



UNIVERSITÀ  
di VERONA



FONDAZIONE  
BRUNO KESSLER

# Seeing the Abstract: Translating the Abstract Language for Vision Language Models

Davide Talon<sup>\*1</sup> Federico Girella<sup>\*2</sup> Ziyue Liu<sup>23</sup> Marco Cristani<sup>2</sup> Yiming Wang<sup>1</sup>  
<sup>1</sup>Fondazione Bruno Kessler <sup>2</sup>University of Verona <sup>3</sup>Polytechnic University of Turin



Politecnico  
di Torino



CODE/DEMO

## The Problem

Humans use both **abstract** and **concrete** words to express feelings, desires, and observable properties. However, Vision-Language Models (VLMs) are primarily pre-trained on concrete-oriented web-image texts, underrepresenting abstract concepts.

### Concrete-oriented

“A sleeveless top with crochet bodice and loose waist”



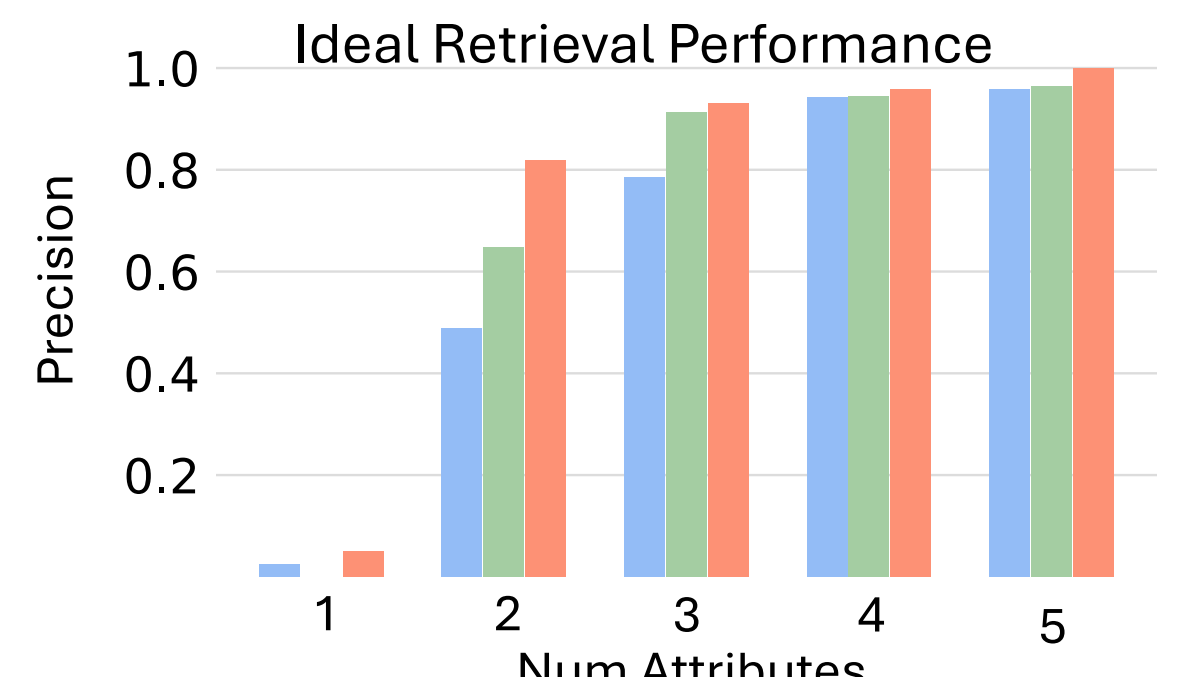
### Abstract-oriented

“Opt for something a little different, such as this **sexy** top! This sleeveless number features a stunning crochet bodice, followed by loose, **airy** fabric”

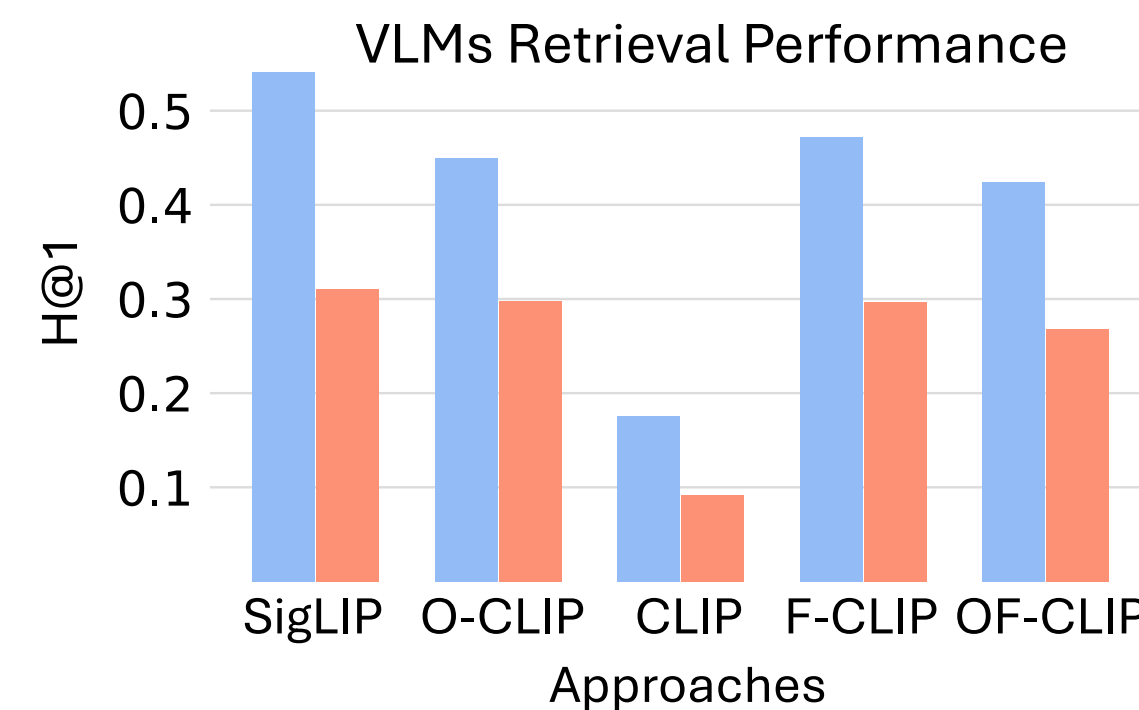


## The Concreteness Gap

Abstracts ideally allow for a more discriminative query

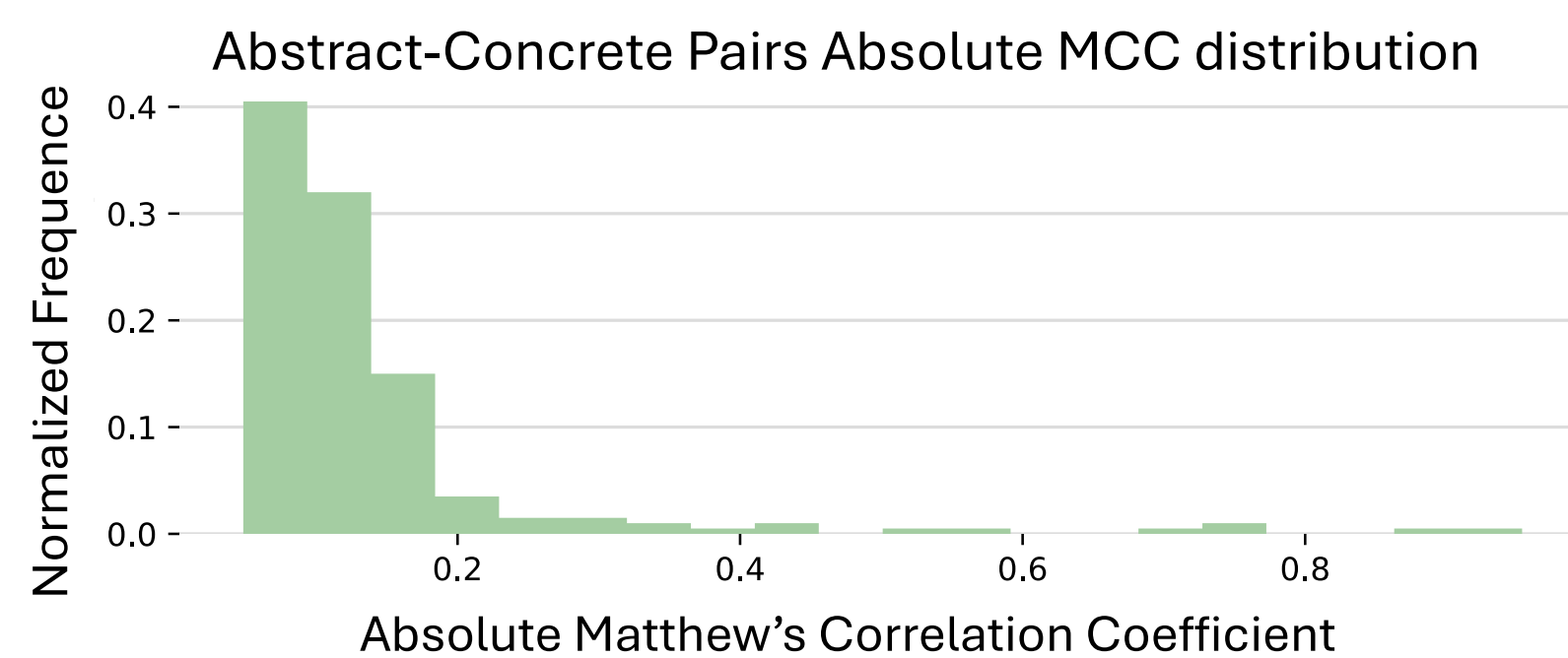


VLMs struggle to leverage abstract attributes



Abstract attributes convey novel information

Fashion descriptions are abstract-oriented



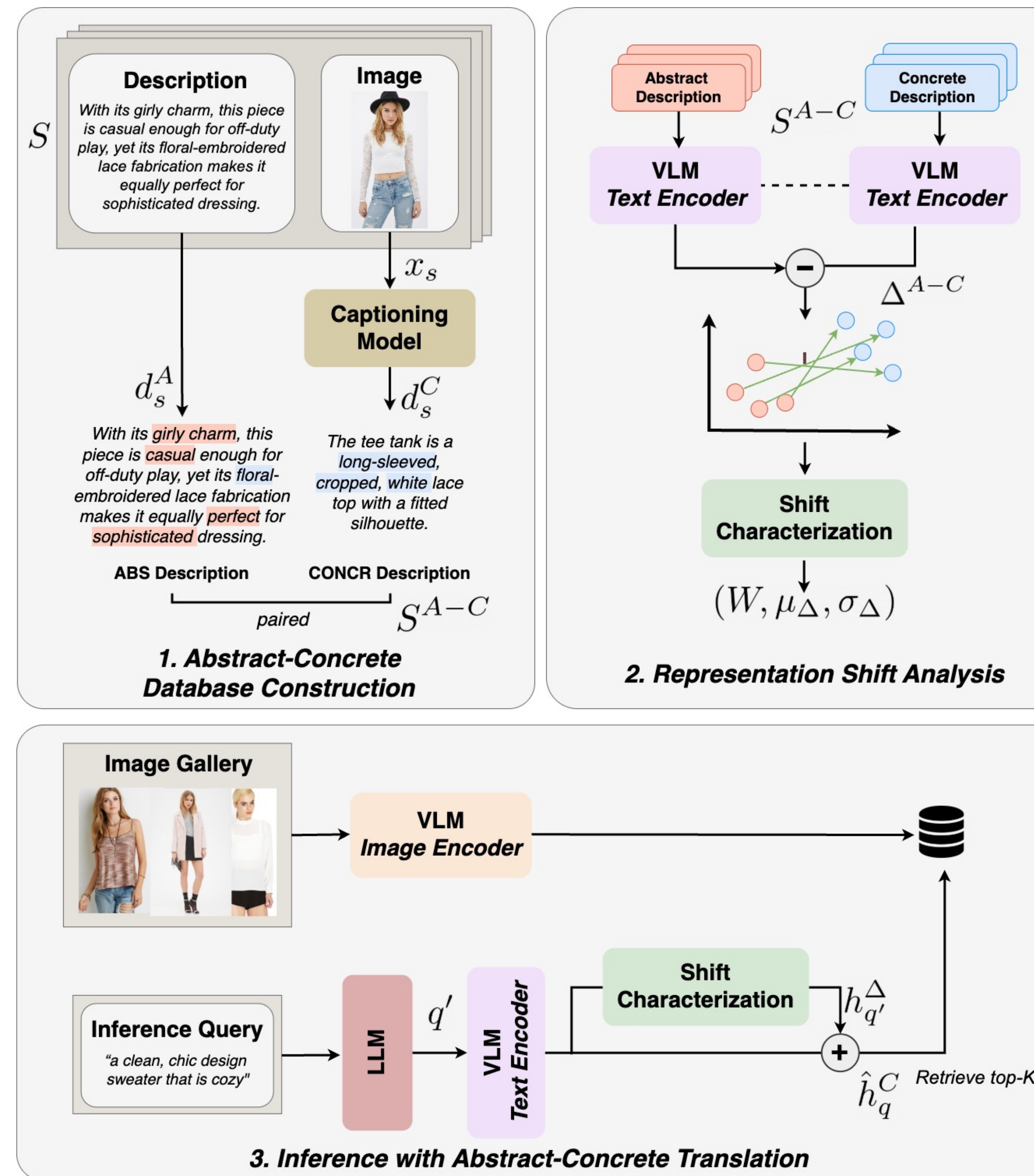
| Model/Dataset                | Concrete occurrences |         |           | Abstract occurrences |         |           |
|------------------------------|----------------------|---------|-----------|----------------------|---------|-----------|
|                              | Unique               | Total   | Per desc. | Unique               | Total   | Per desc. |
| <b>Ground-truth data</b>     |                      |         |           |                      |         |           |
| DeepFashion                  | 625                  | 10,905  | 3         | 716                  | 14,045  | 3         |
| FACAD                        | 2795                 | 202,883 | 2         | 2019                 | 194,528 | 2         |
| LAION 400M                   | 14,217               | 26M     | 0         | 12,844               | 19M     | 0         |
| <b>Captioning Models</b>     |                      |         |           |                      |         |           |
| Qwen2-VL                     | 245                  | 12,819  | 3         | 98                   | 2,623   | 0         |
| CogVLM2                      | 351                  | 19,042  | 5         | 265                  | 10,774  | 2         |
| <b>Large Language Models</b> |                      |         |           |                      |         |           |
| Phi3                         | 425                  | 7,737   | 2         | 281                  | 3,742   | 1         |
| LLaMa 3.1 8B                 | 405                  | 7,778   | 2         | 199                  | 2,675   | 0         |

## TL;DR

VLMs struggle to handle abstract-oriented adjectives, i.e. properties not perceivable with the senses. ACT better leverages them using a training-free strategy.

## Abstract-to-Concrete Translator (ACT)

Construct Abstract-Concrete paired data to learn the representation shift. At inference, **augment** abstract description **embeddings** according to the shift.



## Results

- ACT retrieval is competitive on same/cross-dataset settings
- ACT bridges the latent representation of descriptions for different VLMs
- ACT allows for effective and flexible retrieval with abstracts

| Model            | R@1  | R@5  | R@10 | H@1  | H@5  | H@10 |
|------------------|------|------|------|------|------|------|
| <b>Zero-Shot</b> |      |      |      |      |      |      |
| SigLIP           | .062 | .228 | .322 | .311 | .536 | .639 |
| CLIP             | .019 | .061 | .091 | .092 | .201 | .278 |
| O-CLIP           | .061 | .204 | .291 | .298 | .527 | .637 |
| EVA-CLIP         | .045 | .151 | .218 | .225 | .412 | .517 |
| F-CLIP           | .060 | .205 | .292 | .297 | .512 | .618 |
| OF-CLIP          | .054 | .181 | .260 | .268 | .491 | .606 |

|                                                 |             |             |             |             |             |             |
|-------------------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <b>Same-Dataset (DeepFashion → DeepFashion)</b> |             |             |             |             |             |             |
| SigLIP-ft-df                                    | .083        | .307        | .421        | .417        | .659        | .753        |
| CLIP-ft-df                                      | .029        | .095        | .142        | .140        | .299        | .393        |
| O-CLIP-ft-df                                    | .080        | .272        | .378        | .398        | .640        | .740        |
| EVA-CLIP-ft-df                                  | .063        | .219        | .316        | .317        | .539        | .654        |
| F-CLIP-ft-df                                    | .074        | .254        | .350        | .363        | .603        | .705        |
| OF-CLIP-ft-df                                   | .072        | .245        | .344        | .358        | .607        | .712        |
| <b>ACT-df (Ours)</b>                            | <b>.089</b> | <b>.303</b> | <b>.409</b> | <b>.437</b> | <b>.665</b> | <b>.756</b> |

|                                            |             |             |             |             |             |             |
|--------------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <b>Cross-Dataset (FACAD → DeepFashion)</b> |             |             |             |             |             |             |
| SigLIP-ft-facad                            | .070        | .259        | .364        | .352        | .568        | .681        |
| CLIP-ft-facad                              | .023        | .079        | .122        | .114        | .265        | .358        |
| O-CLIP-ft-facad                            | .065        | .230        | .324        | .324        | .562        | .667        |
| EVA-CLIP-ft-facad                          | .049        | .169        | .249        | .246        | .457        | .561        |
| F-CLIP-ft-facad                            | .060        | .206        | .295        | .295        | .526        | .631        |
| OF-CLIP-ft-facad                           | .059        | .209        | .298        | .298        | .542        | .649        |
| <b>ACT-facad (Ours)</b>                    | <b>.087</b> | <b>.302</b> | <b>.407</b> | <b>.428</b> | <b>.661</b> | <b>.754</b> |

| Model               | Backbone     | R@1         | R@5         | R@10        | H@1         | H@5         |
|---------------------|--------------|-------------|-------------|-------------|-------------|-------------|
| SigLIP              | VIT-B-16     | .062        | .228        | .322        | .310        | .536        |
| <b>SigLIP-ACT</b>   | VIT-B-16     | <b>.089</b> | <b>.303</b> | <b>.409</b> | <b>.437</b> | <b>.665</b> |
| SigLIP              | ViT-L-16-384 | .094        | .348        | .473        | .458        | .679        |
| <b>SigLIP-ACT</b>   | ViT-L-16-384 | <b>.108</b> | <b>.385</b> | <b>.513</b> | <b>.527</b> | <b>.736</b> |
| O-CLIP              | VIT-B-32     | .061        | .204        | .290        | .298        | .527        |
| <b>O-CLIP-ACT</b>   | VIT-B-32     | <b>.062</b> | <b>.210</b> | <b>.298</b> | <b>.312</b> | <b>.537</b> |
| O-CLIP              | VIT-L-14     | .085        | .301        | .411        | .412        | .660        |
| <b>O-CLIP-ACT</b>   | VIT-L-14     | <b>.091</b> | <b>.313</b> | <b>.422</b> | <b>.447</b> | <b>.676</b> |
| O-CLIP              | VIT-H-14     | .091        | .319        | .432        | .447        | .673        |
| <b>O-CLIP-ACT</b>   | VIT-H-14     | <b>.098</b> | <b>.336</b> | <b>.449</b> | <b>.483</b> | <b>.696</b> |
| EVA-CLIP            | EVA02-B-16   | .045        | .151        | .218        | .225        | .412        |
| <b>EVA-CLIP-ACT</b> | EVA02-B-16   | <b>.052</b> | <b>.175</b> | <b>.249</b> | <b>.258</b> | <b>.463</b> |
| EVA-CLIP            | EVA02-L-14   | .063        | .217        | .307        | .308        | .521        |
| <b>EVA-CLIP-ACT</b> | EVA02-L-14   | <b>.072</b> | <b>.252</b> | <b>.345</b> | <b>.352</b> | <b>.574</b> |
| EVA-CLIP            | EVA02-E-14   | .090        | .312        | .423        | .440        | .667        |
| <b>EVA-CLIP-ACT</b> | EVA02-E-14   | <b>.096</b> | <b>.334</b> | <b>.451</b> | <b>.471</b> | <b>.689</b> |

| Representation Shift | Language Rewriting | R@5         | H@1         |
|----------------------|--------------------|-------------|-------------|
| <b>X</b>             | <b>X</b>           | .228        | .311        |
| <b>X</b>             | Llama 3.1-8B       | .294        | .411        |
| Qwen2-VL             | <b>X</b>           | .246        | .347        |
| CogVLM2              | Llama 3.1-8B       | .302        | .433        |
| Img-embeddings       | Llama 3.1-8B       | .266        | .406        |
| Qwen2-VL             | Llama 3.2-3B       | .259        | .365        |
| Qwen2-VL             | Phi3-mini-4K       | .295        | .424        |
| Qwen2-VL             | Llama 3.1-8B       | <b>.303</b> | <b>.437</b> |

