Training-free Zero-shot
Composed Image Retrieval via
***Wei**ghted **Mo**dality Fusion and Similarity*

Ren-Di Wu, Yu-Yen Lin, Huei-Fang Yang
National Sun Yat-sen University, Kaohsiung, Taiwan

Date: Dec. 06, 2024

What is Composed Image Retrieval (CIR)?

Definition

- Reference Image: Base visual input.
  - (eg. A blue T-Shirt)

What is Composed Image Retrieval (CIR)?

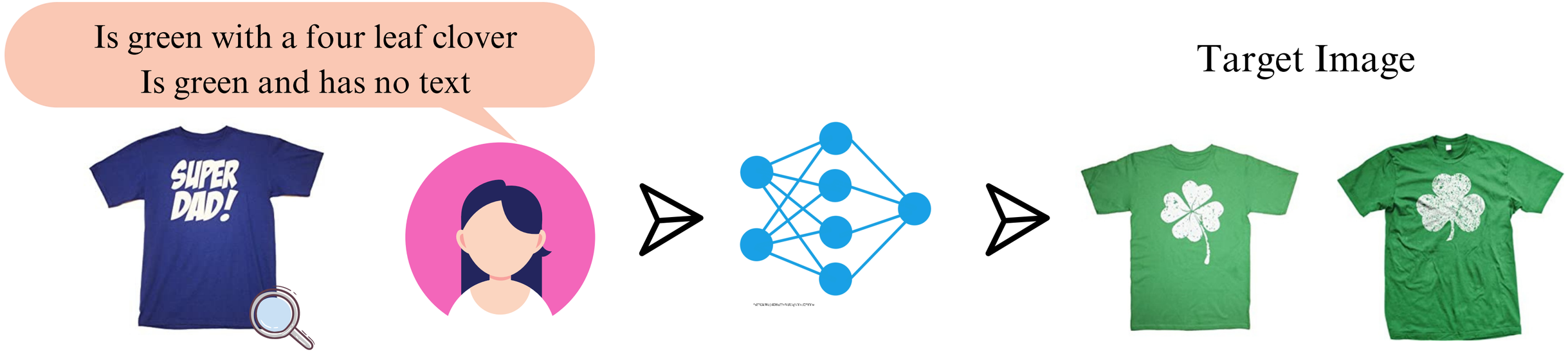
Definition

- Reference Image: Base visual input.
  - (eg. A blue T-Shirt)
- Text Modifier: Describes the desired transformation or addition.
  - (eg. Convert the pattern with 4 leaf clover and turn into green)

What is Composed Image Retrieval (CIR)?

Definition

- Reference Image: Base visual input.
  - (eg. A blue T-Shirt)
- Text Modifier: Describes the desired transformation or addition.
  - (eg. Convert the pattern with 4 leaf clover and turn into green)



Applications of CIR: What can we achieve?

Applications

- Fashion recommendation
  - (e.g., "The user might also like a 4-leaf clover design one").

The user also likes the 4-leaf clover

Recommend this item

Applications of CIR: What can we achieve?

Applications

- Fashion recommendation
  - (e.g., "The user might also like a 4-leaf clover design one").
- Real-world tasks - Image database search engine
  - (e.g., "Find a similar photo but without the people in the picture").

The user also likes the 4-leaf clover

Recommend this item

I want a image without the people

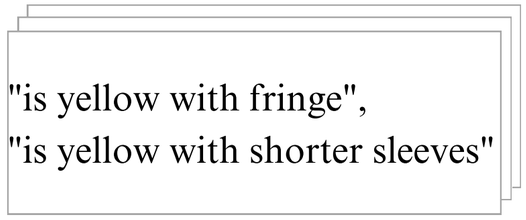
Database find and return relative image

Challenges in CIR

- Requires labor-intensive triplet collection for supervised learning.

Issue: Triplet collection

1. Query Image



2. Modified Descriptions

3. Target Image

Challenges in CIR

- Requires labor-intensive triplet collection for supervised learning.
- Models trained on specific domain often fail to generalize well to different domains.

Issue: Triplet collection

1. Query Image

"is yellow with fringe",
"is yellow with shorter sleeves"

2. Modified Descriptions

3. Target Image

Issue: Dataset is for specific domain

≠

Challenges in CIR

- Requires labor-intensive triplet collection for supervised learning.
- Models trained on specific domains often fail to generalize well to different domains.
- Previous Zero-shot CIR relies heavily on additional datasets pertaining to fusion.

Issue: Triplet collection

1. Query Image

"is yellow with fringe",
"is yellow with shorter sleeves"

2. Modified Descriptions

3. Target Image

Issue: Dataset is for specific domain

≠

Scalability vs. Efficiency

- State-of-the-art (SOTA) methods excel at large scales but are resource-intensive.

Model Name	CLIP B/32	CLIP L/14	CLIP H/14	CLIP G/14
Number Of Parameter	151M	428M	986M	1.37B

Scalability vs. Efficiency

- State-of-the-art (SOTA) methods excel at large scales but are resource-intensive.
- Smaller-scale models are more efficient but often lack performance.

Model Name	CLIP B/32	CLIP L/14	CLIP H/14	CLIP G/14
Number Of Parameter	151M	428M	986M	1.37B

Developing Training-free Approaches for CIR

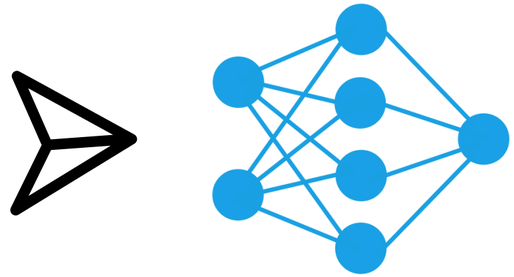
CIR Challenge

Scalability vs. Efficiency

Query Image

"is yellow with fringe",
"is yellow with shorter sleeves"

Modified Descriptions



Target Image

Developing Training-free Approaches for CIR

CIR Challenge

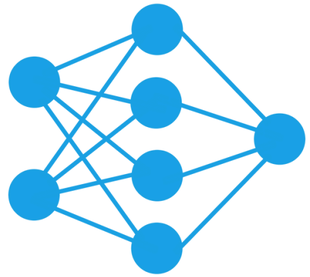
Scalability vs. Efficiency

- A training-free method that maintains
  a. Efficiency
  b. Ensures generalization
  c. Balances with scalability

Query Image

"is yellow with fringe",
"is yellow with shorter sleeves"

Modified Descriptions

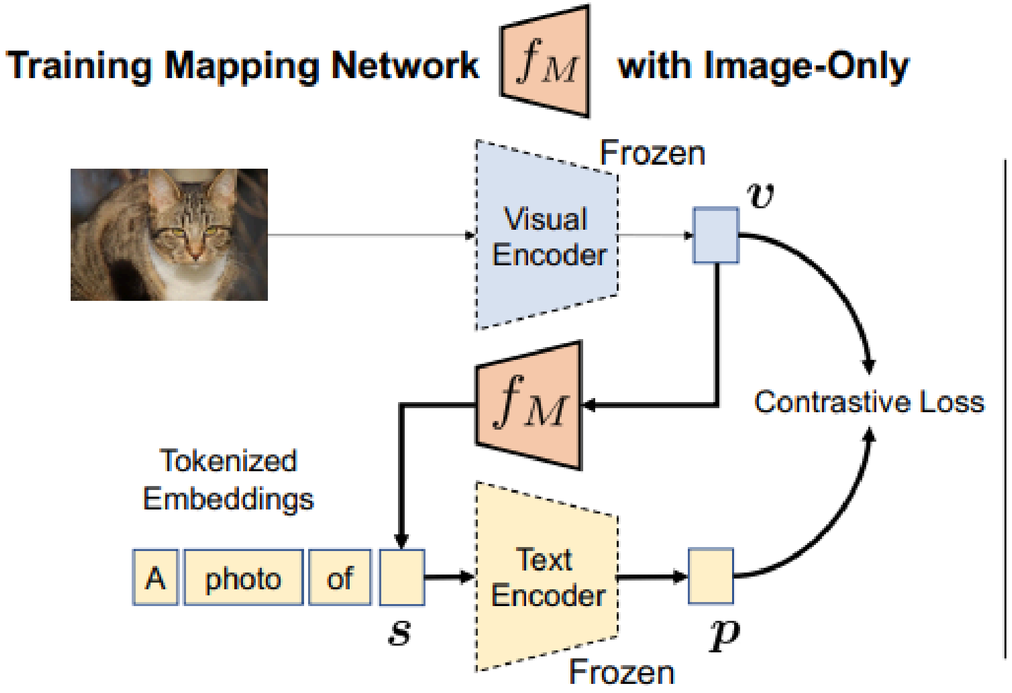


A method to solve this

Target Image

Pic2Word: Pioneering Zero-Shot CIR with Pseudo-Word Mappings

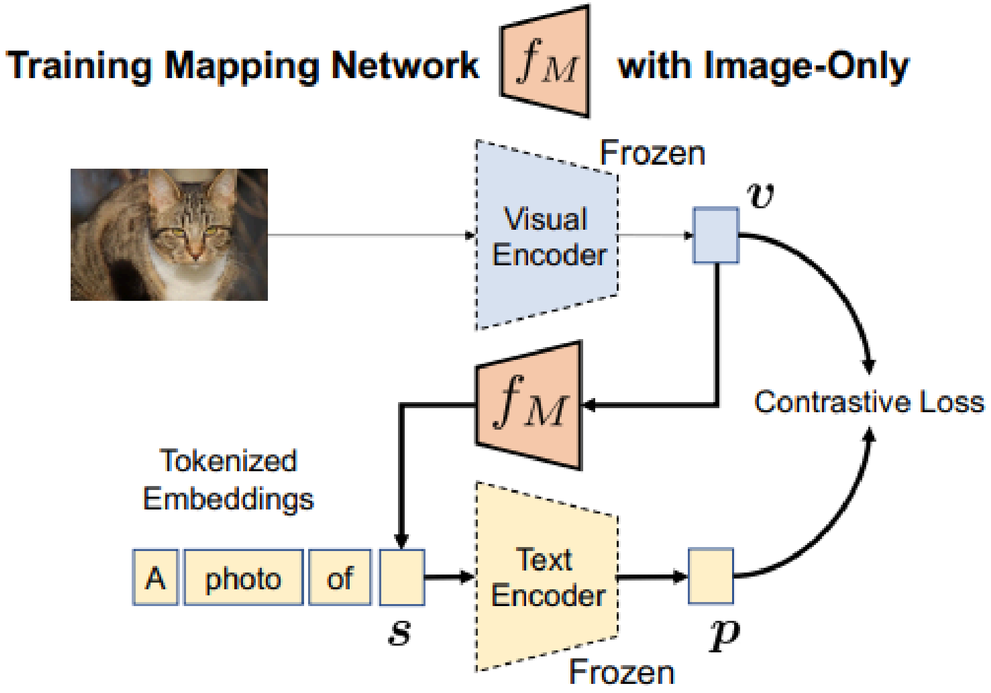
Mapping network to align image and text embeddings.

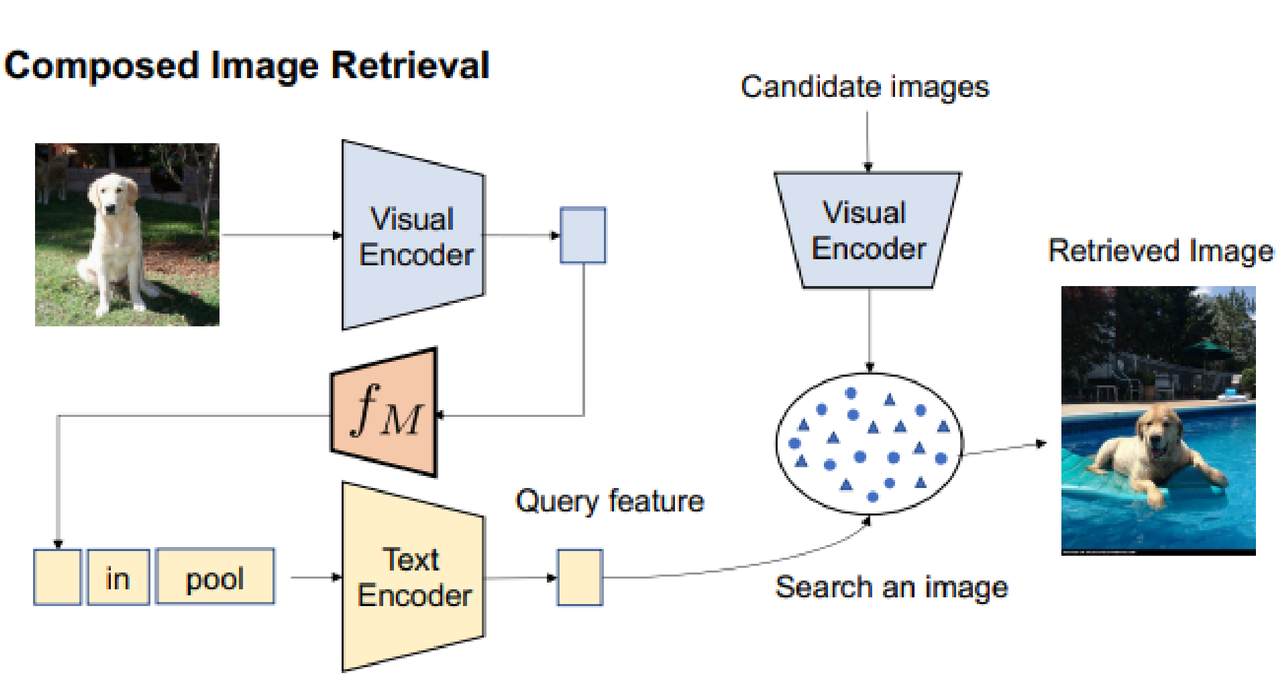


Pic2Word: Pioneering Zero-Shot CIR with Pseudo-Word Mappings

Mapping network to align image and text embeddings.

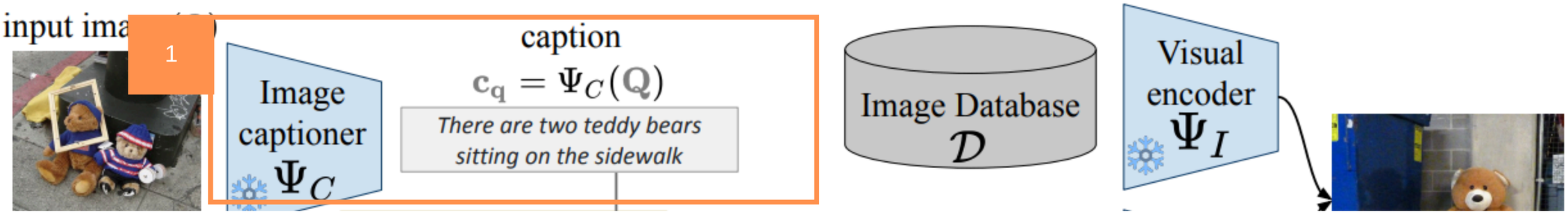
Pseudo-word tokens are used to complete text prompts for retrieval





CIReVL: Achieving Training-Free ZS-CIR via Vision-Language Embeddings

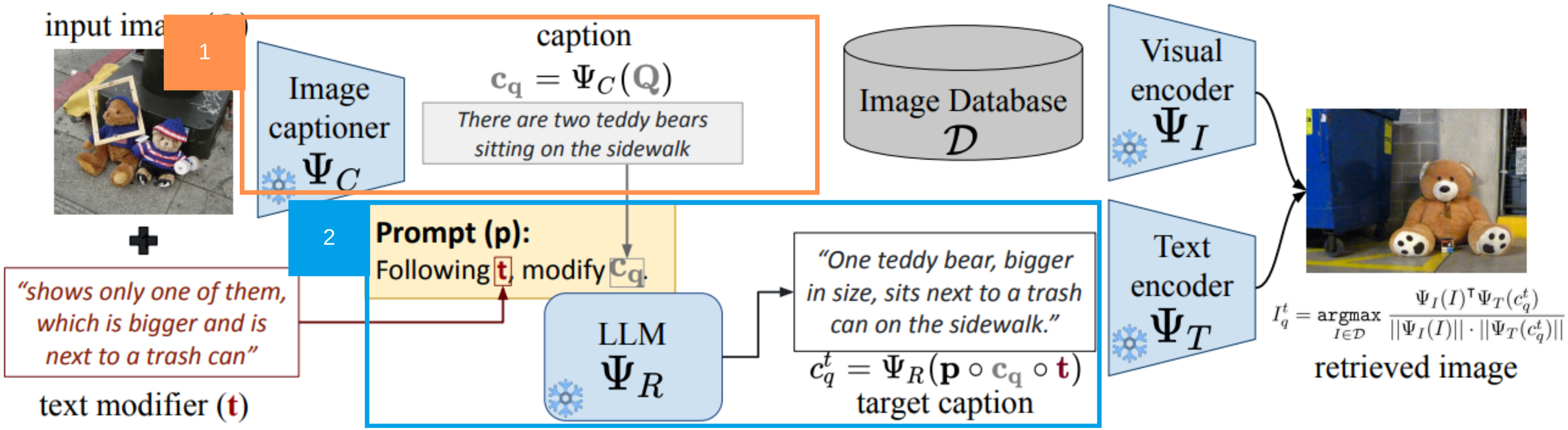
1 Captioning the image using a VLM



CIReVL: Achieving Training-Free ZS-CIR via Vision-Language Embeddings

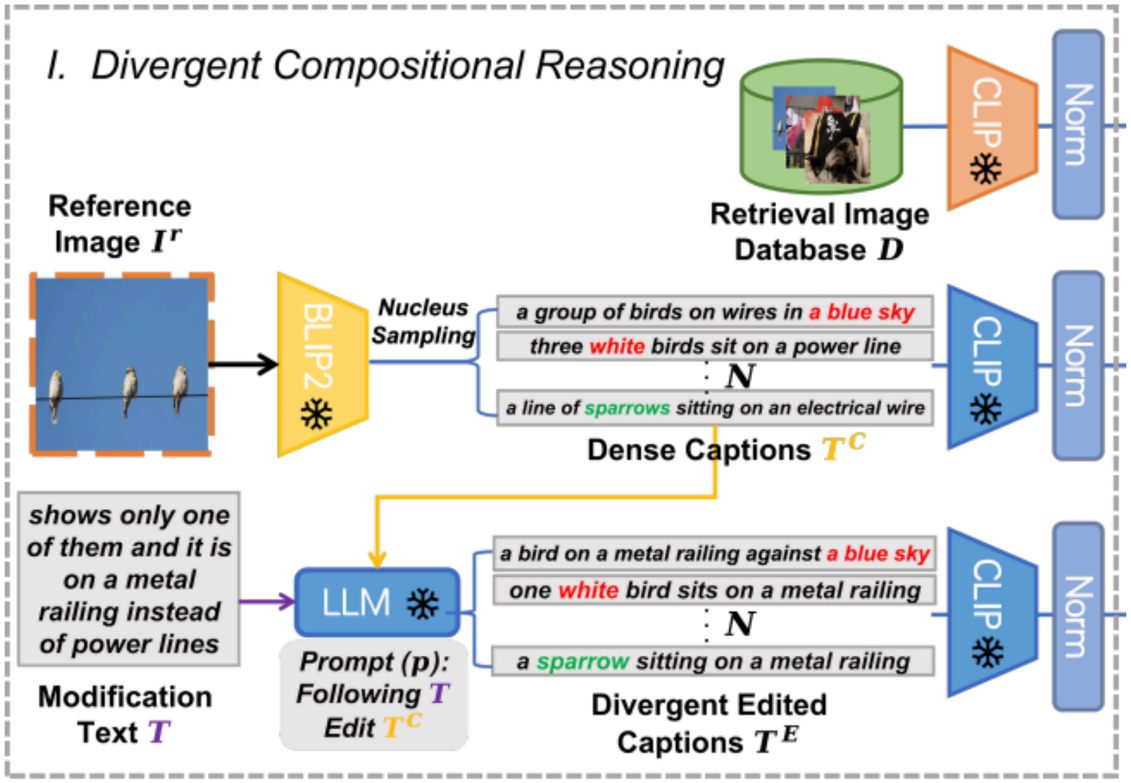
1 Captioning the image using a VLM

2 LLM creates the target image description



LDRE: Advancing ZS-CIR with Diverse Captions and LLM Reasoning

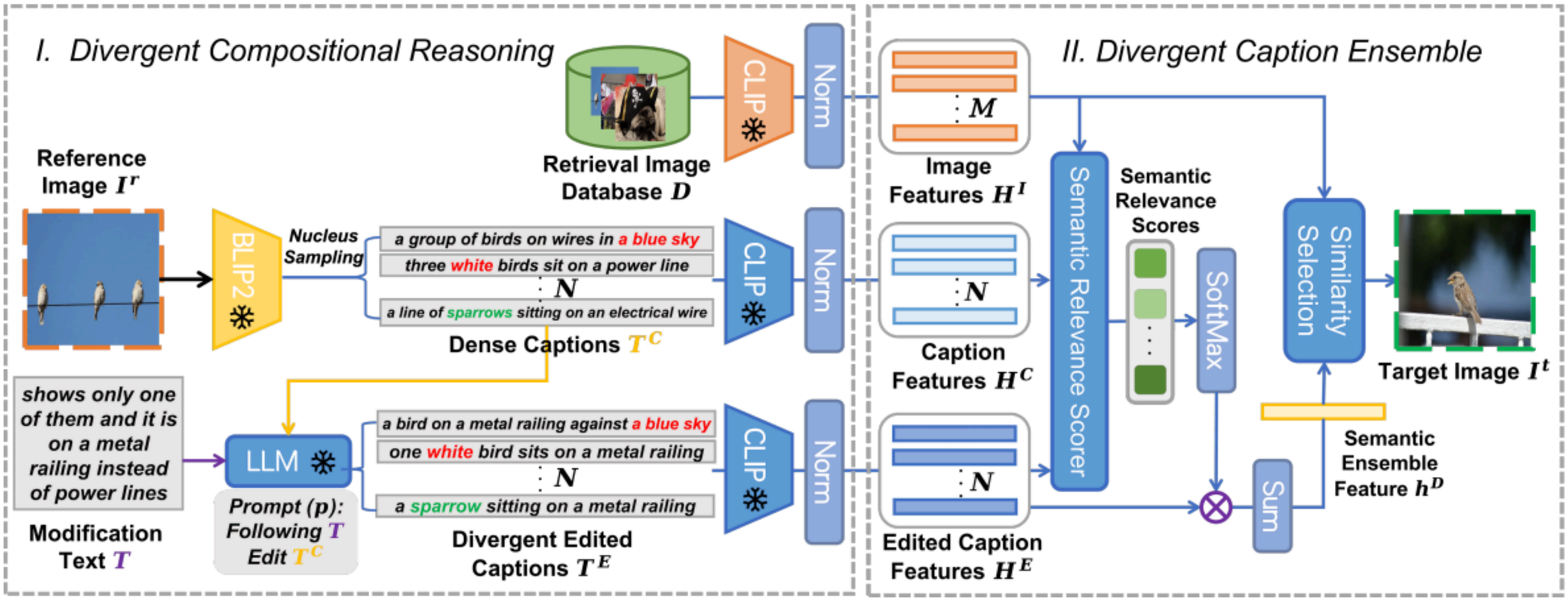
Diverse reasoning to
cover possible semantics.



LDRE: Advancing ZS-CIR with Diverse Captions and LLM Reasoning

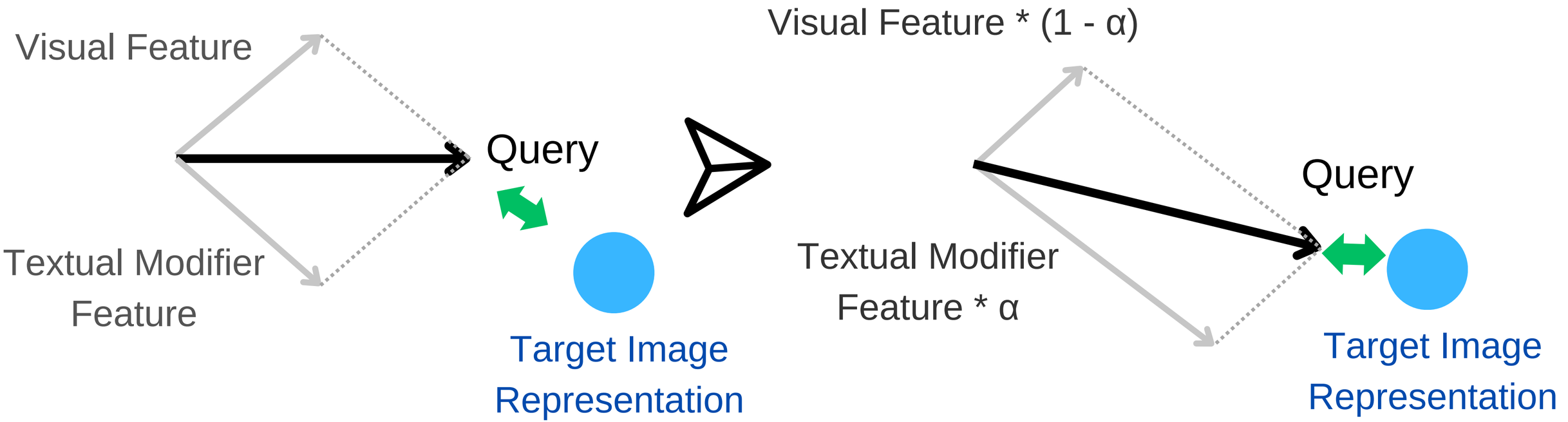
Diverse reasoning to cover possible semantics.

Ensemble integrates captions for robust retrieval.



Our idea to achieve training-free zero-shot CIR: WeiMoCIR

- Modality Fusion: Combines image and text features using weighted averages.
- Weighted Similarity: Balances query-to-image and query-to-caption relevance.



Our idea to achieve training-free zero-shot CIR: WeiMoCIR

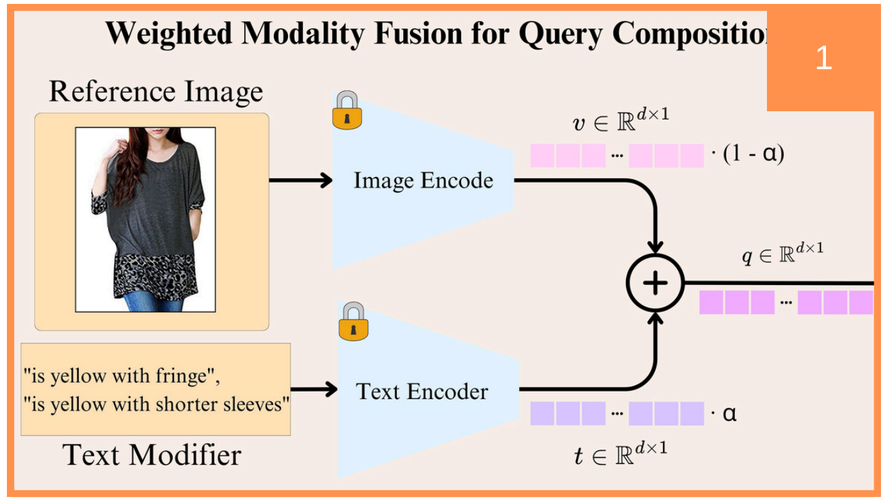
- Modality Fusion: Combines image and text features using weighted averages.
- Weighted Similarity: Balances query-to-image and query-to-caption relevance.
- Enhanced Representations via MLLMs: The idea stems from the fact that image descriptions inherently contain potential pseudo-queries.

Prompt

MLLM

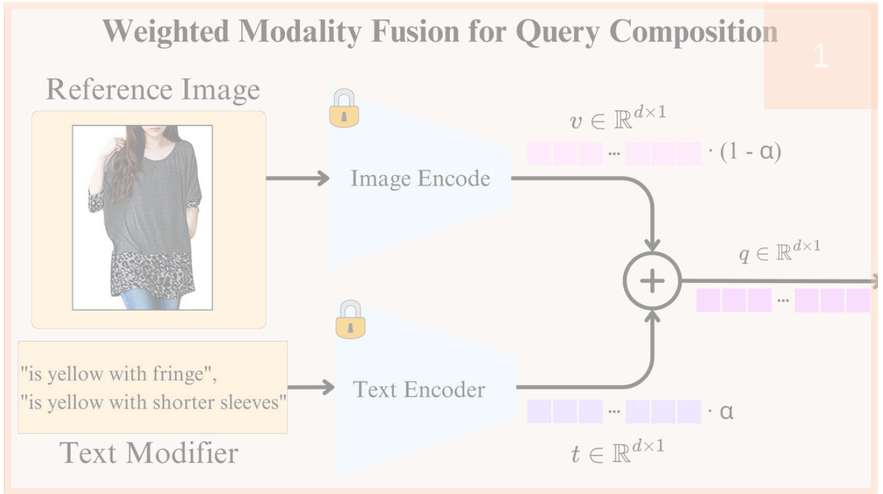
"Yellow short-sleeve t-shirt with fringe at the bottom.",
"V-neck t-shirt with yellow and gray color block design.",
"Yellow t-shirt with a raw hem and distressed detailing."

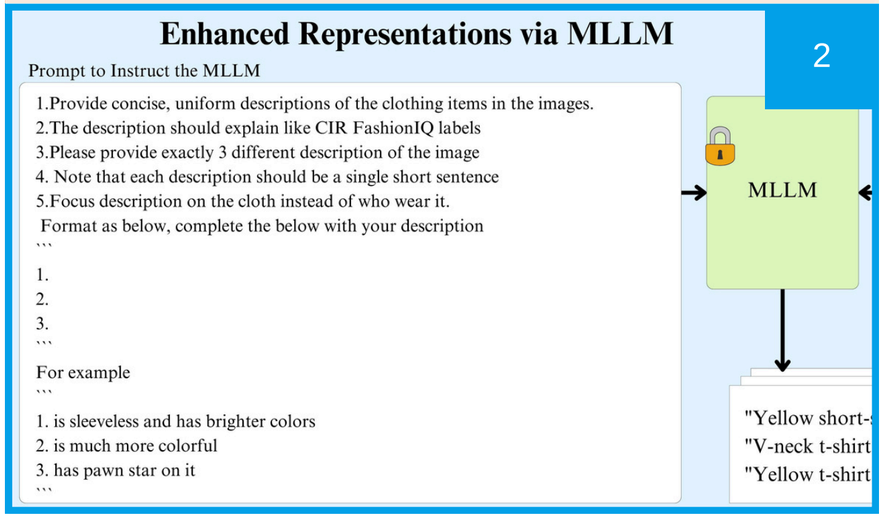
WeiMoCIR pipeline - OverView



1 Weighted Modality Fusion (Query Representation)

WeiMoCIR pipeline - OverView

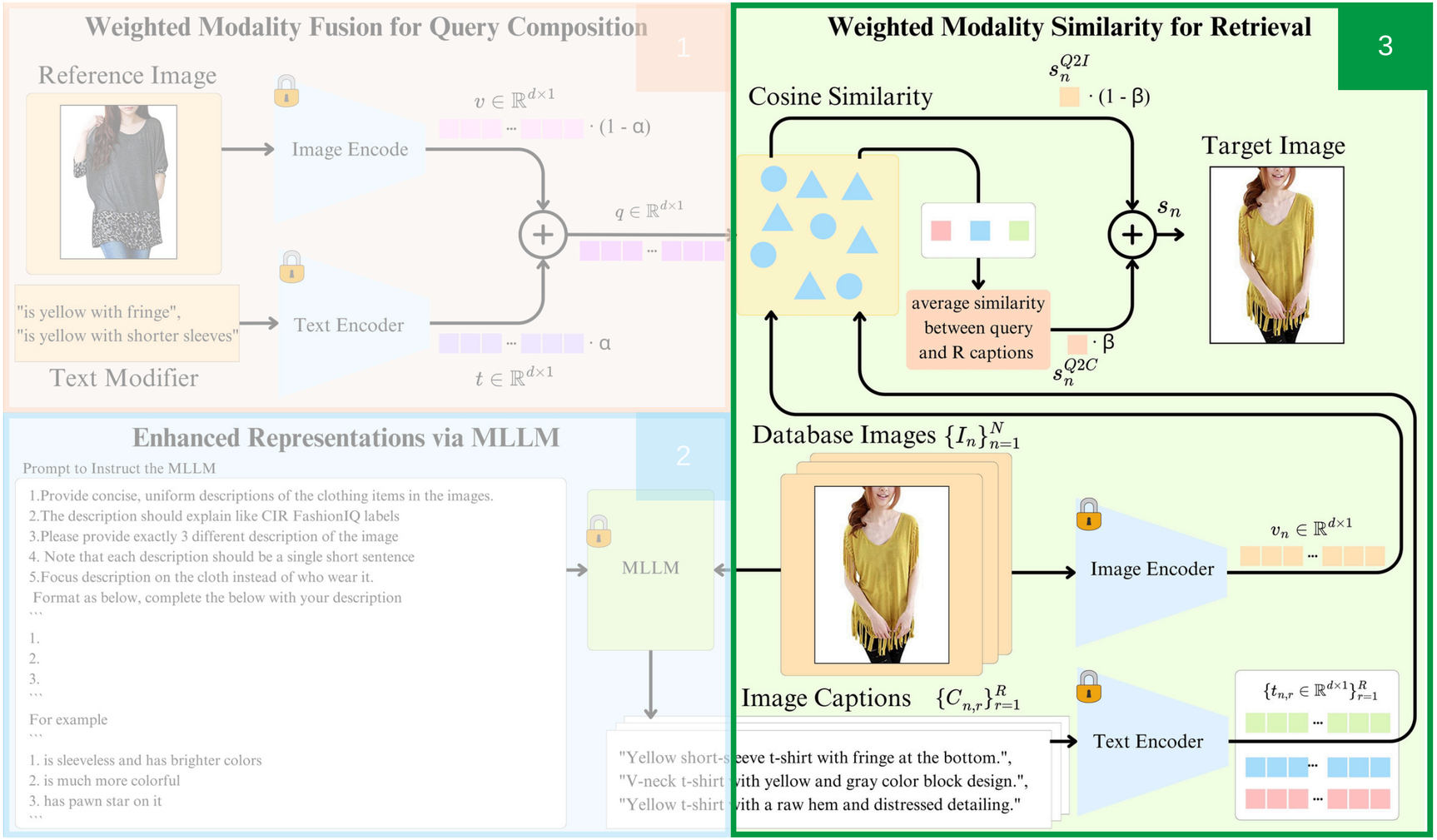




1 Weighted Modality Fusion (Query Representation)

2 Enhanced Representations via MLLMs (Database Representation)

WeiMoCIR pipeline - OverView

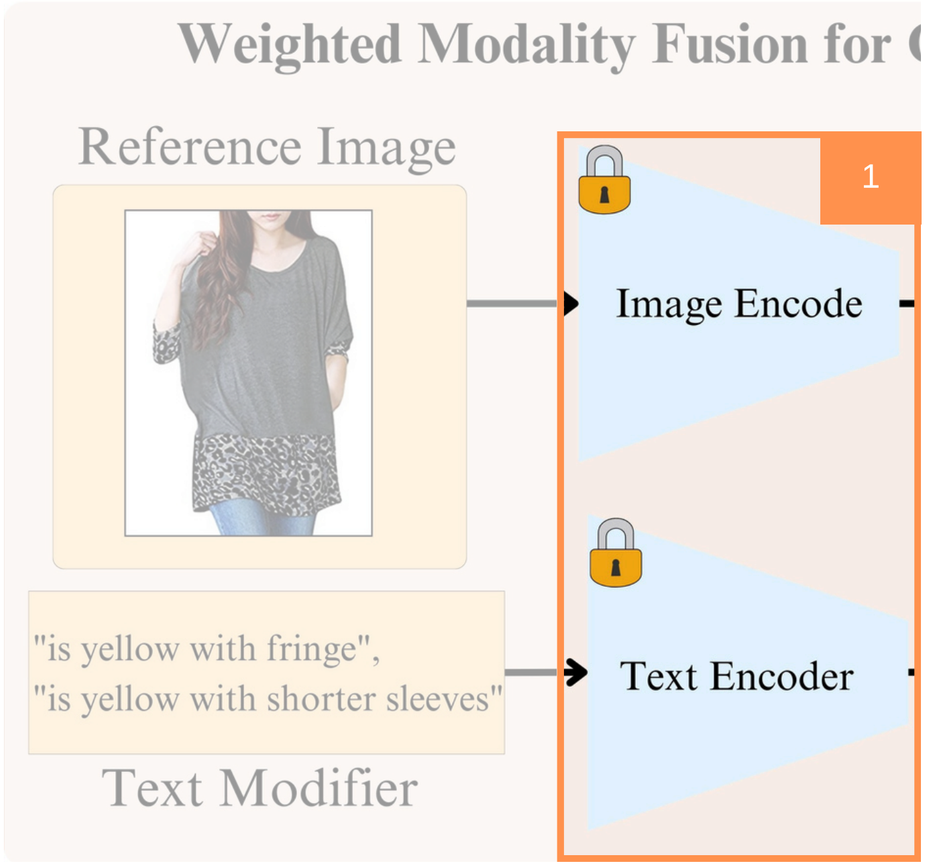


1 **Weighted Modality Fusion** (Query Representation)

2 **Enhanced Representations via MLLMs** (Database Representation)

3 **Weighted Modality Similarity for retrieval** (Similarity Computation)

Weighted Modality Fusion



1

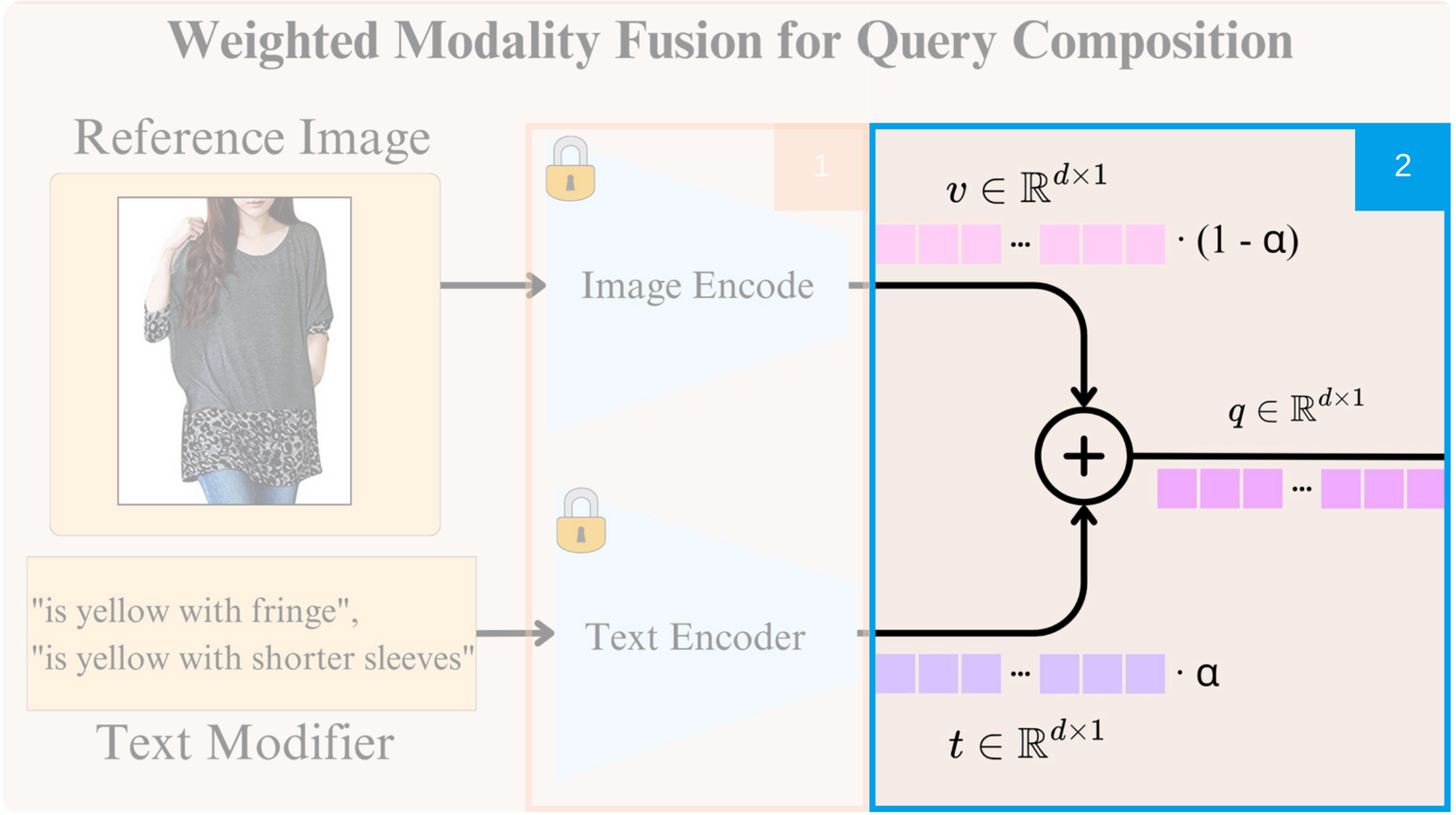
Replaceable VLM

Vision encoder and text encoder are used as feature extractors.

Visual Feature $v_n = f_\theta(I_n) \in \mathbb{R}^{d\times 1}$

Text Feature $t_{n,r} = f_\phi(C_{n,r}) \in \mathbb{R}^{d\times 1}$

Weighted Modality Fusion



1

Replaceable VLM

Vision encoder and text encoder are used as feature extractors.

Visual Feature $v_n = f_\theta(I_n) \in \mathbb{R}^{d\times1}$

Text Feature $t_{n,r} = f_\phi(C_{n,r}) \in \mathbb{R}^{d\times1}$

2

Combines image and text into a unified query

$$q = \mathcal{M}(v,t) = (1-\alpha)\cdot v + \alpha \cdot t$$

α: Balances contributions of visual and textual features.

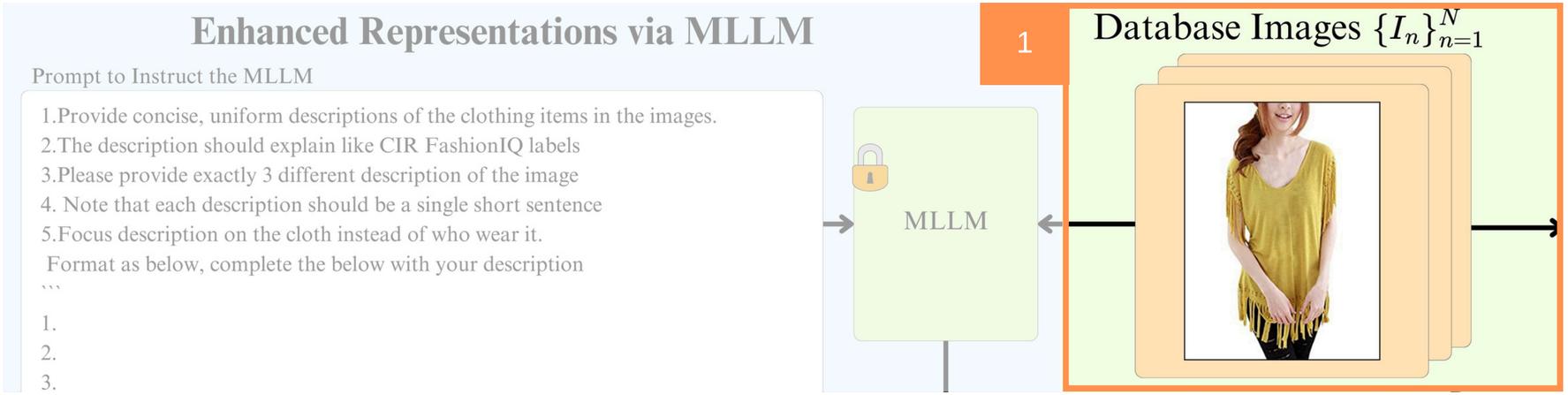
Enhanced Representations via MLLMs

- Generate captions for database images using MLLMs
- Captions provide semantic context for image features.

1

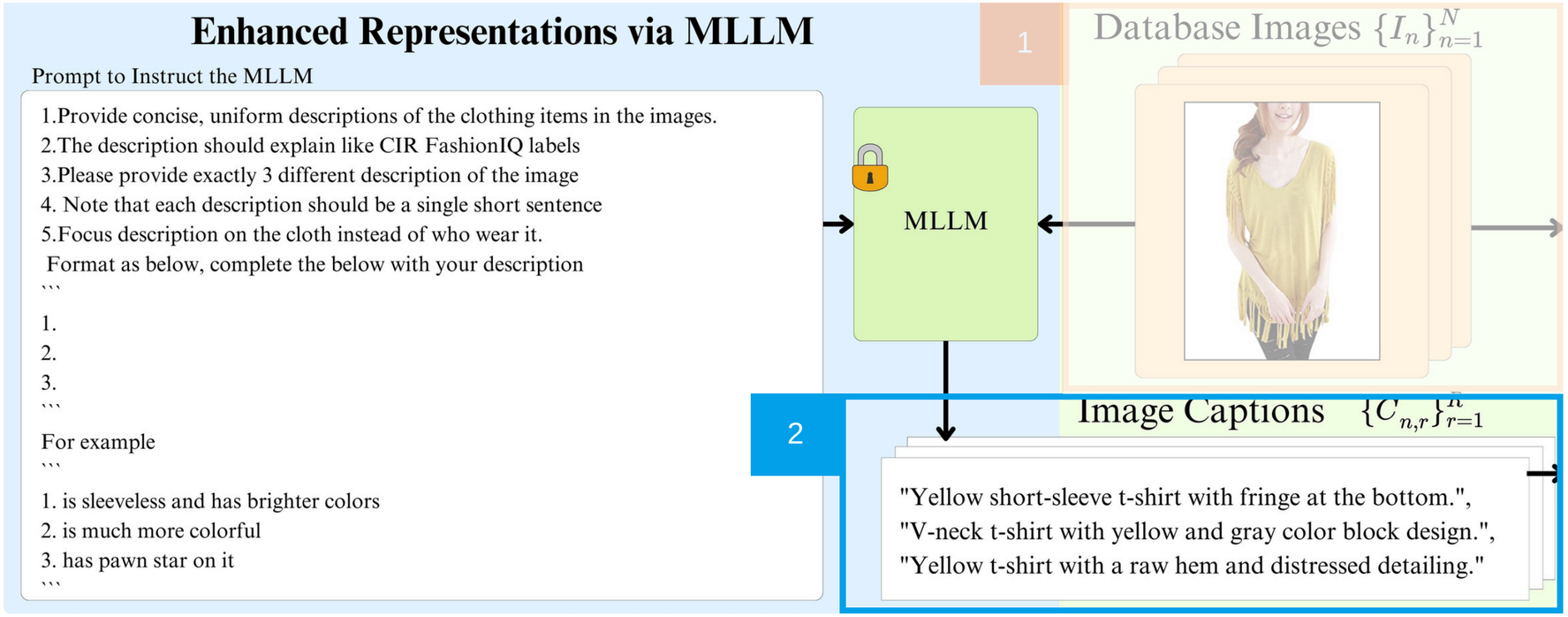
$$\{I_n\}_{n=1}^{N}$$

N pieces of
Database Images



Enhanced Representations via MLLMs

- Generate multiple captions for each database image using MLLMs (In this experience we set R = 3)
- Captions provide semantic context for image features.



1

$$\{I_n\}_{n=1}^{N}$$

N pieces of Database Images

MLLM

2

$$\{C_{n,r}\}_{r=1}^{R}$$

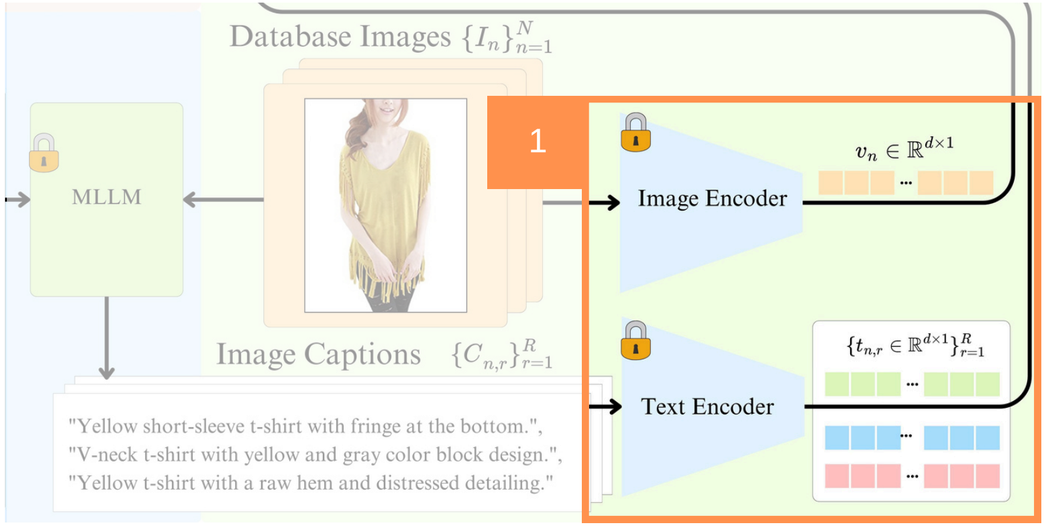
A set of R captions for N pieces of Database Images

Weighted Modality Similarity for retrieval

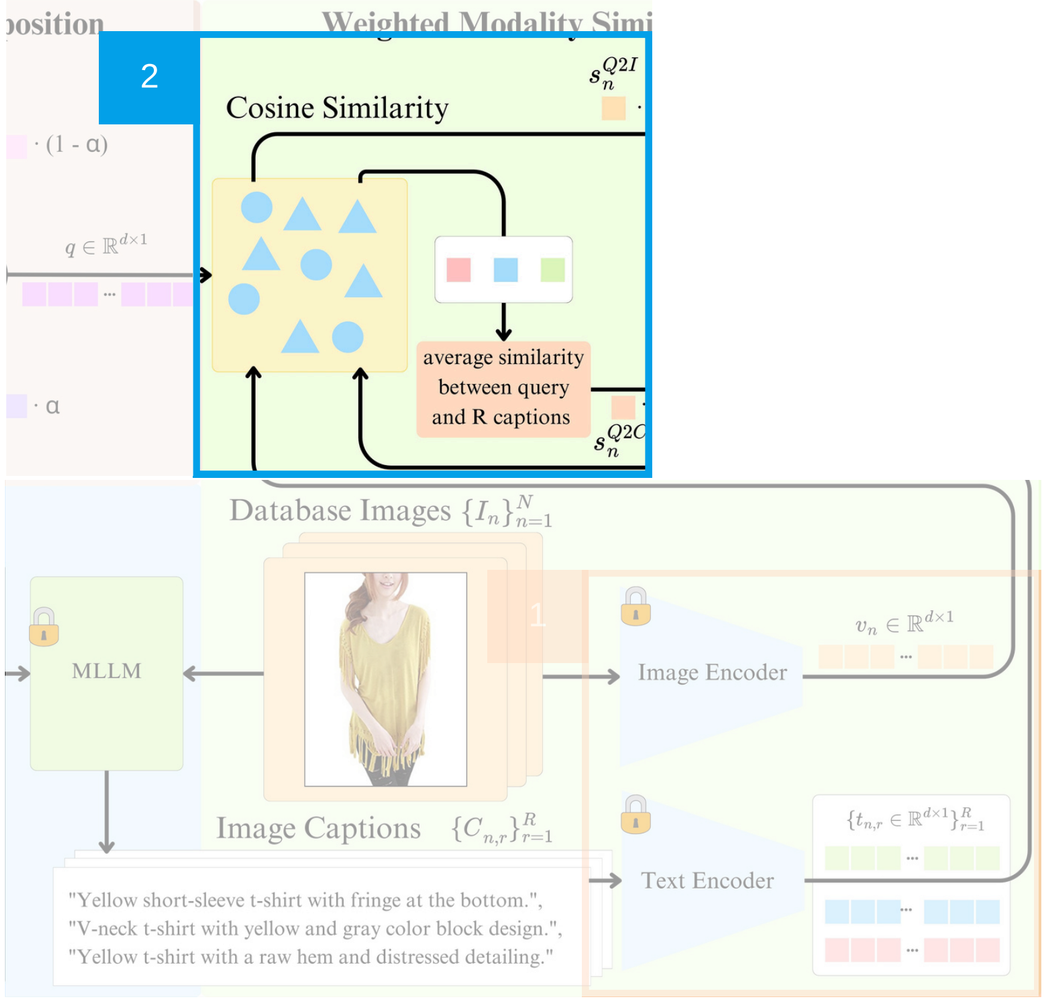
1

Feature Extraction

- Visual Feature $v_n = f_\theta(I_n) \in \mathbb{R}^{d\times 1}$
- Text Features. $t_{n,r} = f_\phi(C_{n,r}) \in \mathbb{R}^{d\times 1}$



Weighted Modality Similarity for retrieval



1

Feature Extraction

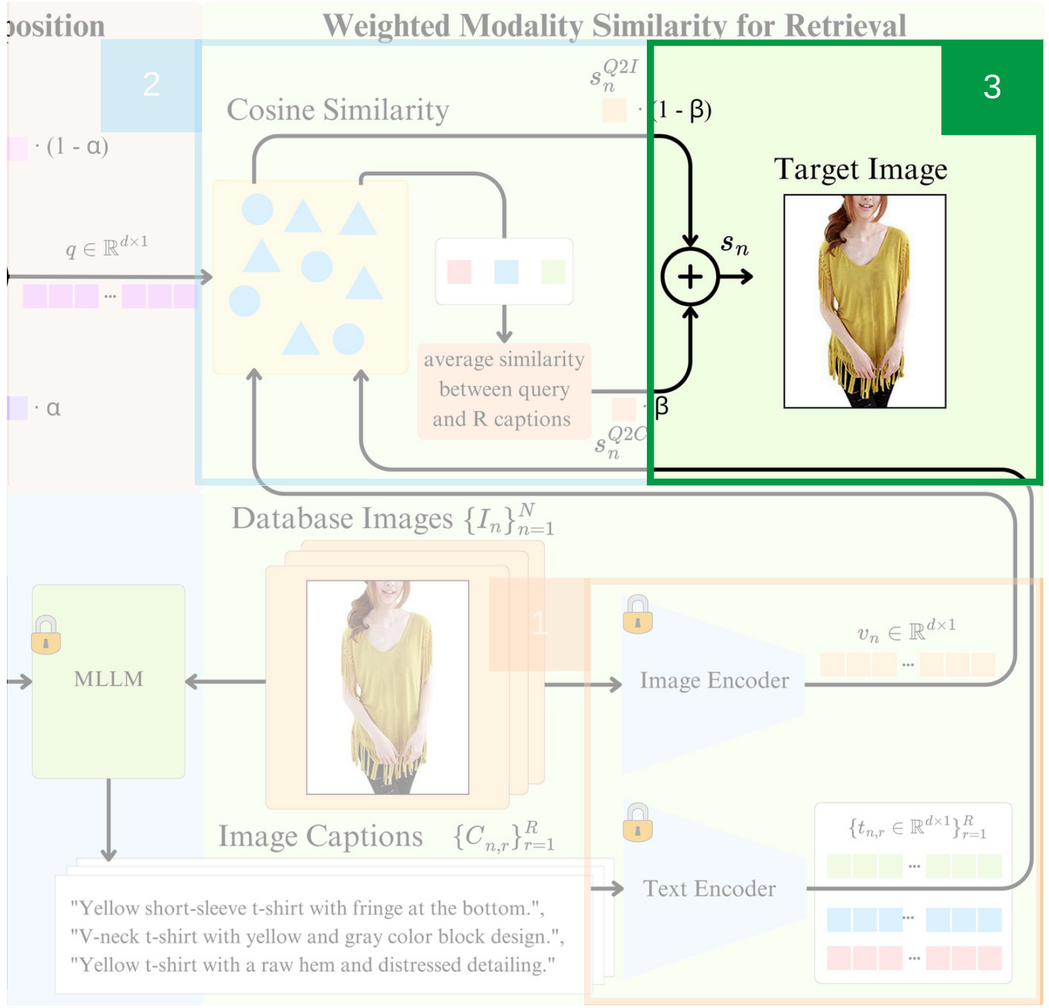
- Visual Feature $v_n = f_\theta(I_n) \in \mathbb{R}^{d\times1}$
- Text Features. $t_{n,r} = f_\phi(C_{n,r}) \in \mathbb{R}^{d\times1}$

2

Combines

- Query-to-Image $s_n^{Q2I} = \text{sim}(q, v_n)$
- Query-to-Captions $s_n^{Q2C} = \frac{1}{R}\sum_{r=1}^{R}\text{sim}(q, t_{n,r})$

Weighted Modality Similarity for retrieval



1

Feature Extraction

- Visual Feature $v_n = f_\theta(I_n) \in \mathbb{R}^{d\times1}$
- Text Features. $t_{n,r} = f_\phi(C_{n,r}) \in \mathbb{R}^{d\times1}$

2

Combines

- Query-to-Image $s_n^{Q2I} = \mathrm{sim}(q, v_n)$
- Query-to-Captions $s_n^{Q2C} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{sim}(q, t_{n,r})$

3

Weighted Similarity Score

$$s_n = (1-\beta)\cdot s_n^{Q2I} + \beta \cdot s_n^{Q2C}$$

Introduction to dataset

- FashionIQ: Focuses on fashion-related composed image retrieval.

Dataset	Domain
FashionIQ	Fashion images
CIRR	Real-world data

Introduction to dataset

- FashionIQ: Focuses on fashion-related composed image retrieval.
- CIRR (Composed Image Retrieval on Real-life images): Real-world, complex descriptors for composed retrieval.

Dataset	Domain
FashionIQ	Fashion images
CIRR	Real-world data

Evaluation Metrics and Baseline

Evaluation Metrics

- Recall@K: Measures how often the correct result is in the top-K retrieved items.

Dataset	Domain	Metric
FashionIQ	Fashion images	R@10, R@50
CIRR	Real-world data	R@1, R@5, R@10, R@50, Rsubset@1, Rsubset@2, Rsubset@3

Evaluation Metrics and Baseline

Evaluation Metrics

- Recall@K: Measures how often the correct result is in the top-K retrieved items.

Key Baselines

- Existing training-free method
- Existing Zero-Shot CIR methods
- Compare across different scale VLM

Dataset	Domain	Metric
FashionIQ	Fashion images	R@10, R@50
CIRR	Real-world data	R@1, R@5, R@10, R@50, Rsubset@1, Rsubset@2, Rsubset@3

Comparison on FashionIQ (ViT L/14, 428M params)

In the CLIP L/14, we even achieve better than the previous SOTA

Method	Traning-Free	Average R@10	Average R@50
Pic2Word	X	24.70	43.70
SEARLE		27.61	47.90
LinCIR		26.28	46.49
CIReVL	V	28.55	48.57
LDRE		28.51	50.54
WeiMoCIR (Ours)		31.54	50.99

Comparison on FashionIQ (ViT G/14, 1.37B params)

Although the SOTA achieve much better results, we still outperformance than other training free apoach

Method	Traning-Free	Average R@10	Average R@50
Pic2Word	X	31.28	51.89
SEARLE		34.81	55.71
LinCIR		45.11	65.69
CIReVL	V	32.19	52.36
LDRE		32.49	55.46
WeiMoCIR (Ours)		37.03	57.29

Comparison on CIRR (ViT L/14, 428M params)

We get great performance at the smaller model in CIRR

Method	Traning-Free	R@1	R@5	R@10	R@50
Pic2Word	X	23.90	51.70	65.30	87.80
SEARLE		24.87	52.31	66.29	88.58
LinCIR		25.04	53.25	66.68	-
CIReVL	V	24.55	52.31	64.92	86.34
LDRE		26.53	55.57	67.54	88.50
WeiMoCIR (Ours)		30.94	60.87	73.08	91.61

Comparison on CIRR (ViT G/14, 1.37B params)

But, we still have the room for improvement on G/14

Method	Backbone	Traning-Free	R@1	R@5	R@10	R@50
Pic2Word	ViT G/14	X	30.41	58.12	69.23	-
SEARLE	ViT G/14		34.80	64.07	75.11	-
LinCIR	ViT G/14		35.25	64.72	76.05	-
CIReVL	ViT G/14	V	34.65	64.29	75.06	91.66
LDRE	ViT G/14		36.15	66.39	77.25	93.95
WeiMoCIR (Ours)	ViT G/14		31.04	60.41	72.27	90.89

Comparison with the training-free methods (ViT B/32, 151M params)

Method	Dataset	Average R@10	Average R@50
CIReVL	FashionIQ	28.29	49.35
LDRE	FashionIQ	24.81	45.63
WeiMoCIR (Ours)	FashionIQ	29.86	49.82

Method	Dataset	R@1	R@5	R@10	R@50
CIReVL	CIRR	23.94	52.51	66.00	86.95
LDRE	CIRR	25.69	55.13	69.04	89.90
WeiMoCIR (Ours)	CIRR	26.31	57.69	70.36	91.01

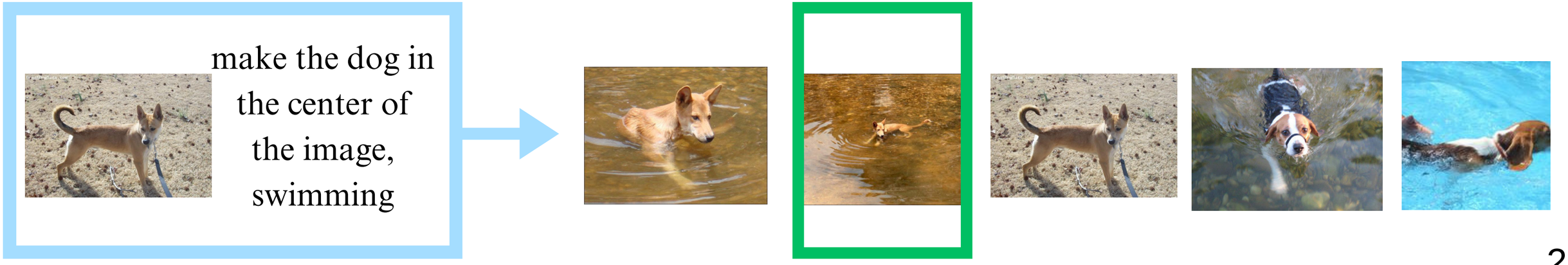
Qualitative results: Successful cases

- Text modifiers include color and style adjustments.

Qualitative results: Successful cases

- Text modifiers include color and style adjustments.

- Queries involve complex scene and action modifications.



Qualitative results: Failure cases

- Subtle Differences: Difficulty distinguishing fine-grained attributes.

Qualitative results: Failure cases

- Subtle Differences: Difficulty distinguishing fine-grained attributes.

Is more black with less graphics
It is black and has a simple print



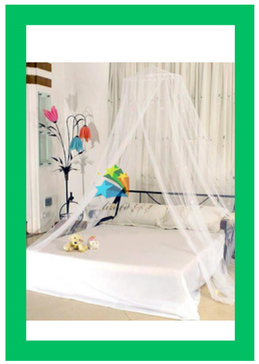



- Object Positioning: Struggles with relational queries.

The mattress is on the floor instead of on a bed





Impact of Different Pretrained VLMs

- Performance depends heavily on the alignment between VLM pretraining datasets and target CIR datasets.

CIRR			Recall@K				$Recall_{Subset}$@K		
Backbone	α	β	R@1	R@5	R@10	R@50	R@1	R@2	R@3
CLIP ViT L/14	0.80	0.10	30.94	60.87	73.08	91.61	58.55	79.06	90.07
CLIP ViT H/14	0.80	0.10	29.11	59.76	72.34	91.18	57.23	79.08	89.76
CLIP ViT G/14	0.80	0.10	31.04	60.41	72.27	90.89	58.84	78.92	89.64
BLIP w/ ViT-B	0.95	0.20	25.16	52.55	64.94	86.96	56.58	77.40	88.75
BLIP w/ ViT-B†	0.95	0.20	33.37	62.63	73.30	92.19	63.98	82.46	91.81
BLIP w/ ViT-B‡	0.95	0.20	**36.51**	**66.75**	**77.88**	93.45	**65.06**	**82.63**	**92.60**
BLIP w/ ViT-L	0.95	0.20	24.46	53.04	66.70	88.99	50.92	73.78	86.65
BLIP w/ ViT-L†	0.95	0.20	30.94	61.64	73.49	92.60	57.88	78.53	90.02
BLIP w/ ViT-L‡	0.95	0.20	32.07	63.08	75.28	**93.49**	58.63	79.13	90.53

† Finetuned on Flickr30k ‡ Finetuned on COCO

Effects of MLLM-Generated Captions

- Using multiple captions in most cases yields better results compared to a single caption.
- Combining captions provides a more complete description of image content.

Captions ($t_{n,r}$)	CLIP R@10	CLIP R@50	BLIP R@10	BLIP R@50
$t_{n,1}$	31.44	50.82	22.48	40.21
$t_{n,2}$	30.77	50.18	21.33	39.48
$t_{n,3}$	30.84	49.96	22.07	38.80
$t_{n,1} \cup t_{n,2}$	31.84	50.94	22.79	40.83
$t_{n,2} \cup t_{n,3}$	31.05	50.12	22.39	40.15
$t_{n,1} \cup t_{n,3}$	**31.77**	50.98	**22.95**	40.50
$t_{n,1} \cup t_{n,2} \cup t_{n,3}$	31.54	**50.99**	22.93	**41.00**

Impact of the MLLM-generated captions

- We found out that the best caption quality is located at the 1st one, the performance will drop compared to 1st captions.

Captions ($t_{n,r}$)	CLIP R@10	CLIP R@50	BLIP R@10	BLIP R@50
$t_{n,1}$	31.44	50.82	22.48	40.21
$t_{n,2}$	30.77	50.18	21.33	39.48
$t_{n,3}$	30.84	49.96	22.07	38.80
$t_{n,1} \cup t_{n,2}$	31.84	50.94	22.79	40.83
$t_{n,2} \cup t_{n,3}$	31.05	50.12	22.39	40.15
$t_{n,1} \cup t_{n,3}$	**31.77**	50.98	**22.95**	40.50
$t_{n,1} \cup t_{n,2} \cup t_{n,3}$	31.54	**50.99**	22.93	**41.00**

Conclusion

Key Contributions	Fusion	Combines image and text features for flexible query representation.
	Captions	MLLMs enrich database images with diverse semantic descriptions.
	Similarity	Balances query-to-image and query-to-caption matches for robust retrieval.

Conclusion

Key Contributions	Fusion	Combines image and text features for flexible query representation.
	Captions	MLLMs enrich database images with diverse semantic descriptions.
	Similarity	Balances query-to-image and query-to-caption matches for robust retrieval.
Achieves		• Achieves ZS-CIR without additional training. • Simplifies workflows while delivering competitive results. • Adaptable to diverse tasks and datasets.

Code is available at
https://github.com/whats2000/WeiMoCIR

Future Work

Observation

- Performance at rank-50 is highly effective, indicating the method retrieves relevant candidates at a broader scale.
- However, refining results at the top-rank (eg. 1, 5) still requires improvement.

Proposed Solution

- Reranker to refine

Future Work

Observation

- Performance at rank-50 is highly effective, indicating the method retrieves relevant candidates at a broader scale.
- However, refining results at the top-rank (eg. 1, 5) still requires improvement.

Proposed Solution

- Reranker to refine

Observation

- Incorporating textual descriptions improves retrieval performance by providing semantic context.
- Current captions are free-form and lack structured representation.

Proposed Solution

- Structured descriptions

Outline

Issue

- CIR Problem: Requires labor-intensive triplet collection.
- ZS-CIR Challenge: Most solutions still need pretraining on large datasets.
- Need: Training-free, scalable with model evolve.

Movtivation

- Avoid resource-intensive pretraining.
- Improve text-image alignment.
- Simplify and Efficiency retrieval with pre-trained models.

Method

- Fusion: Weighted text-image combination.
- Enhancement: Captions via MLLMs.
- Similarity: Combines query-to-image and query-to-caption.

Contribution

- Linear fusion of image and text.
- Introduces multimodal enhancement for retrieval.
- A new training-free ZS-CIR method.

Result

- Matches or surpasses pre-trained methods, especially small scale VLM
- MLLMs improve similarity scoring.

Future Work

- Improve high-rank results with rerankers.
- Utilize structured captions for better relevance scoring.

Compare with other method

Method	Fusion Approach	Focus	Training-Free	Handling Modalities
Pic2Word	Maps image → pseudo words	Query	x	Text-only
CIReVL	Caption + Text Modifier	Query	v	Text-dominant, limited visual context
LDRE	Diverse captions for expansion	Query Expansion	v	Expands queries via textual reasoning
WeiMoCIR	Weighted image + text fusion	Database + Query	v	Balances visual and textual relevance

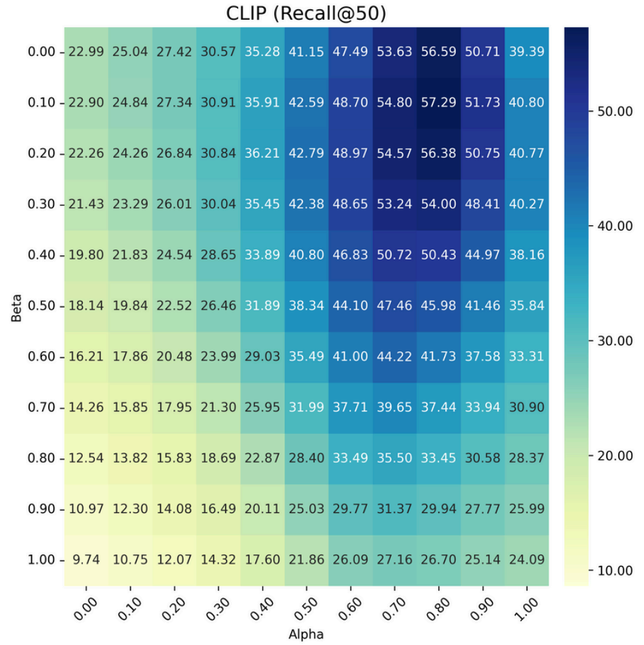
Effects of α and β on Performance

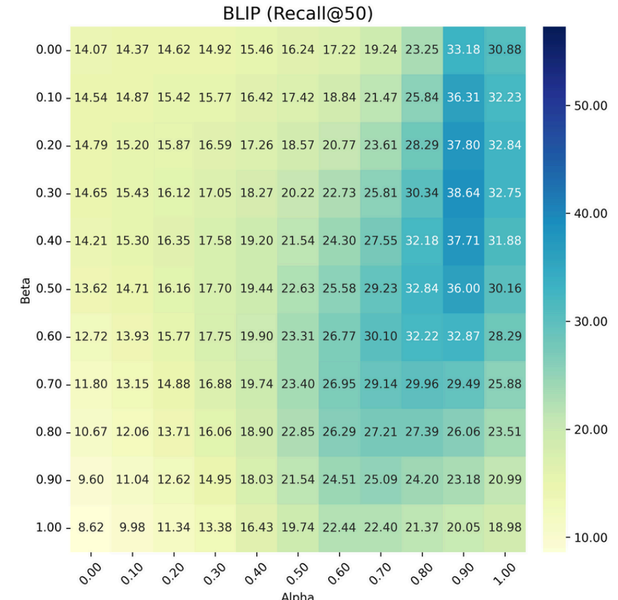
α (Query Modality Fusion)

- Balances visual and textual contributions.
- Higher α: More weight on text.
- Best performance observed with moderate to high α, emphasizing textual contributions.

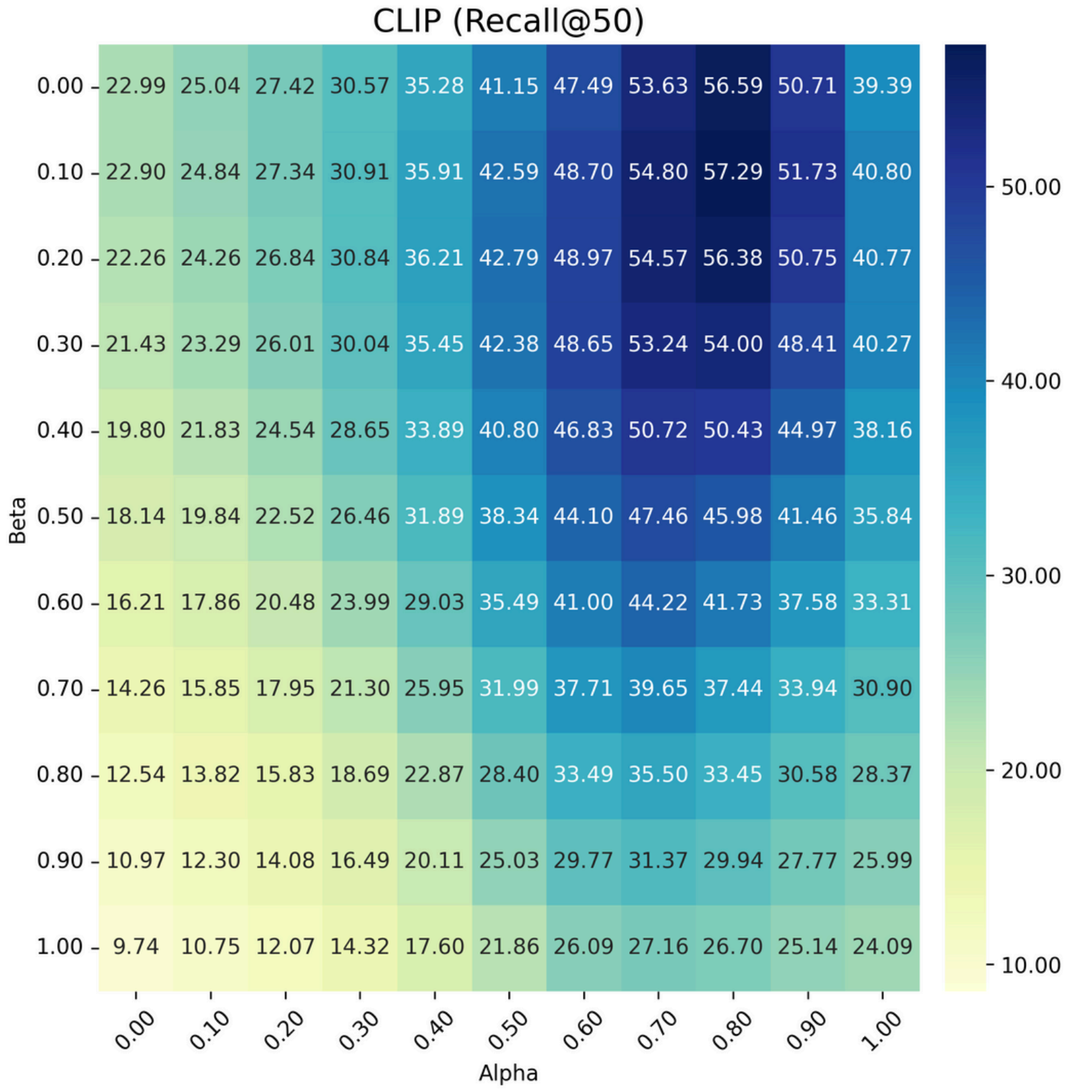
β (Similarity Weight)

- Balances query-to-image and query-to-caption similarities.
- Best performance was achieved with higher β, indicating the importance of caption-based information.

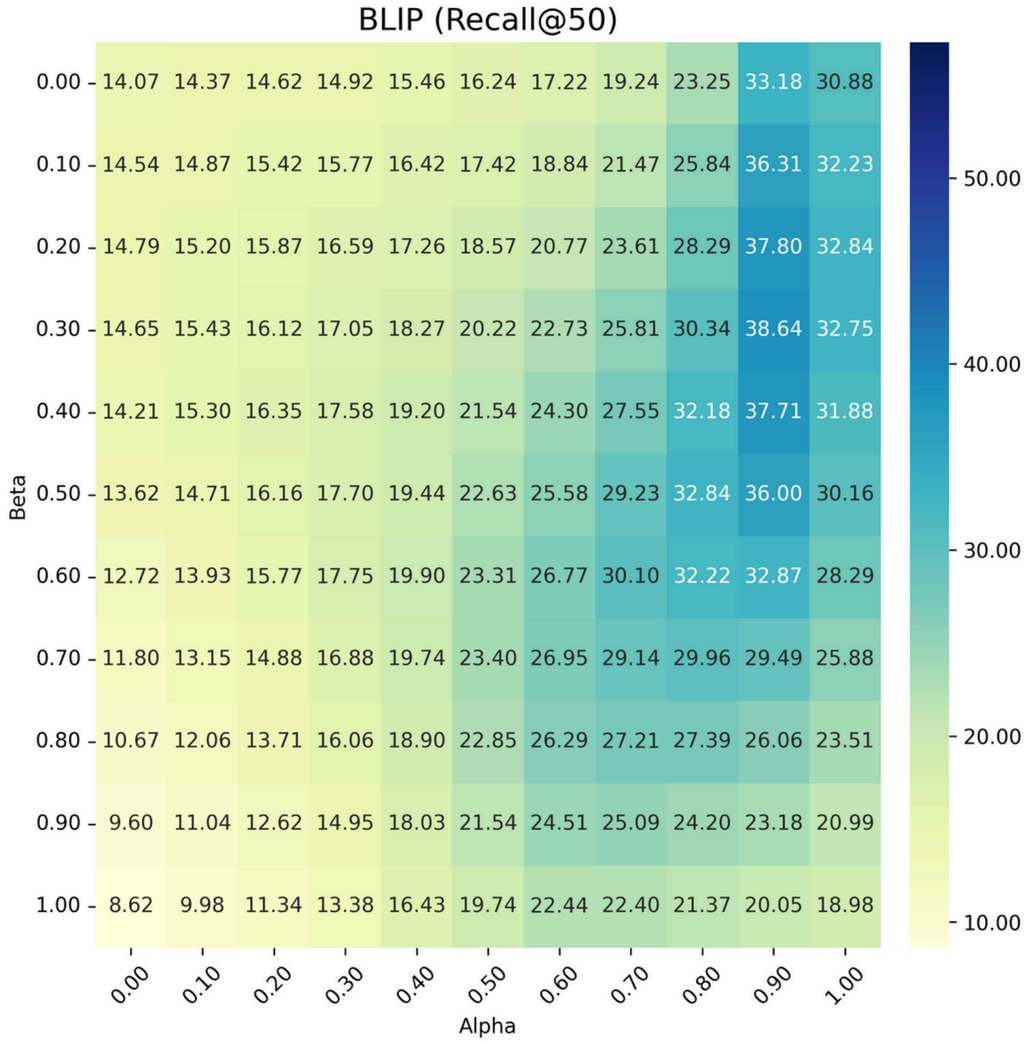




CLIP


CLIP (Recall@50)
Beta
Alpha
0.00 22.99 25.04 27.42 30.57 35.28 41.15 47.49 53.63 56.59 50.71 39.39
0.10 22.90 24.84 27.34 30.91 35.91 42.59 48.70 54.80 57.29 51.73 40.80
0.20 22.26 24.26 26.84 30.84 36.21 42.79 48.97 54.57 56.38 50.75 40.77
0.30 21.43 23.29 26.01 30.04 35.45 42.38 48.65 53.24 54.00 48.41 40.27
0.40 19.80 21.83 24.54 28.65 33.89 40.80 46.83 50.72 50.43 44.97 38.16
0.50 18.14 19.84 22.52 26.46 31.89 38.34 44.10 47.46 45.98 41.46 35.84
0.60 16.21 17.86 20.48 23.99 29.03 35.49 41.00 44.22 41.73 37.58 33.31
0.70 14.26 15.85 17.95 21.30 25.95 31.99 37.71 39.65 37.44 33.94 30.90
0.80 12.54 13.82 15.83 18.69 22.87 28.40 33.49 35.50 33.45 30.58 28.37
0.90 10.97 12.30 14.08 16.49 20.11 25.03 29.77 31.37 29.94 27.77 25.99
1.00 9.74 10.75 12.07 14.32 17.60 21.86 26.09 27.16 26.70 25.14 24.09
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
50.00
40.00
30.00
20.00
10.00

BLIP


BLIP (Recall@50)
Beta
Alpha
0.00 14.07 14.37 14.62 14.92 15.46 16.24 17.22 19.24 23.25 33.18 30.88
0.10 14.54 14.87 15.42 15.77 16.42 17.42 18.84 21.47 25.84 36.31 32.23
0.20 14.79 15.20 15.87 16.59 17.26 18.57 20.77 23.61 28.29 37.80 32.84
0.30 14.65 15.43 16.12 17.05 18.27 20.22 22.73 25.81 30.34 38.64 32.75
0.40 14.21 15.30 16.35 17.58 19.20 21.54 24.30 27.55 32.18 37.71 31.88
0.50 13.62 14.71 16.16 17.70 19.44 22.63 25.58 29.23 32.84 36.00 30.16
0.60 12.72 13.93 15.77 17.75 19.90 23.31 26.77 30.10 32.22 32.87 28.29
0.70 11.80 13.15 14.88 16.88 19.74 23.40 26.95 29.14 29.96 29.49 25.88
0.80 10.67 12.06 13.71 16.06 18.90 22.85 26.29 27.21 27.39 26.06 23.51
0.90 9.60 11.04 12.62 14.95 18.03 21.54 24.51 25.09 24.20 23.18 20.99
1.00 8.62 9.98 11.34 13.38 16.43 19.74 22.44 22.40 21.37 20.05 18.98
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
50.00
40.00
30.00
20.00
10.00