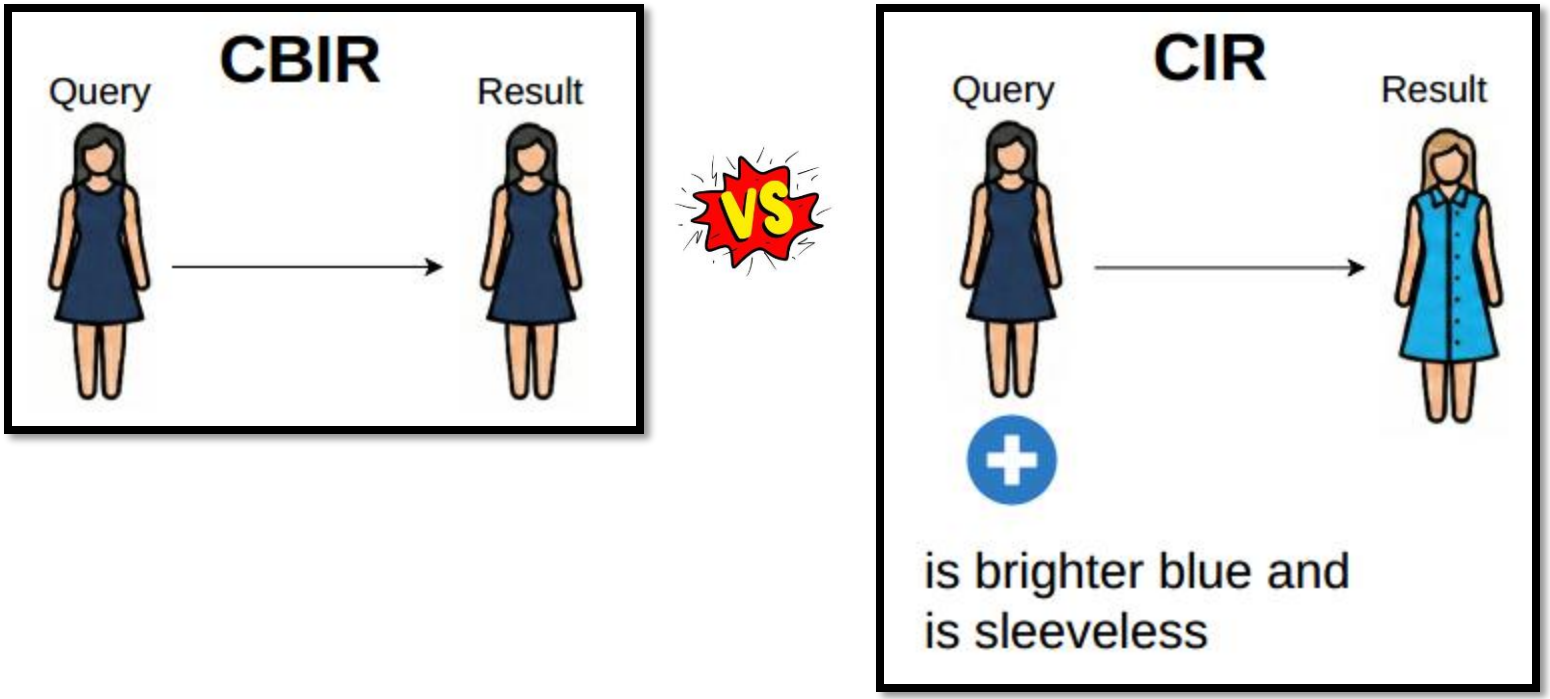
PDV: Prompt Directional Vectors for Zero-shot Composed Image Retrieval

Osman Tursun, Sinan Kalkan, Simon Denman, Clinton Fookes

Composed Image Retrieval (CIR)



CIR Vs Zero-shot CIR

Composed Image Retrieval (CIR)

- **Supervised Training**
Requires labeled triplets
(Reference + Text → Target)

- **Learned Fusion Module**
Trains a dedicated transformation

- **Data-Dependent**
Limited by annotation cost
(e.g., FashionIQ)

- **High In-Domain Accuracy**
Strong on seen categories

Zero-shot CIR

- **Training-Free**
No task-specific triplet data

- **Pretrained Vision-Language Model**
Uses models like CLIP[1] directly

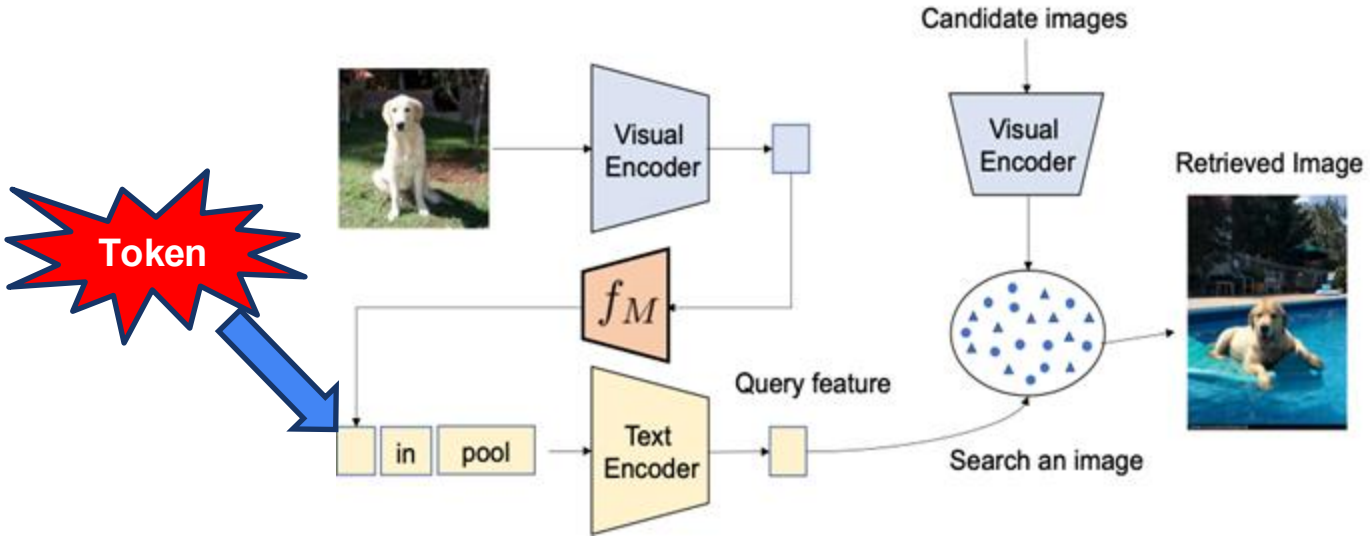
- **Highly Scalable**
Generalizes to open-domain queries

- **Robust Out-of-the-Box**
Good performance across diverse queries

[1] Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

Zero-shot Composed Image Retrieval

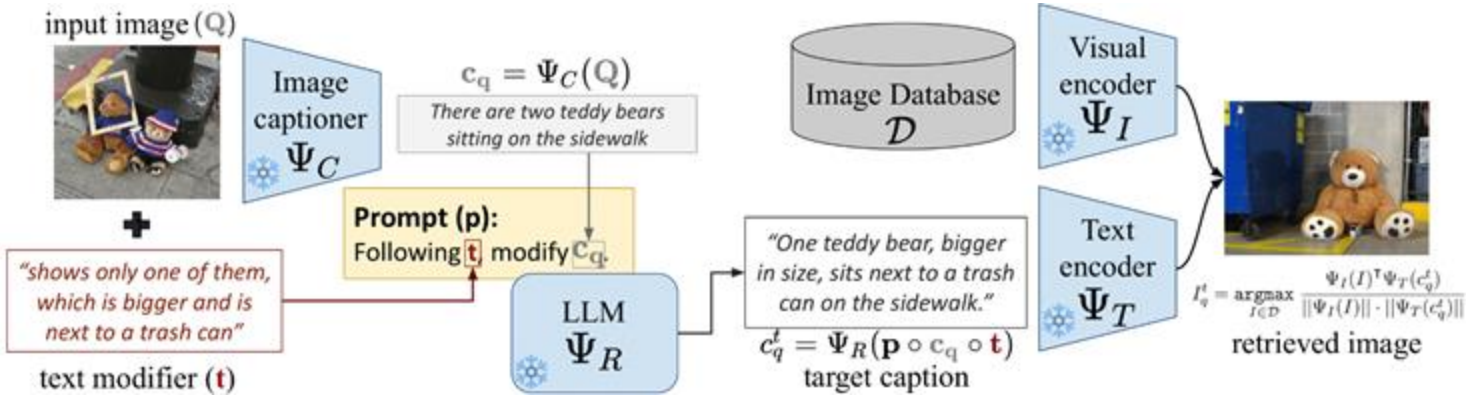
- *Token based approach e.g., Pic2Word [1], LinCIR [2]*



[1] Saito, K., et al., Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval, CVPR 2023
[2] Ge, Y., et al., Language-Only Training of Zero-Shot Composed Image Retrieval, CVPR 2024.

Zero-shot Composed Image Retrieval

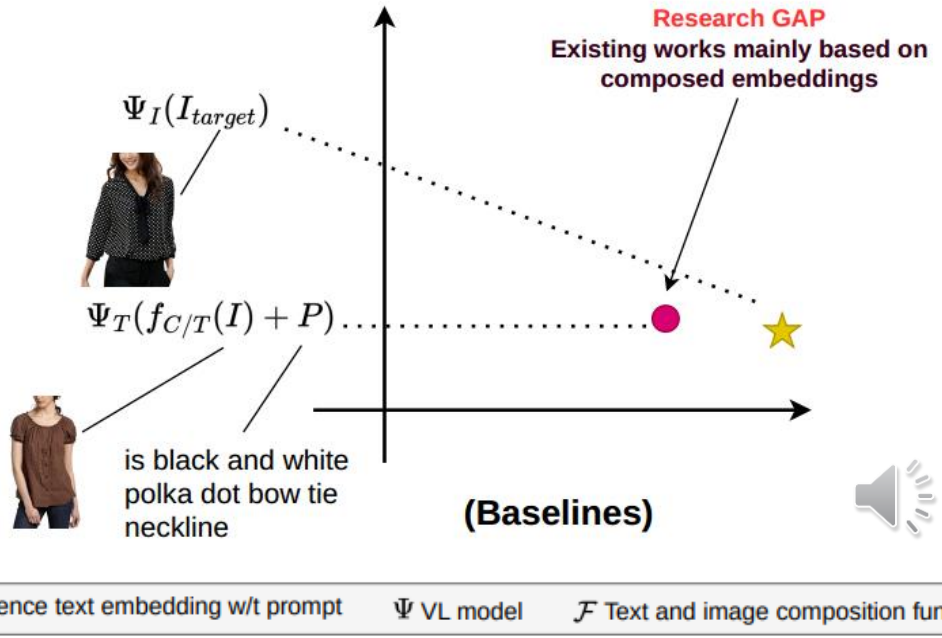
- Token based approach e.g., Pic2Word, LinCIR
- *Caption based approach e.g., CIReVL [1], SEIZE [2]*



[1] Karthik, S., et al., Vision-by-Language for Training-Free Compositional Image Retrieval, ICLR 2024
[2] Yang et al., Semantic Editing Increment Benefits Zero-Shot Composed Image Retrieval, ACM MM 2024

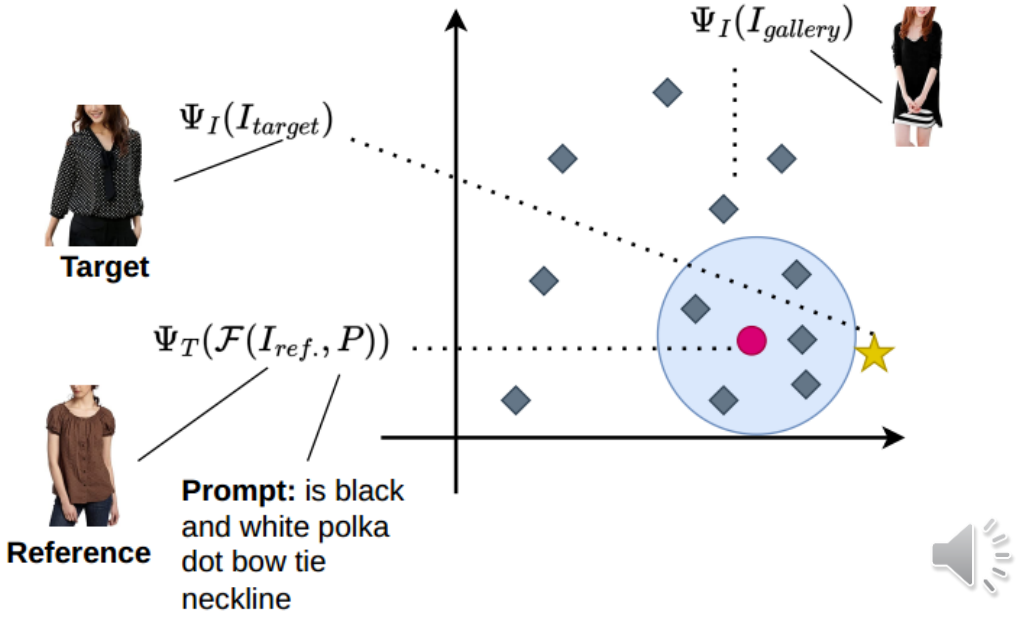
Research Gaps in ZS-CIR

- *Static query embedding representations.*



Research Gaps in ZS-CIR

- *Static query embedding representations.*



Research Gaps in ZS-CIR

- Static query embedding representations
- *Insufficient utilization of image embeddings*

Supervision	Methods	Dress		Shirt		TopTee		Average	
		R10	R50	R10	R50	R10	R50	R10	R50
ZERO-SHOT	Image-only	5.4	13.9	9.9	20.8	8.3	17.7	7.9	17.5
	Text-only	13.6	29.7	18.9	31.8	19.3	37.0	17.3	32.9
	Image+Text	16.3	33.6	21.0	34.5	22.2	39.0	19.8	35.7
	Pic2Word	**20.0**	**40.2**	**26.2**	**43.6**	**27.9**	**47.4**	**24.7**	**43.7**

Saito, K., et al., Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval, CVPR 2023

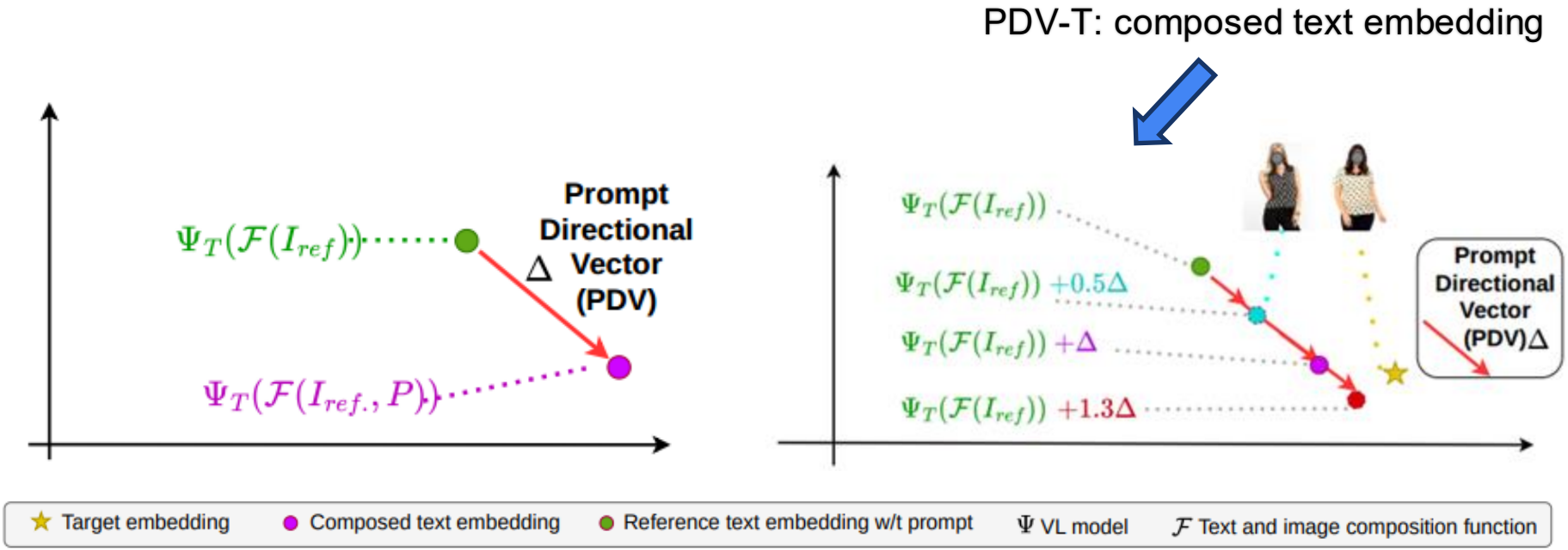
Research Gaps in ZS-CIR

- Static query embedding representations
- Insufficient utilization of image embeddings
- *Suboptimal fusion of text and image embeddings.*

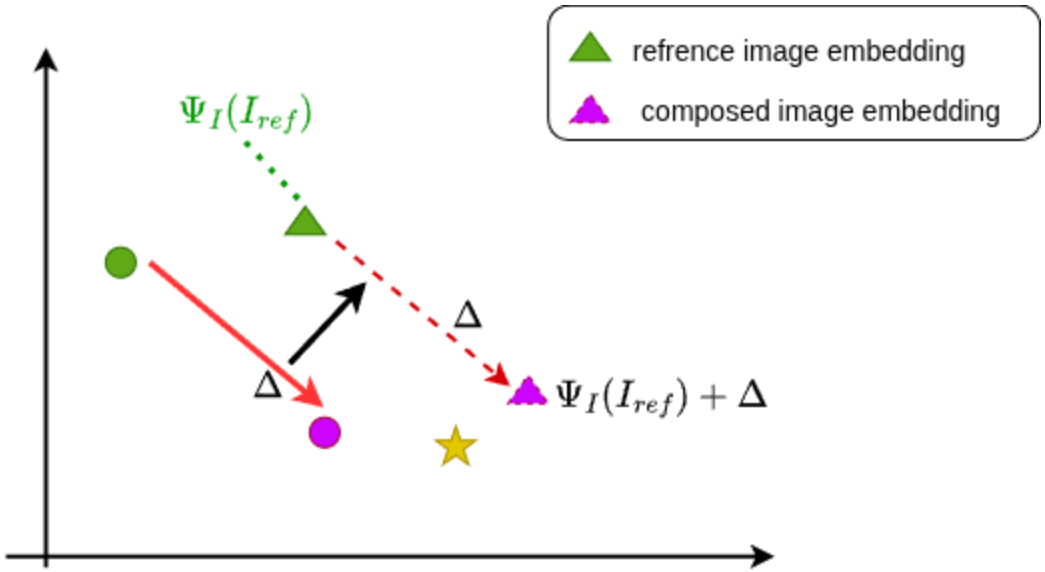
Supervision	Methods	Dress		Shirt		TopTee		Average	
		R10	R50	R10	R50	R10	R50	R10	R50
ZERO-SHOT	Image-only	5.4	13.9	9.9	20.8	8.3	17.7	7.9	17.5
	Text-only	13.6	29.7	18.9	31.8	19.3	37.0	17.3	32.9
	Image+Text	16.3	33.6	21.0	34.5	22.2	39.0	19.8	35.7
	Pic2Word	**20.0**	**40.2**	**26.2**	**43.6**	**27.9**	**47.4**	**24.7**	**43.7**

Saito, K., et al., Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval, CVPR 2023

PDV: Prompt Directional Vectors

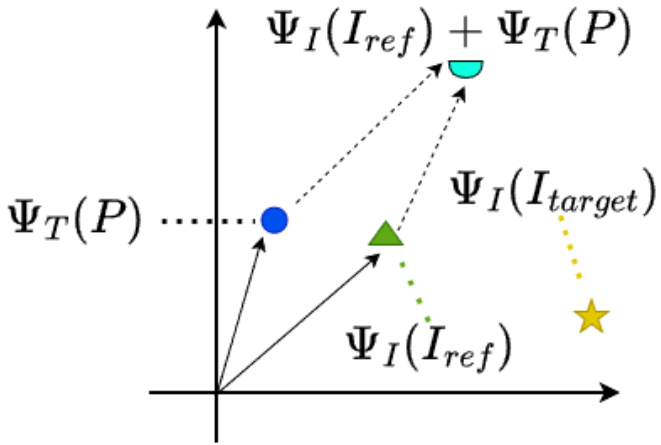


PDV-I: Composed Image Embedding

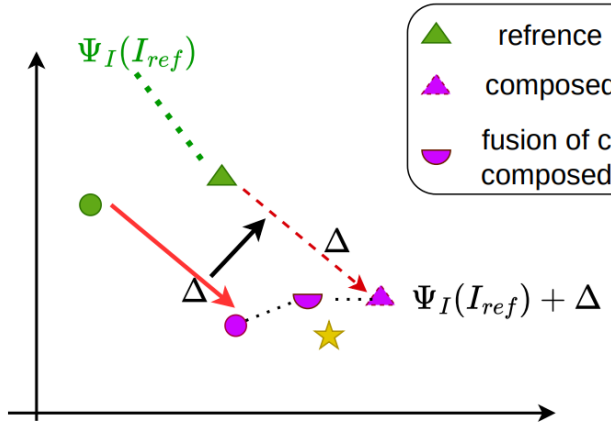


Target embedding | Composed text embedding | Reference text embedding w/t prompt | Ψ VL model | $\mathcal{F}$ Text and image composition function

PDV-F: Fusion of PDV-T and PDV-I



Navie Fusion



PDV-F

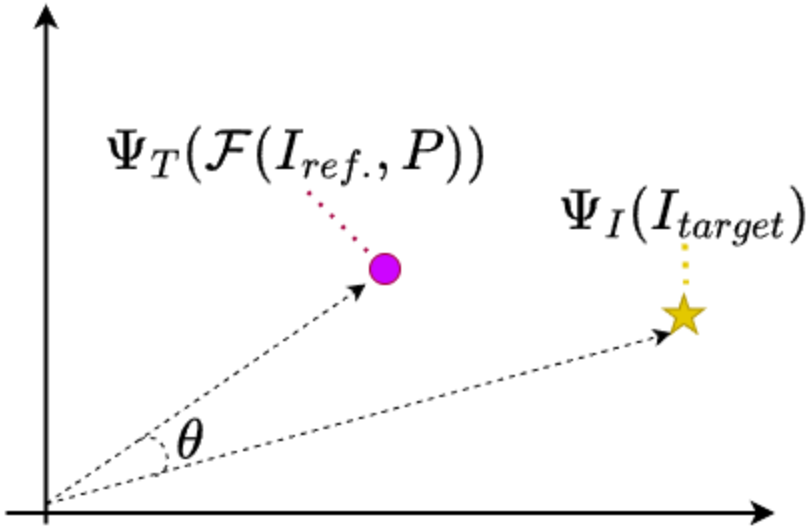
★ Target embedding ● Composed text embedding ● Reference text embedding w/t prompt Ψ VL model $\mathcal{F}$ Text and image composition function

PDV and Its Variants

Component	Definition	Special cases
PDV	$\Delta_{\text{PDV}} = \Psi_T(\mathcal{F}(I_{ref}, P)) - \Psi_T(\mathcal{F}(I_{ref}))$	—
PDV-T (text)	$\Phi_{\text{PDV-T}} = \Psi_T(\mathcal{F}(I_{ref})) + \alpha_T \Delta_{\text{PDV}}$	$\alpha_T = 1 \Rightarrow$ baseline ZS-CIR
PDV-I (image)	$\Phi_{\text{PDV-I}} = \Psi_I(I_{ref}) + \alpha_I \Delta_{\text{PDV}}$	$\alpha_I = 0 \Rightarrow$ image-only (CBIR)
PDV-F (fusion)	$\Phi_{\text{PDV-F}} = (1-\beta)\,\Phi_{\text{PDV-I}} + \beta\,\Phi_{\text{PDV-T}}$	β=0: PDV-I; β=1: PDV-T

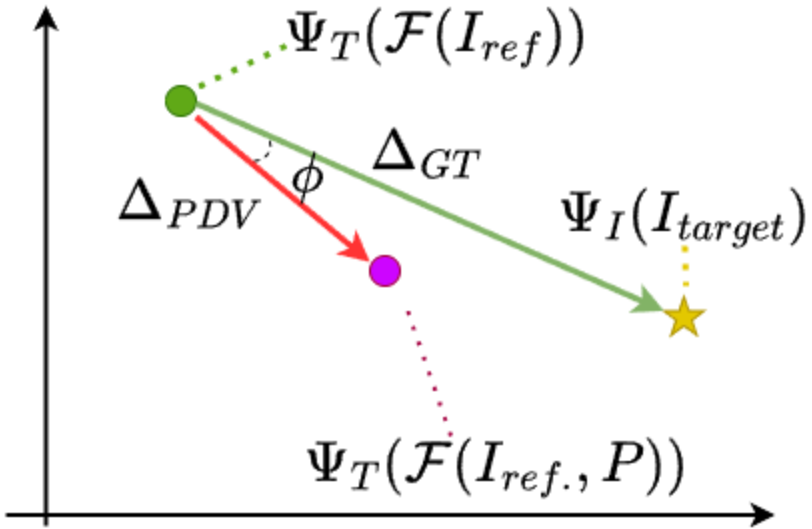
Table 1: Summary of PDV and its variants. α_T controls prompt strength (α_T>1 amplify, α_T<1 attenuate).

When and Why PDV Works



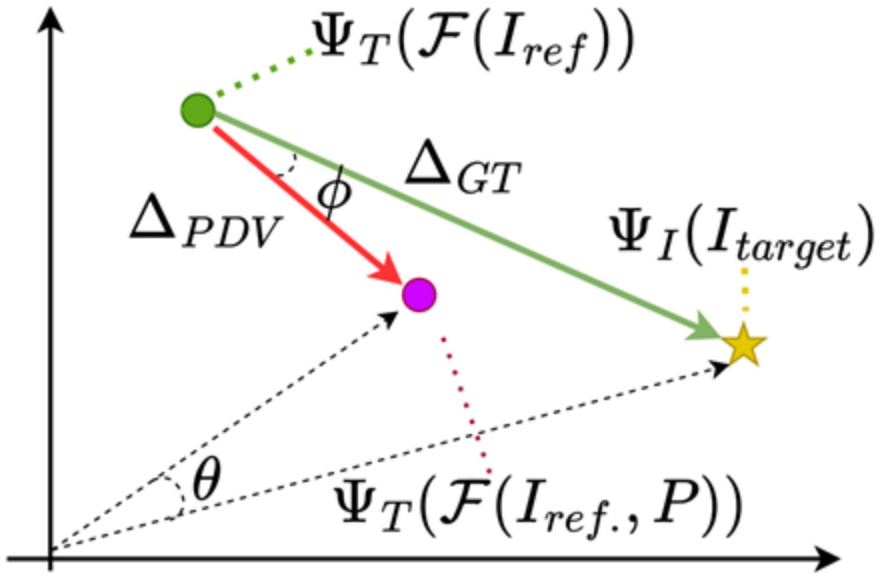
Target embedding | Composed text embedding | Reference text embedding w/t prompt | Ψ VL model | $\mathcal{F}$ Text and image composition function

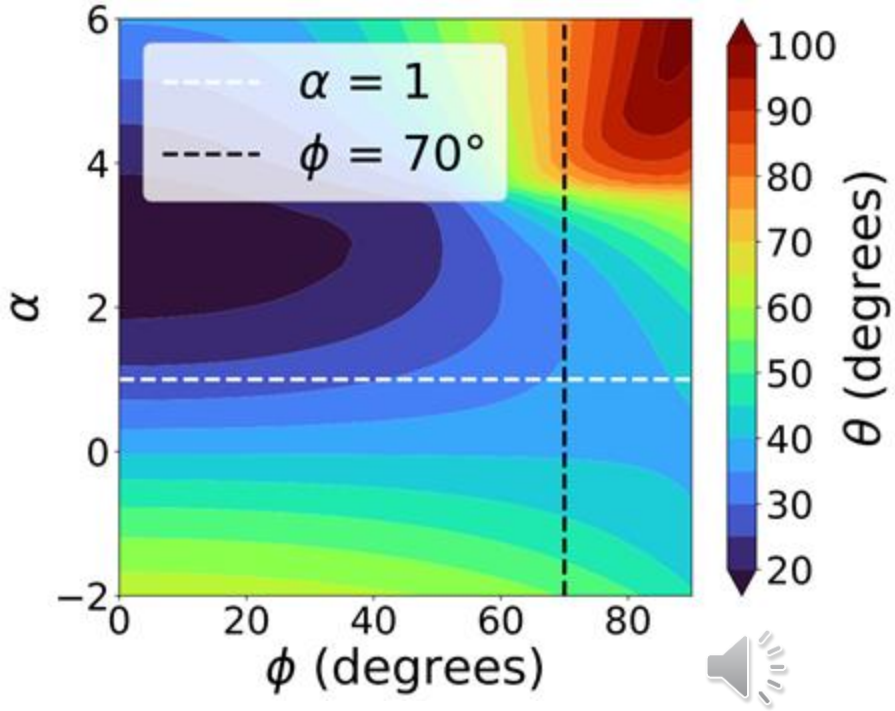
When and Why PDV Works



★ Target embedding ● Composed text embedding ● Reference text embedding w/t prompt Ψ VL model $\mathcal{F}$ Text and image composition function

When and Why PDV Works





★ Target embedding ● Composed text embedding ● Reference text embedding w/t prompt Ψ VL model $\mathcal{F}$ Text and image composition function

Experiment Setups

Baselines:

- Token-based methods: Pic2Word, SEARLE [1]
- Caption-based methods: CIReVL, LDRE [2]

Datasets:

- FashionIQ, CIRCO, CIRR

[1] Baldrati et al., Zero-Shot Composed Image Retrieval with Textual Inversion, ICCV 2023
[2] Yang et al. LDRE: LLM-based Divergent Reasoning and Ensemble for Zero-Shot Composed Image Retrieval, SIGIR 2024

Quantiative Results

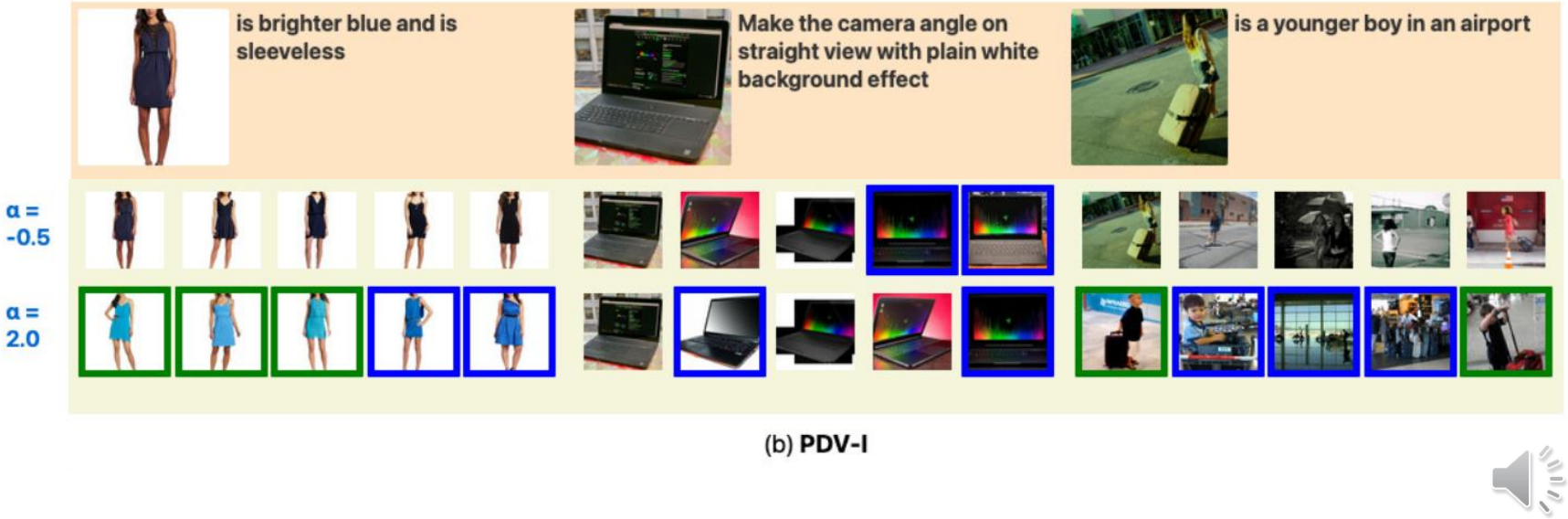
Arch	Baseline + PDV	Fashion-IQ (Avg)		CIRCO				CIRR					
		R@10	R@50	mAP@5	mAP@10	mAP@25	mAP@50	R@1	R@5	R@10	R@50	R_s@1	R_s@2
ViT B/32	SEARLE	+3.6	+2.7	+6.8	+5.6	+5.1	+4.7	+2.2	+0.5	+0.8	+0.0	+3.7	+1.9
	CIReVL	+23.1	+14.2	+33.2	+33.7	+33.2	+32.0	+38.9	+22.2	+13.9	+6.3	+9.4	+4.6
	LDRE	–	–	−0.1	+4.1	+4.5	+4.3	+14.1	+8.8	+5.4	+1.7	+4.9	+2.2
ViT L/14	Pic2Word	+7.2	+2.6	+13.7	+15.8	+14.8	+14.3	+0.9	+1.2	+1.5	+1.0	−0.8	−0.4
	SEARLE	+10.5	+6.2	+7.7	+6.6	+6.8	+6.3	+5.8	+2.2	+0.4	−0.3	+3.9	+2.0
	CIReVL	+32.0	+18.3	+38.2	+40.0	+37.9	+37.4	+47.6	+26.5	+18.6	+6.9	+14.3	+6.8
	LDRE	–	–	+13.0	+11.7	+11.4	+10.8	+13.0	+8.2	+6.5	+2.5	+5.4	+2.9
ViT G/14	CIReVL	+35.3	+20.2	+12.1	+14.0	+13.5	+13.1	+10.1	+5.7	+3.8	+1.2	+2.1	+0.6
	LDRE	–	–	+4.7	+6.1	+5.2	+5.1	+13.7	+7.9	+4.5	+1.4	+5.2	+3.2

Relative improvement (%) of PDV-F over baselines across all datasets and metrics.

Qualitative Results: PDV-T

(a) **PDV-T**

Qualitative Results: PDV-I



(b) **PDV-I**

Qualitative Results: PDV-F

Conclusion and Other Contributions

- PDV is an effective plug-and-play approach for ZS-CIR.
- PDV addresses three key limitations of existing methods: static query embeddings, insufficient utilization of image embeddings, and suboptimal image-text fusion.
- PDV further introduces auto-tuning of α and efficient iterative retrieval via gallery filtering.

For more information:

- Paper ID: 190.
- Arxiv: `https://arxiv.org/abs/2502.07215`
- Code: `https://github.com/neouyghur/PDV`