garg, H., Gupta, S., Hsieh, Y., and Tock, T. (2010). The YouTube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pages 293–296. ACM.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Lovász, L. (1993). Random walks on graphs: A survey. In Miklós, D., Sós, V. T., and Szőnyi, T., editors, *Combinatorics, Paul Erdős is Eighty*, volume 2, pages 1–46. János Bolyai Mathematical Society.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). Boolean retrieval. *Introduction to information retrieval*, pages 1–18.

Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L. B., Lozhkov, A., Tazi, N., et al. (2025). Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). HowTo100M: Learning a text-video embedding by watching missing moments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640. IEEE/CVF.

Nguyen, T. T. and Veer, E. (2024). Why people watch user-generated videos? a systematic review and meta-analysis. *International Journal of Human-Computer Studies*, 181:103–144.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477. IEEE.

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

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. (2015). Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542. IEEE.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wray, M., Doughty, H., and Damen, D. (2021). On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3650–3660.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. (2021). Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). 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)*, pages 5288–5296. IEEE.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702. IEEE.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919. AAAI Press.

Zohar, O., Farré, M., Marafioti, A., Noyan, M., Cuenca, P., Zakka, C., and Joshua (2025). SmolVLM-2: Bringing video understanding to every device. Hugging Face Blog. `https://huggingface.co/blog/smolvlm2`. Accessed: 2025-06-24.