WHU-NERCMS@TRECVID 2023 Ad-hoc Search Task

Jiangshan He

riverhill@whu.edu.cn

Hubei Key Laboratory of Multimedia and Network Communication Engineering

National Engineering Center for Multimedia Software

School of Computer Science, Wuhan University

November 14, 2023

Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Introduction

- Ad-hoc Search (AVS)
  - As many as possible shots that match the input
  - Compared to Known Item Search（KIS）

ID	Topic
735	A toy vehicle
746	A man riding a scooter
749	A person wearing any kind of face or head mask
750	A man with an earring in his left ear

ID	Query
vbs23-kis-t8	View down from the helmet camera of a mountain biker, as he spins around on a path along a narrow ridge. He spins by jumping on the back wheel. The ridge is flanked by sea. We hear the biker narrating the scene.

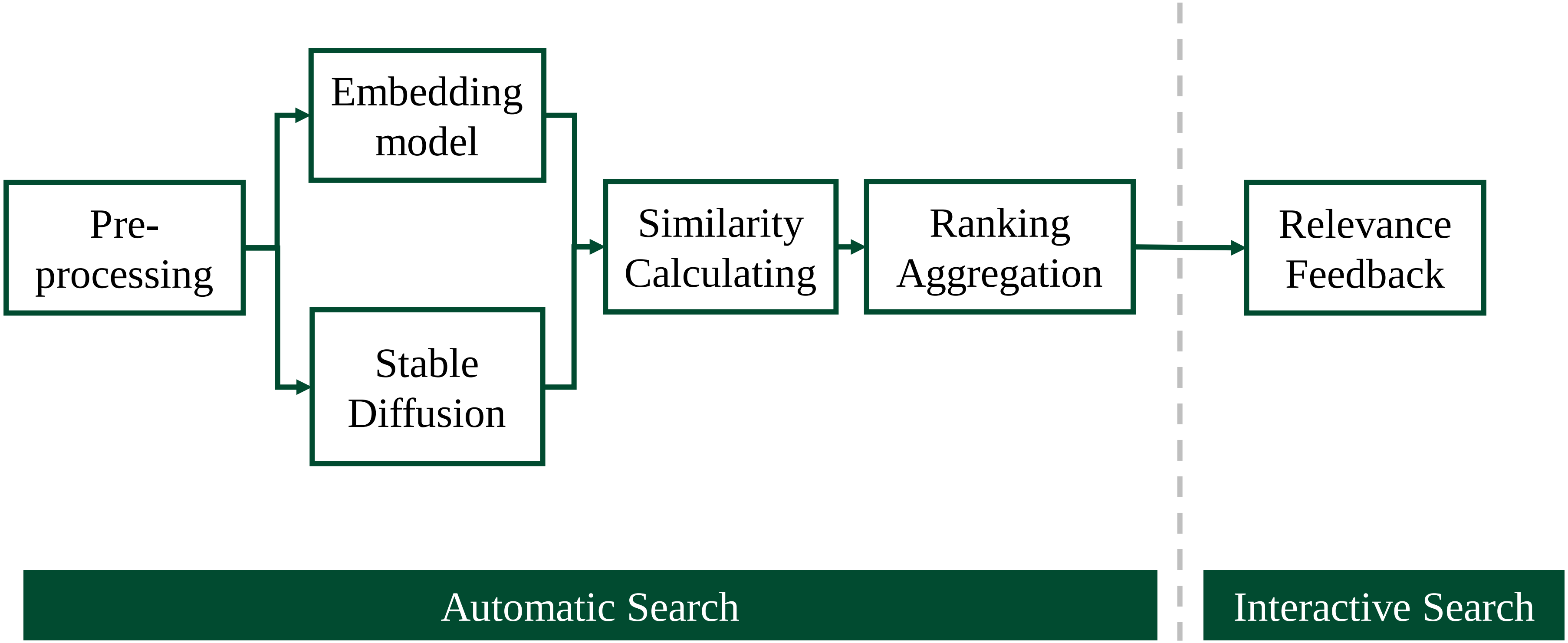
Examples of 2023 AVS topics & VBS KIS-T query

Outline

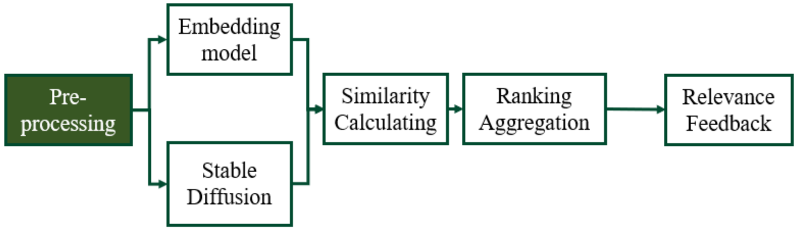
- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Solution

- Framework



Solution


Pre-processing

Embedding model

Stable Diffusion

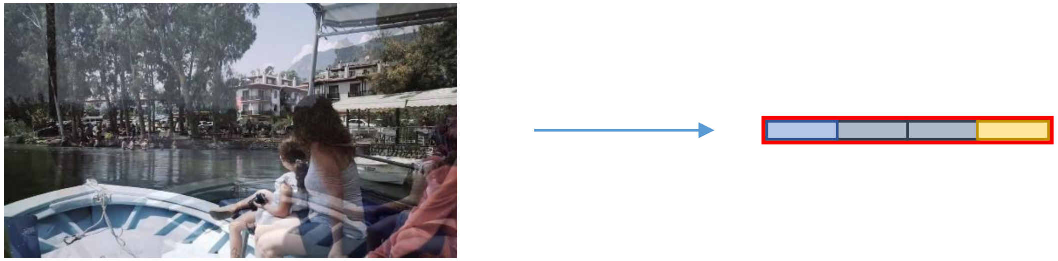
Similarity Calculating

Ranking Aggregation

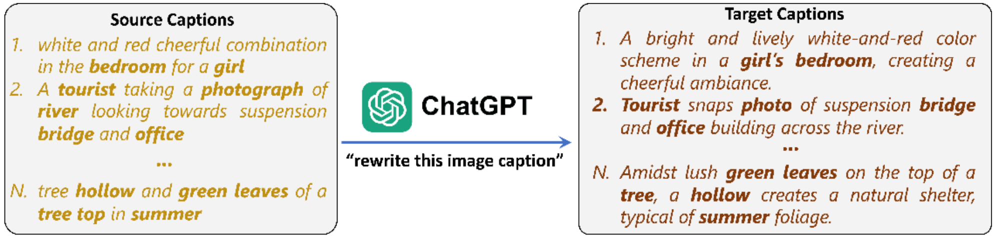
Relevance Feedback

■ Step 1 : Pre-processing

- Keyframes (Official dataset) →Image Embeddings


shot17235_9_RKF.png

- Embedding models:
    - ➢ CLIP [Radford+, 2021](8)
    - ➢ SLIP [Mu+, 2021](5)
    - ➢ BLIP [Li+,2022](4)
    - ➢ BLIP-2 [Li+,2023](1)
    - ➢ LaCLIP [Fan+,2023](1)


Source Captions

1. *white and red cheerful combination in the **bedroom** for a **girl***
2. *A **tourist** taking a **photograph** of **river** looking towards suspension **bridge** and **office***

...

N. *tree **hollow** and **green leaves** of a **tree top** in **summer***

"rewrite this image caption"

Target Captions

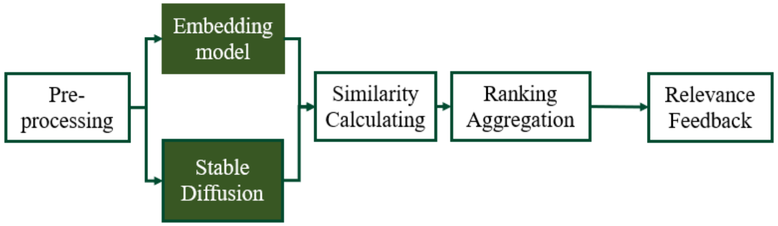
1. *A bright and lively white-and-red color scheme in a **girl's bedroom**, creating a cheerful ambiance.*
2. ***Tourist** snaps **photo** of suspension **bridge** and **office** building across the river.*

...

N. *Amidst lush **green leaves** on the top of a **tree**, a **hollow** creates a natural shelter, typical of **summer** foliage.*

LaCLIP

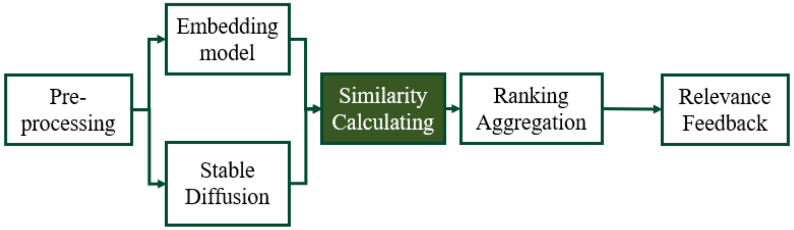
Solution



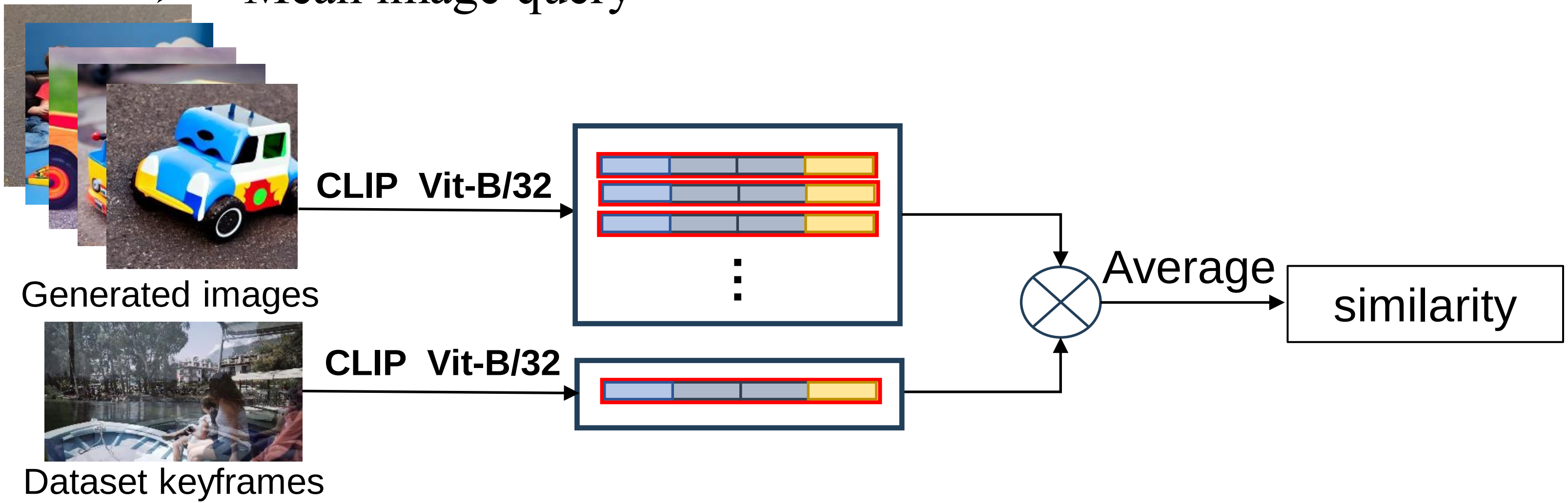
Step 2&3:Embedding model & Stable Diffusion

- Extract text embeddings
  - From official topics
  - Various models
- Generate abundant images
  - model : stable-diffusion-v1-5
  - 1000 images for one topic
  - “ A toy vehicle”

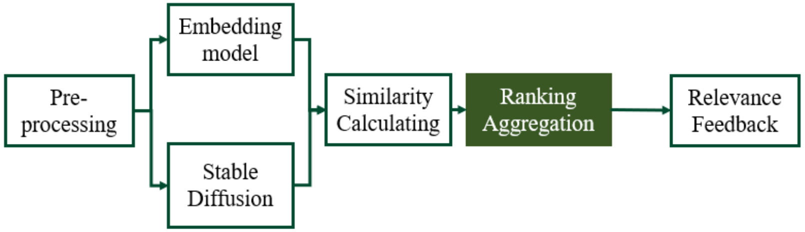
Solution



- Step 4: Similarity Calculating
  - Corresponding embeddings
    - Cosine similarity
    - Average of their similarities for different pre-trained models
    - BLIP : ViT-B(COCO), ViT-B(Flickr30k), ViT-L(COCO), ViT-L(Flickr30k)
    - “Mean image query”



Solution



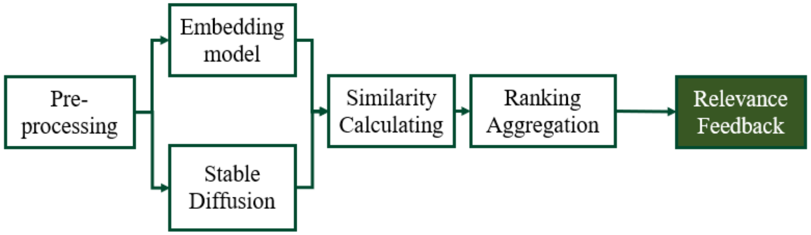
Step 5: Ranking Aggregation

- 5 similarity lists —— Embedding models
- 1 similarity list —— Stable Diffusion
- Weights based on performance in 2022 AVS topics

Type	Run ID	infAP	Weight(C:S:B:B2:L:D)
Automatic	F_1	0.292	10:3:16:4:3:3
	F_2	**0.292**	10:3:16:4:3:6
	F_3	0.291	10:3:20:4:3:3
	F_4	0.290	13:3:16:4:3:3

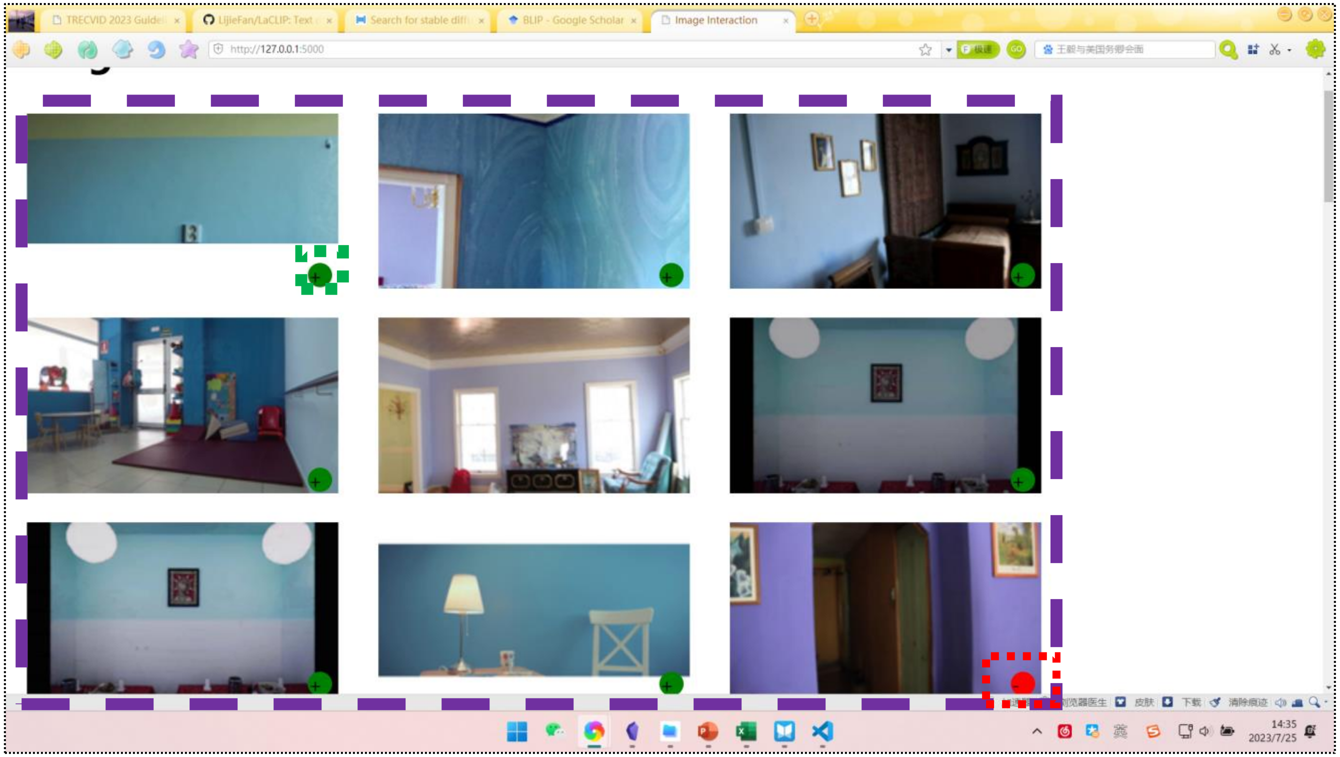
Concrete weights

Solution


Pre-processing
Embedding model
Stable Diffusion
Similarity Calculating
Ranking Aggregation
Relevance Feedback

Step 6: Relevance Feedback

- Judge TOP 30 results **for 3 iterations**.
- Simple GUI

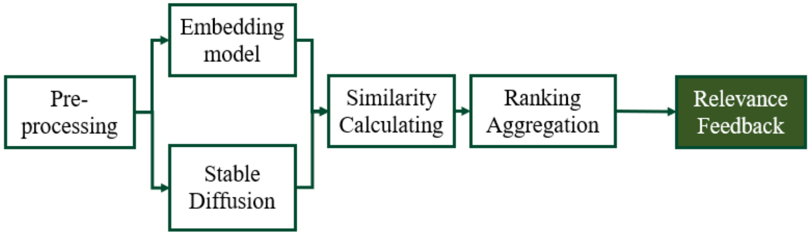


① Feedback Area

② Positive Feedback

③ Positive Feedback

Solution



■ Step 6: Relevance Feedback

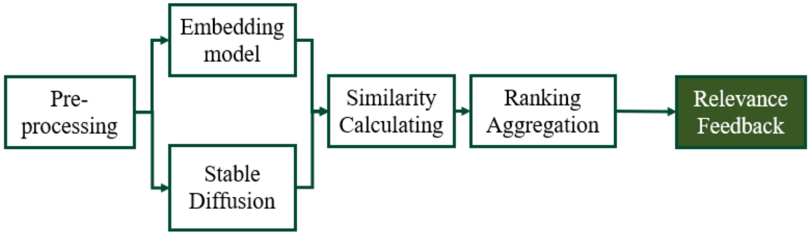
- Interactive Ranking Aggregation (IRA)
  - ➢ Positive feedback increases weight
  - ➢ Negative feedback decreases weight

$$\tilde{w}_i = \frac{1}{|\Phi_+|}\sum_{d_j \in \Phi_+} s_i^{d_j} - \frac{1}{|\Phi_-|}\sum_{d_k \in \Phi_-} s_i^{d_k}$$

  - ➢ A smooth update of the weight

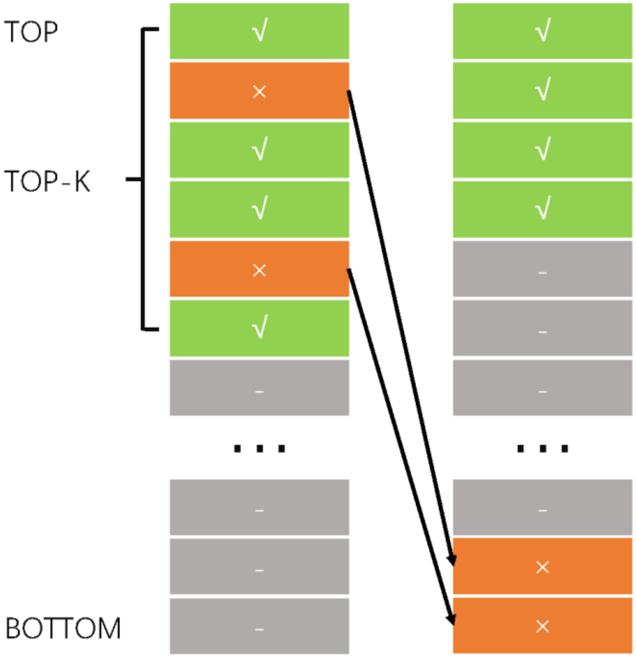
$$w_i = \alpha\tilde{w}_i + (1-\alpha)w_i, \alpha = 0.9$$

Solution



Step 6: Relevance Feedback

- Top-K feedback
  - Positive feedback puts first
  - Negative feedback puts last
- Two acceleration schemes
  - Only positive feedback (final choice)
  - Only negative feedback



Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Model Selection Strategy

- More diverse
- Fewer poor pre-trained model

Abbreviation	Description	Abbreviation	Description
C	CLIP	S	SLIP
B	BLIP	B2	BLIP-2
L	LaCLIP	D	Diffusion

Model Selection Strategy

- More diverse

Type	infAP	Type	infAP
C	0.1603	S	0.1286
B	0.1857	B2	0.1585
L	0.0931	D	0.0788

Type	infAP
C+S	0.1835
C+S+D	0.1891
C+B+S+D	0.2363
C+B+S+D+B2	0.2604
C+B+S+D+B2+L	**0.2636**

Model Selection Strategy

- Fewer poor pre-trained model

Type	Pre-trained type	infAP	Fusion infAP		
BLIP	ViT-Base	0.0745	0.149	--	0.1724
	ViT-Large	0.0769		--	
	ViT-B (Flickr30k)	0.1293		**0.1857**	
	ViT-B (COCO)	0.1333			
	ViT-L (Flickr30k)	0.1447	--		
	ViT-L (COCO)	0.1623	--		

Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Experiments

- The Same Modality
  - As “mean image query”
  - Imagination of what might appear in the dataset？
  - Text vs Text

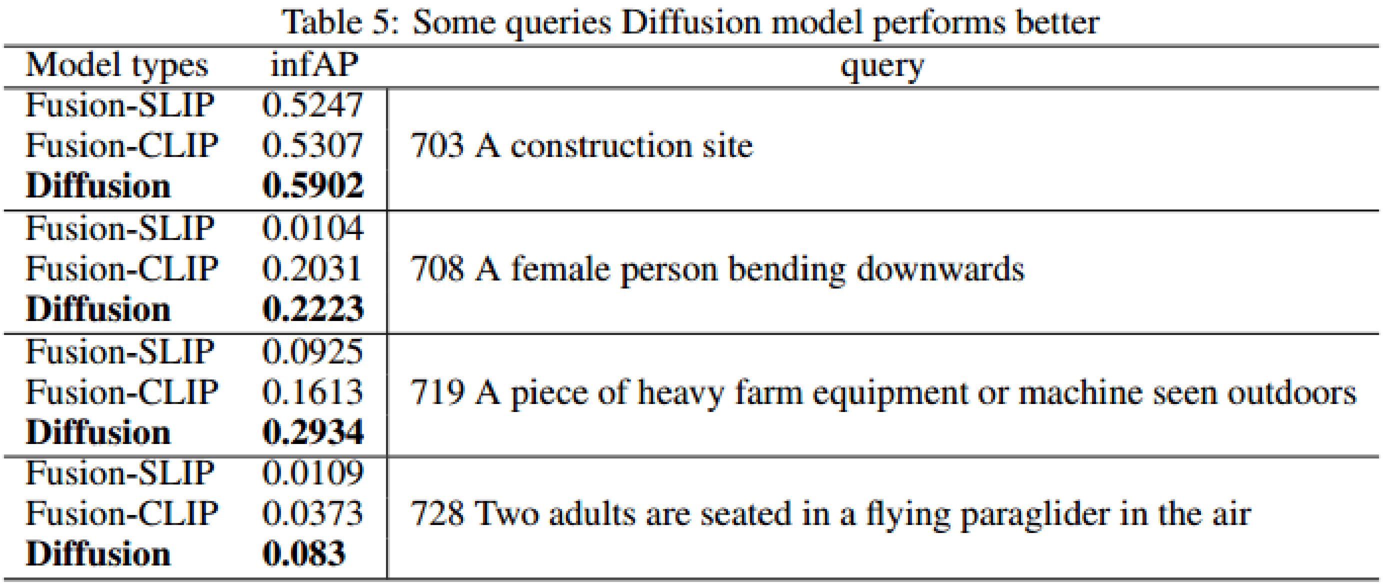


Table 5: Some queries Diffusion model performs better

Model types	infAP	query
Fusion-SLIP	0.5247	703 A construction site
Fusion-CLIP	0.5307	
Diffusion	**0.5902**	
Fusion-SLIP	0.0104	708 A female person bending downwards
Fusion-CLIP	0.2031	
Diffusion	**0.2223**	
Fusion-SLIP	0.0925	719 A piece of heavy farm equipment or machine seen outdoors
Fusion-CLIP	0.1613	
Diffusion	**0.2934**	
Fusion-SLIP	0.0109	728 Two adults are seated in a flying paraglider in the air
Fusion-CLIP	0.0373	
Diffusion	**0.083**	

Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Experiments

Interactive Algorithm

- Result

Priority	Automatic run infAP	Interactive run infAP	Performance
1	0.292	0.299	+0.007
2	**0.292**	0.298	+0.006
3	0.291	**0.299**	**+0.008**
4	0.290	0.296	+0.006

- Drawbacks
  - Extreme conditions
  - Algorithm cannot access complex semantic information.

Outline

- Introduction
- Solution
- Experiments
  - Model Selection Strategy
  - The Same Modality
  - Interactive Algorithm
- Conclusion and Future Work

Conclusion and Future Work

- Conclusion
  - Multiple Embedding model
  - Stable Diffusion
  - Fusion by weights
  - Interactive Ranking Aggregation
- Future Work
  - Interactive algorithm in terms of semantic
  - Reduce search time
  - LLM

Thanks for your time!

A team’s work presented by

Jiangshan He, Hong Zhang,

Zhengqian Wu and Chao Liang*

(* indicates corresponding author)

Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University