LUMOS-DM: Landscape-Based Multimodal Scene Retrieval Enhanced by Diffusion Model

Viet-Tham Huynh[1,2], Trong-Thuan Nguyen[1,2], Quang-Thuc Nguyen[1,2], Mai-Khiem Tran[1,2], Tam V. Nguyen[3], and Minh-Triet Tran[1,2(✉)]

[1] Software Engineering Laboratory and Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
{hvtham,ntthuan,nqthuc,tmkhiem}@selab.hcmus.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam
tmtriet@hcmus.edu.vn
[3] Department of Computer Science, University of Dayton, Dayton, U.S.A.
tamnguyen@udayton.edu

Abstract. Information retrieval is vital in our daily lives, with applications ranging from job searches to academic research. In today's data-driven world, efficient and accurate retrieval systems are crucial. Our research focuses on video data, using a system called LUMOS-DM: Landscape-based Multimodal Scene Retrieval Enhanced by Diffusion Model. This system leverages Vision Transformer and Diffusion Models, taking user-generated sketch images and text queries as input to generate images for video retrieval. Initial testing on a dataset of 100 h of global landscape videos achieved an 18.78% at Top-20 accuracy rate and 36.45% at Top-100 accuracy rate. Additionally, video retrieval has various applications, including generating data for advertising and marketing. We use a multi-modal approach, combining sketch and text descriptions to enhance video content retrieval, catering to a wide range of user needs.

Keywords: Information Retrieval · Stable Diffusion · ControlNet · Vision Transformer · Sketch · Text

1 Introduction

Information retrieval [6,8,12,17] is far from being a novelty in today's world. It plays a pivotal role in our daily lives, with numerous practical applications such as job hunting, academic research, and accessing work-related documents. In an era where the daily influx of information is rapidly expanding, one of the paramount concerns in information retrieval is ensuring the continuous and timely updating of data. In addition, the Diffusion Model (DM) [7], notably the Stable Diffusion [13] variant, is a cutting-edge generative modeling framework in machine learning. It excels at progressively enhancing random noise to generate high-quality and diverse data samples. Fine-tuning diffusion steps, provides precise control over the trade-off between sample quality and diversity, making it

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14557, pp. 145–158, 2024.
https://doi.org/10.1007/978-3-031-53302-0_11

a crucial tool for applications like image synthesis and text generation, driving innovation in AI and data science.

Video retrieval offers a multitude of possibilities, including the synthesis of new data for various purposes, such as advertising campaigns and marketing efforts. It is employed to retrieve appropriate content that can be used to create new media with the potential to go viral. Therefore, retrieval is essential. In this paper, our focus lies on the cross-modal of video retrieval, particularly for natural and outdoor scenes. We adopt a multi-modal approach, utilizing both sketch and text descriptions to depict queries for various scenes. Users have the ability to quickly sketch or draw the main features of a scene, such as a waterfall or a house with a windmill, and provide textual descriptions. By combining these two different sources, we synthesize a scene that closely mimics reality, which serves as input for our retrieval process. We employ this approach to enhance the retrieval of video content.

In an era of unprecedented data expansion, creating a swift and efficient information retrieval system is vital. Our research focuses on extracting concise and contextually relevant video clips from the dynamic landscape of video data. We introduce a straightforward, yet remarkably effective solution in this paper, emphasizing video content retrieval. Leveraging the power of the diffusion model, specifically the Stable Diffusion variant, empowers us to generate diverse, high-quality video clips. By integrating this generative model, our information retrieval system identifies relevant video segments and enhances search results with visually captivating and contextually pertinent content. This synergy of generative modeling and information retrieval pioneers a new approach to navigating the data deluge, ensuring that the extracted video clips align seamlessly with user expectations and preferences. Our proposed solution leverages the Vision Transformer (ViT) and Stable Diffusion Model to enhance information retrieval. Using user-generated sketch images and queries as input, our system generates images for video retrieval. The outcome is a video frame closely resembling the sketch image, accompanied by a relevant description, revolutionizing content retrieval.

To evaluate our system, we also carefully collected a dataset of 100 h of YouTube videos featuring various landscapes from different regions around the world. We conducted extensive testing on this dataset, and the results demonstrated that our proposed baseline method achieved a high accuracy rate (above 18.78%) in information retrieval and search.

The key results achieved in this paper are as follows:

- We successfully collected 100 h of landscape videos from various countries around the world. In this paper, we provide a list of YouTube links to the 100 h of collected videos.
- We propose a LUMOS-DM: Landscape-based Multimodal Scene Retrieval Enhanced by Diffusion Model method with input consisting of both sketches and text queries. This method has demonstrated its effectiveness with an 18.78% accuracy (Top-20) on the dataset we collected. Furthermore, we also tested on Top-50 and Top-100, respectively, resulting in 27.8% and 36.45%, respectively.

The structure of this paper is as follows. In Sect. 2, we briefly review recent work on information retrieval and the types of input data used in information retrieval. We present our process to collect data in Sect. 3. Then we propose in Sect. 4 our method using Vision Transformer, namely LUMOS-DM: Landscape-based Multimodal Scene Retrieval Enhanced by Diffusion Model, to conduct information retrieval. The experimental results in the landscape dataset using our proposed method are in Sect. 5. Finally, Sect. 6 discusses the conclusion and future work.

2 Related Work

2.1 Benchmarks

Information Retrieval Using Image. A novel CBIR method based on the transfer learning-based visual geometry group (VGG-19) approach, genetic algorithm (GA), and extreme learning machine (ELM) classifier was proposed by Bibi et al. [2] in 2022. The problem of semantic-based image retrieval of natural scenes was addressed by Alqasrawi [1]. It is acknowledged that the convolutional part of the network contains many neurons, with the majority having minimal impact on the final classification decision. In this context, a novel algorithm was proposed by Staszewski et al. [16], enabling the extraction of the most significant neuron activations and their utilization in constructing effective descriptors. In 2023, Chang et al. [3] introduced a Deep Supervision and Feature Retrieval network (Dsfer-Net) designed for bitemporal change detection. MindEye, an innovative fMRI-to-image approach for retrieving and reconstructing viewed images from brain activity, was presented by Scotti et al. [14]. Gong et al. [6] introduced Boon, a novel cross-modal search engine that combines the GPT-3.5-turbo large language model with the VSE network VITR (Vision Transformers with Relation-focused learning) to enhance the engine's capability in extracting and reasoning with regional relationships in images. This approach also provides textual descriptions alongside corresponding images when using image queries. Other influential contributions in this field include the work by Pradhan et al. [12].

Information Retrieval Using Text. A method for retrieving patents relevant to an initial set of patents, utilizing state-of-the-art techniques in natural language processing and knowledge graph embedding, was proposed by Siddharth et al. [15]. These methods struggle with mitigating ambiguity in video-text correspondence when describing a video using only one feature, necessitating multiple simultaneous matches with various text features. Lin et al. [10] introduced a Text-Adaptive Multiple Visual Prototype Matching Model to tackle this challenge in 2022. In 2023, Lin et al. [11] explored a principled model design space along two axes: representing videos and fusing video and text information. Chen et al. [5] utilized tags as anchors to enhance video-text alignment. However, the primary objective of the text-to-video retrieval task is to capture complementary audio and video information relevant to the text query, rather than solely

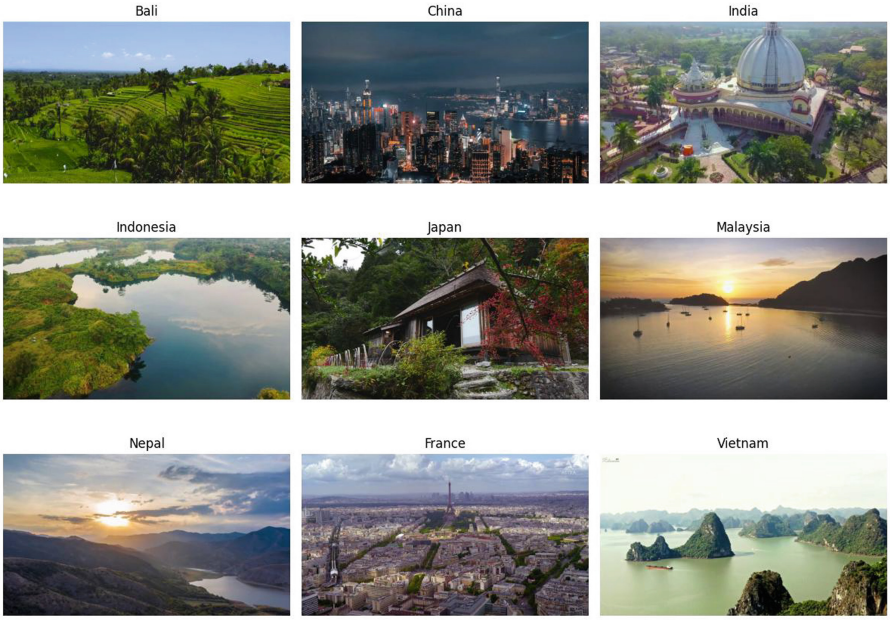

Fig. 1. Some illustrations of data collected from different countries

improving audio and video alignment. To address this concern, Ibrahimi et al. [8]. Introduced TEFAL in 2023, a method that produces both audio and video representations conditioned on the text query. To facilitate effective retrieval with aspect strings, Sun et al. [17] introduced mutual prediction objectives between the item aspect and content text.

2.2 Discussion

To capture more ranking information, Chen et al. [4] proposed a novel ranking-aware uncertainty approach for modeling many-to-many correspondences using provided triplets. In light of the existing research progress in text-pedestrian image retrieval, Li et al. [9] proposed a progressive feature mining and external knowledge-assisted feature purification method to address the associated challenges comprehensively. These two methods have achieved some quite outstanding results. However, Chen et al.'s method focuses on human subjects, while Li's method focuses on fashion. With this research, we want to focus on the subject of natural landscapes. The highlight of LUMOS-DM is that we will try to combine text and sketch input data instead of combining text and natural images like the two methods above.

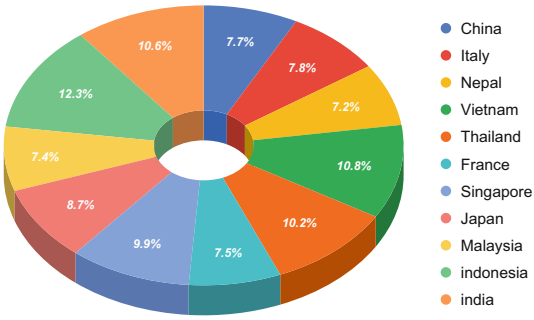

Fig. 2. The percentage of video hours by country in our collection

3 Landscape Dataset

3.1 Dataset Overview

Dataset Collection. We embarked on an extensive data collection endeavor, curating a rich repository of videos that showcased the unique landscapes of several diverse countries across the globe. Among the nations included in our comprehensive dataset are France, Vietnam, Japan, Thailand, Singapore, Malaysia, and more. Figure 1 illustrates the dataset we collected through 100h of video.

Dataset Statistics. Our collective efforts culminated in acquiring an impressive 100 h of video content, with the corresponding breakdown of video hours for each country thoughtfully illustrated in Fig. 2.

Our data collection strategy was meticulously designed to achieve two distinct objectives. First and foremost, we aimed to imbue our dataset with diversity, drawing from the rich tapestry of natural and urban environments found across these nations. This diversity ensures that our dataset encapsulates a broad spectrum of terrains, climates, and cultural influences, enabling a more comprehensive analysis.

In parallel, our data collection approach was strategically directed towards capturing the essence of each location. To achieve this, we meticulously selected videos featuring iconic landmarks, renowned architectural marvels, and the breathtaking natural wonders that define each region. By doing so, we have diversified our dataset and accentuated the distinctive characteristics that make each country unique.

In summary, our data collection efforts spanned multiple countries, focusing on both diversity and the representation of notable locations. This carefully cultivated dataset serves as the foundation for our research, enabling a robust exploration of landscape imagery and its connection to information retrieval.

3.2 Dataset Construction Pipeline

We initially converted the data collected from videos into images to prepare for the experimental phase. We opted to create image data by extracting frames

Fig. 3. Description of the image created from CLIP-Interrogator tool

from videos at a rate of 0.5 frames per second. In total, we obtained 360,000 images. We will refer to this dataset as the *LandscapeDataset*.

We further transformed the *LandscapeDataset* into sketch-like representations using the Holistically-Nested Edge Detection technique to facilitate experimentation involving the fusion of sketches and text for information retrieval.

Additionally, from the *LandscapeDataset*, we generated descriptive sentences for the images using the CLIP-Interrogator tool (Fig. 3).

With these preparatory steps completed, we are fully equipped for information retrieval through the fusion of text and sketch to generate images using Diffusion Models.

4 Methodology

4.1 Stable Diffusion Model

Diffusion probabilistic models (DM) [7] represents a novel and potent paradigm in the realm of probabilistic generative modeling. These models manifest as parameterized Markov chains, fine-tuned via variational inference, with the primary objective of generating samples that closely align with the underlying data distribution at a finite temporal horizon. Intriguingly, the transitions within this chain, denoted as $p_\theta(x_{t-1}|x_t)$, are systematically trained to reverse a diffusion process in a Markov chain that progressively injects noise into the data, counter to the sampling direction represented by $q_\theta(x_t|x_{t-1})$, ultimately leading to the gradual erosion of the original signal. Notably, when the diffusion mechanism involves a subtle infusion of Gaussian noise, the sampling chain

transitions can be elegantly streamlined to conditional Gaussians, simplifying the neural network parameterization. This simplification is encapsulated in the following objective function:

$$L_{DM} = E_{x,\epsilon_\theta \sim N(0,1),t}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2] \tag{1}$$

Furthermore, the framework employs denoising autoencoders $(\theta(x_t, t); t = 1, \ldots, T)$ trained meticulously to predict denoised variants of input data (x_t) derived from their noisy counterparts (x). This synthesis of principles from physics, variational inference, and neural networks underpins a powerful approach that holds significant promise for high-quality image synthesis and generative modeling, notwithstanding the substantial computational demands it entails.

Stable Diffusion. [13] advances the Diffusion model's foundational principles, significantly enhancing generative capabilities. Central to this innovation is integrating a Variational Autoencoder (VAE) with an encoder and decoder, elevating image fidelity. The VAE's encoder transforms data into a lower-dimensional latent form, efficiently reconstructed by the decoder. Stable Diffusion also employs a ResNet-enhanced U-Net architecture, preserving image integrity while reducing noise. A text encoder with CLIP text embeddings and cross-attention layers also enhances contextual awareness, conditioning image generation on textual cues. This pioneering approach merges diffusion principles, VAEs, U-Net structures, and text embeddings, yielding high-quality, contextually informed images with unmatched precision.

4.2 Architecture

ControlNet. [19], a vital component of Stable Diffusion, enhances image synthesis by providing precise control over content, style, and context. It employs advanced neural network structures, including attention mechanisms and disentangled representations, to focus on specific image regions and manipulate factors like pose and lighting. During training, ControlNet's loss function, similar to Stable Diffusion's but with added text (c_T) and latent conditions (c_F), ensures output consistency with specified criteria, enabling precise control and superior image synthesis:

$$L_{ControlNet} = E_{x,\epsilon_\theta \sim N(0,1),t,c_T,c_F}[||\epsilon - \epsilon_\theta(x_t, t, c_T, c_F)||_2^2] \tag{2}$$

This loss function is meticulously designed to optimize the model's performance by minimizing the discrepancy between the generated images and the target images, while also considering the alignment with textual prompts and control parameters. The loss function serves as a crucial metric for monitoring the model's progress and convergence behavior, ensuring that the generated images exhibit high fidelity and align closely with the intended context and narrative.

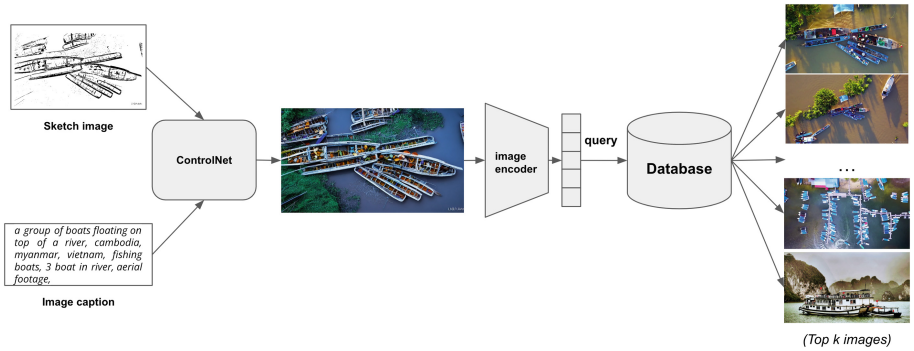

Fig. 4. System overview

Landscape-Based Multimodal Scene Retrieval Enhanced by Diffusion Model: We propose a novel system with three integral components, each contributing to the seamless synthesis and retrieval of landscape images from user inputs. We begin the process with the input sketch image and the user's descriptive sentence. We perform feature extraction from a sketch, utilizing techniques that capture salient visual attributes, contours, and structural elements. Simultaneously, we extract features from the user's textual input, harnessing natural language processing methods to identify keywords and context within the sentence. The significance of this dual feature extraction is twofold:

- Empower us to distill essential visual and semantic information from both modalities.
- Ensure that the extracted data have the same dimensionality, enabling seamless integration. The result is a pair of feature vectors, one representing the sketch and the other encapsulating the textual input, quantifying the input data's essence.

Building on these feature vectors, we leverage the ControlNet architecture, a cutting-edge generative model, to expedite the image generation process. ControlNet expertly synthesizes an image that aligns with the user's input, bridging the gap between the sketch and the textual description. Figure 4 comprehensively illustrates our system's workflow, highlighting the intricate synergy between these components. Furthermore, armed with the newly generated images, we embark on a subsequent phase dedicated to meticulous feature extraction. This process involves a detailed analysis of the visual attributes and characteristics embedded within the synthesized images. These extracted image features form the foundation for our querying procedure, which revolves around our Landscape Dataset - a meticulously curated repository housing an extensive assortment of landscape images. Implementing robust techniques, including similarity measures and content-based image retrieval methodologies, facilitates the querying process. These techniques are instrumental in identifying and ranking the top k images that closely align with the user's input, with the flexibility to adjust the value of k according to specific preferences and requirements.

5 Experiments

5.1 Experimental Settings

In pursuit of a comprehensive investigation into the matter, we embarked on an exhaustive exploration utilizing various tools at our disposal. To ensure a rigorous examination, we undertook a systematic approach by experimenting with various combinations of components. For the vital aspect of information representation, we diligently employed no fewer than three distinct models to generate the essential feature maps from both the source data and query data. These models included ViT-B-16, which was employed exclusively for images, ViT-B-32, which was utilized for both images and text, and EfficientNet-B4 [18], specially chosen for its efficacy in processing image data. This judicious selection of models was undertaken to ensure a comprehensive analysis yielding insightful results. Using these models as components, we have created six experiments reported in Table 1:

- **ViT ControlNet:** In this experiment, we use ViT B-16 to compute the feature map from ControlNet-generated query image. Thus, this feature map should contain text and sketch data information. ViT B-16 is also used to compute the feature maps of the whole source dataset. Thus, these computed feature maps exist in the same feature space, all having the same dimension of (1000) and can be compared with each other.
- **ViT Edge:** This experiment is similar to ViT ControlNet but differs at the input stage. Instead of using fused ControlNet-generated images, only the edge map is used, and the ControlNet image generation step is skipped altogether. Because both the dataset feature map and the edge map features are generated with ViT B-16, they, therefore, are also in the same feature space. This makes it possible to make comparisons of the edge maps against the image dataset.
- **EfficientNetB4-ControlNet:** This experiment is also similar to ViT ControlNet. However, we replace ViT-B16 with EfficientNet-B4. EfficientNet-B4 is used to compute the feature maps of ControlNet-generated images and the source dataset. The feature space is thusly the same with dimensions of (1792, 7, 7), and we make comparisons in this space.
- **ViT-CLIP-ControlNet:** We use CLIP-ViT B32, a cross-modal model capable of fusing both textual and visual inputs into a common feature space. ControlNet-generated images and the source dataset are used to compute feature maps in this experiment. The resultant feature maps have a dimension of (1, 512) and are comparable across the query and the source dataset.
- **ViT-CLIP-Prompt:** This method is similar to ViT-CLIP-ControlNet. However, instead of using visual inputs, we only use textual inputs from the ViT L-14 generated prompts to generate feature maps. We can perform the same comparisons using the source dataset feature maps from ViT-CLIP-ControlNet.

– **Vit-CLIP-Edge:** Similar to ViT-CLIP-ControlNet, we use CLIP-ViT B32 as feature map generator. However, in this experiment, we only use the edge maps instead of generating images of ControlNet. Thus, the information about the prompt is absent in the generated feature map.

Table 1. Overview of the experiments

Experiment name	Textual input	Visual input	Feature space (feature size)
ViT-Edge	(none)	Edge maps	ViT-B-16 (1000)
ViT-ControlNet	(ViT-L-14) generated prompts	ControlNet-generated images	ViT-B-16 (1000)
EfficientNetB4-ControlNet	ViT-L-14 generated prompts	ControlNet-generated images	EfficientNet-B4 (1792, 7, 7)
ViT-CLIP-ControlNet	(none)	ControlNet-generated images	ViT-B-32 (1, 512)
ViT-CLIP-Prompt	ViT-L-14 generated prompts	(none)	ViT-B-32 (1, 512)
ViT-CLIP-Edge	(none)	Edge maps	ViT-B-32 (1, 512)

5.2 Implementation Details

After obtaining the data described in Sect. 3.2, we conducted experiments on one-third of the dataset. The images selected for experimentation were chosen randomly. We conducted the experiments and extracted feature vectors for both the images and their corresponding descriptive sentences. With the outputs obtained during this phase, we evaluated the performance using metrics such as Top-1, Top-5, and Top-20, as explained in the evaluation methodology outlined in Sect. 5.1. These experimental results are presented in Sect. 5.3.

5.3 Experimental Results

Qualitative Observations. We investigate and assess the performance of our innovative method, leveraging a carefully curated testing dataset as depicted in Fig. 5. Our evaluation process is thoughtfully designed to encompass a wide spectrum of real-world scenarios, ensuring our approach's robustness and versatility. To elaborate on LUMOS-DM, we begin with the original image, positioned on the left side, serving as the primary target for retrieval. Within our extensive image repository, we initiate the process by subjecting the image to holistic edge detection as a pivotal initial step. The outcome of this stage, illustrated in the middle, is an edge map that accentuates the structural contours and distinctive

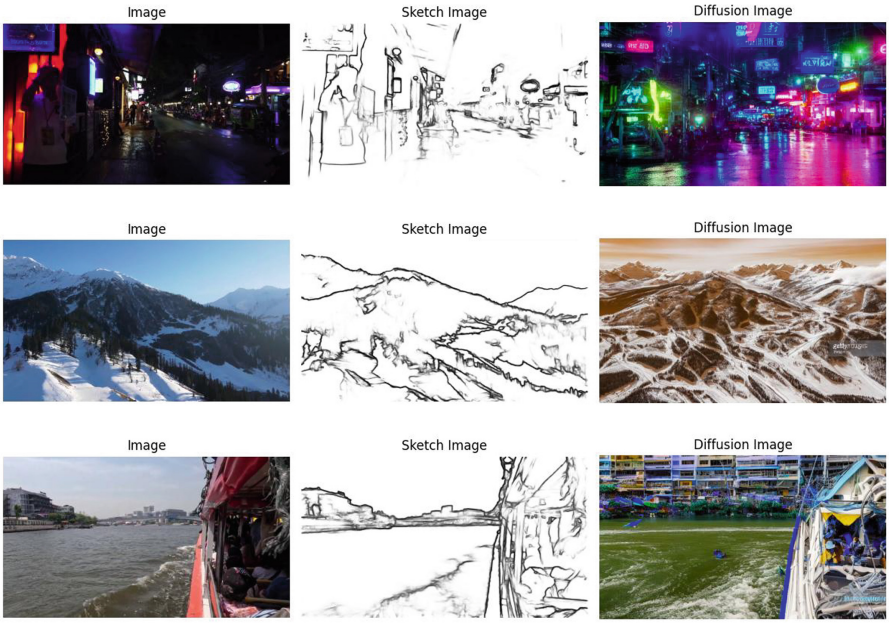

Fig. 5. Illustration images in order: original image (left), sketch image (middle) and image created from diffusion model (right)

features within the image. However, our innovation doesn't halt here; we further enhance the interpretability and predictive prowess of LUMOS-DM by generating a descriptive textual prompt using the ViT model. This prompt is seamlessly integrated with the previously derived edge map, culminating in creating the prediction image, elegantly showcased on the right side of the visual spectrum. This intricate sequence of transformations represents the core of LUMOS-DM, where both visual and semantic information harmoniously converge to yield accurate and meaningful predictions. Subsequently, the ensuing evaluation delves into the performance of this method across diverse scenarios, shedding light on its inherent strengths and capabilities within the domain of image retrieval and prediction.

Quantitative Observations. Figure 6 visually represents the initial effectiveness of various methods and models in interacting with a query image. This preliminary observation sets the stage for a detailed exploration of specific accuracy metrics. Notably, our VIT-CLIP-ControlNet achieves a remarkable Top-1 accuracy of 3.22%, underscoring its capacity to consistently rank the most pertinent image as the top prediction. Extending our analysis to Top-5 accuracy, we maintain strong performance at 9.27%, affirming the model's ability to encompass the correct result within the top quintet of predictions. Notably, Top-20 accuracy soars to 18.79%, accentuating its proficiency in capturing pertinent

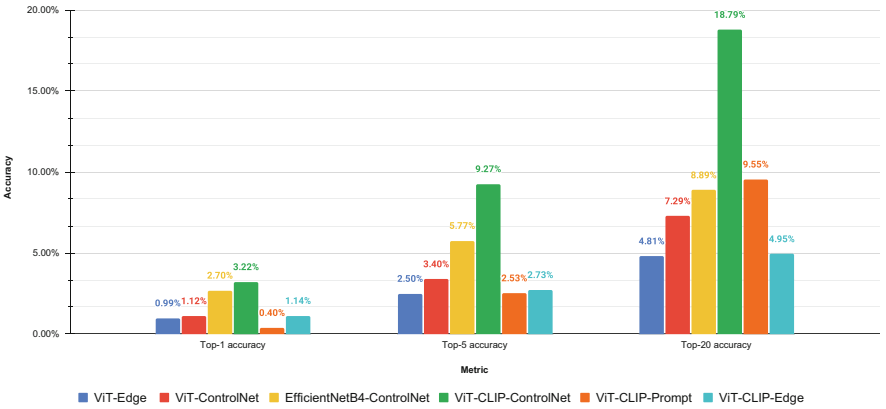

Fig. 6. Accuracy of different retrieval methods across different metrics on a dataset of 21,000 pictures.

images across a broader spectrum of predictions-particularly valuable in contexts featuring expansive image repositories. Our meticulous comparative analysis substantiates VIT-CLIP-ControlNet's supremacy over alternative methods; for instance, ViT-Edge attains a modest 0.99% Top-5 accuracy, while ViT-CLIP-Prompt fares slightly better at 2.53%, yet both fall significantly short of our model's performance. Furthermore, statistical significance tests substantiate the reliability and significance of these performance disparities, solidifying the excellence of VIT-CLIP-ControlNet in image retrieval and prediction tasks.

6 Conclusion

In this paper, we introduce the Landscape Dataset, which was collected from YouTube and comprises a total duration of 100 h of video content. This dataset encompasses various landscapes from different countries around the world. We propose a LUMOS-DM: Landscape-based Multimodal Scene Retrieval Enhanced by Diffusion Model method using Vision Transformer and Stable Diffusion. This method enhances information retrieval efficiency by expanding the input modalities to include descriptive sentences and sketch-like images. This facilitates more user-friendly information retrieval from videos.

LUMOS-DM opens up numerous potential applications in our daily lives. One of the simplest examples of this can be seen in how LUMOS-MD can significantly support individuals involved in affiliate marketing. Content creators will no longer face the challenges of searching for videos. Instead, they only need to provide a description and a basic sketch of the scene they desire, and LUMOS-MD will suggest corresponding video clips. This streamlines the content creation process and facilitates reaching customers more efficiently.

References

1. Alqasrawi, Y.: Bridging the gap between local semantic concepts and bag of visual words for natural scene image retrieval (2022)
2. Bibi, R., Mehmood, Z., Munshi, A., Yousaf, R.M., Ahmed, S.S.: Deep features optimization based on a transfer learning, genetic algorithm, and extreme learning machine for robust content-based image retrieval. PLoS ONE **17**(10), e0274764 (2022)
3. Chang, S., Kopp, M., Ghamisi, P.: Dsfer-Net: a deep supervision and feature retrieval network for bitemporal change detection using modern hopfield networks (2023)
4. Chen, J., Lai, H.: Ranking-aware uncertainty for text-guided image retrieval. ArXiv abs/2308.08131 (2023). https://api.semanticscholar.org/CorpusID:260926537
5. Chen, Y., Wang, J., Lin, L., Qi, Z., Ma, J., Shan, Y.: Tagging before alignment: integrating multi-modal tags for video-text retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, pp. 396–404 (2023). https://doi.org/10.1609/aaai.v37i1.25113, https://ojs.aaai.org/index.php/AAAI/article/view/25113
6. Gong, Y., Cosma, G.: Boon: a neural search engine for cross-modal information retrieval (2023)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
8. Ibrahimi, S., Sun, X., Wang, P., Garg, A., Sanan, A., Omar, M.: Audio-enhanced text-to-video retrieval using text-conditioned feature alignment (2023)
9. Li, H., Yang, S., Zhang, Y., Tao, D., Yu, Z.: Progressive feature mining and external knowledge-assisted text-pedestrian image retrieval (2023)
10. Lin, C., et al.: Text-adaptive multiple visual prototype matching for video-text retrieval. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022). https://openreview.net/forum?id=XevwsaZ-4z
11. Lin, X., et al.Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval (2023)
12. Pradhan, J., Pal, A.K., Hafizul Islam, S.K., Bhaya, C.: DNA encoding-based nucleotide pattern and deep features for instance and class-based image retrieval. IEEE Trans. Nanobiosc. **23**, 190–201 (2023)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
14. Scotti, P.S., et al.: Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors (2023)
15. Siddharth, L., Li, G., Luo, J.: Enhancing patent retrieval using text and knowledge graph embeddings: a technical note (2022)
16. Staszewski, P., Jaworski, M., Cao, J., Rutkowski, L.: A new approach to descriptors generation for image retrieval by analyzing activations of deep neural network layers. IEEE Trans. Neural Netw. Learn. Syst. **33**(12), 7913–7920 (2022)
17. Sun, X., et al.: Pre-training with aspect-content text mutual prediction for multi-aspect dense retrieval (2023)

18. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
19. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)