code & data

Synthesize, Diagnose, and Optimize: Towards Fine-Grained Vision-Language Understanding

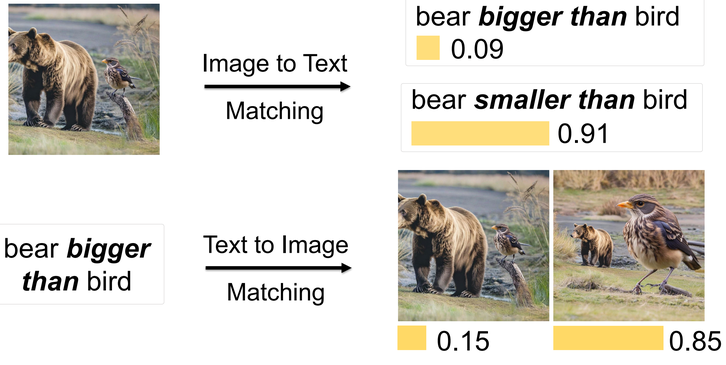
Wujian Peng[1,2] Sicheng Xie[1,2] Zuyao You[1,2] Shiyi Lan[3] Zuxuan Wu[1,2]

[1]Fudan University, [2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, [3]NVIDIA

CVPR JUNE 17-21, 2024 SEATTLE, WA

Overview

Pretrained VLMs struggle with **fine-grained compositional understanding.**



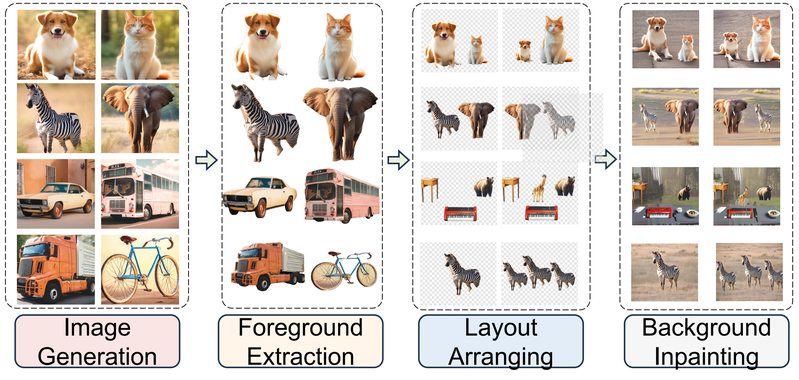
It is essential to evaluate VLMs from **both the visual and language perspectives** simultaneously.

query text: a photo of ***two / three*** dogs

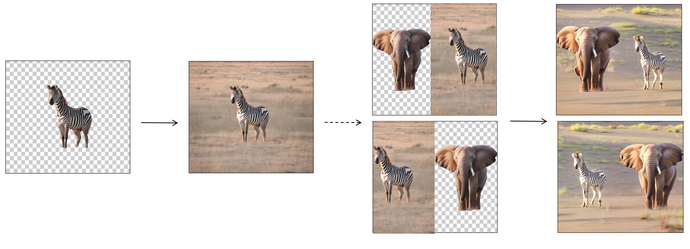
Ensuring that candidates **have only one variable is crucial** for eliminating ambiguity and enhancing reliability.

Method

We **decompose** the complex data construction process **into** several **manageable steps**.



We ensure consistency through a **shared initial background**.



We introduce a **bin-modal hard negatives** loss to optimize the VLMs.

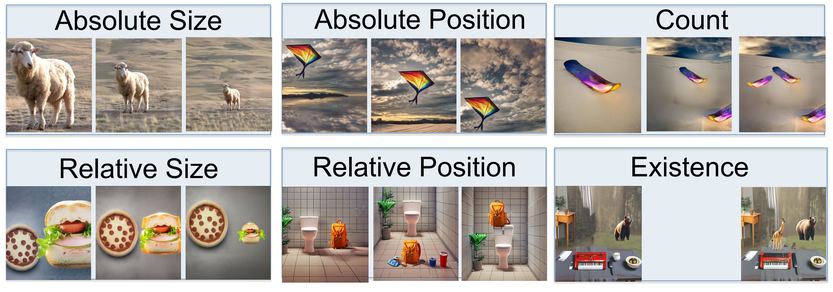
$$L_{hn}^{I2T} = -\sum_i log \frac{s(I_i, T_i)}{\sum_{T_j \in \mathbf{T}} s(I_i, T_j) + \sum_{T_k^{hn} \in \mathbf{T^{hn}}} s(I_i, T_k^{hn})}$$

$$L_{hn}^{T2I} = -\sum_i log \frac{s(I_i, T_i)}{\sum_{I_j \in \mathbf{I}} s(I_j, T_i) + \sum_{I_k^{hn} \in \mathbf{I^{hn}}} s(I_k^{hn}, T_i)}$$

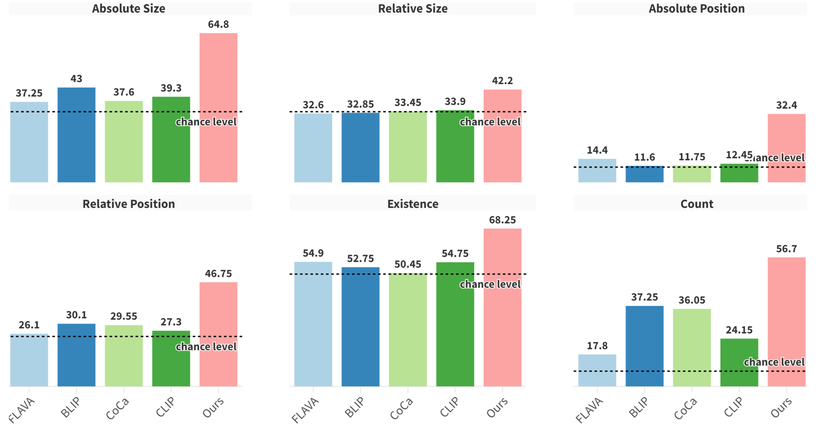
$$L_{clip} + \mu(L_{hn}^{I2T} + L_{hn}^{T2I})$$

Results

Synthesize: we constructed the SPEC benchmark, which consists of 6 subsets and a total of 15,000 image-text pairs.



Diagnose: we evaluated four leading VLMs on SPEC and highlighting their deficiencies in fine-grained understanding.



Optimize: We fine-tuned CLIP with a **hard-negative loss**, significantly enhancing its performance on SPEC.