VISLA Benchmark: Evaluating Embedding Sensitivity to Semantic and Lexical Alterations

Sri Harsha Dumpala [*1,2] **Aman Jaiswal** [*1] **Chandramouli Sastry** [1,2] **Evangelos Milios** [1] **Sageev Oore** [1,2]
Hassan Sajjad [1]

Abstract

Despite their remarkable successes, state-of-the-art language models face challenges in grasping certain important semantic details. This paper introduces the VISLA (Variance and Invariance to Semantic and Lexical Alterations) benchmark, designed to evaluate the semantic and lexical understanding of language models. VISLA presents a 3-way semantic (in)equivalence task with a triplet of sentences associated with an image, to evaluate both vision-language models (VLMs) and unimodal language models (ULMs). An evaluation involving 34 VLMs and 20 ULMs reveals surprising difficulties in distinguishing between lexical and semantic variations. Spatial semantics encoded by language models also appear to be highly sensitive to lexical information. Notably, text encoders of VLMs demonstrate greater sensitivity to semantic and lexical variations than unimodal text encoders. Our contributions include the unification of image-to-text and text-to-text retrieval tasks, an off-the-shelf evaluation without fine-tuning, and assessing LMs' semantic (in)variance in the presence of lexical alterations. The results highlight strengths and weaknesses across diverse vision and unimodal language models, contributing to a deeper understanding of their capabilities.

1. Introduction

Embeddings derived from state-of-the-art large language models (LLMs) form the foundation of several downstream applications, and even achieve human-level performance for some tasks (Zhou et al., 2023b). Despite such success, LLMs are limited in their precise understanding of the semantics of the language. For instance, they exhibit different behaviors for semantically equivalent sentences composed with different syntactic/lexical structures (Krishna et al., 2023; Cao et al., 2021; Zou et al., 2023; Meng et al., 2022). These challenges persist in vision-language models (VLMs) such as CLIP (Radford et al., 2021), particularly in the form of visio-linguistic compositionality – the difficulty in matching images to text describing their composition (Thrush et al., 2022; Yuksekgonul et al., 2023). Specifically, the image representation from VLMs is found to be more biased towards matching the word(s) from the text rather than the semantics derived from their composition.

Despite prior attempts to assess VLMs in visio-linguistic compositional reasoning, it remains unclear if the information gap lies in the image or text encoder. Further, the ambiguity persists regarding whether the text representation is enriched with compositional semantic information and how invariant this representation might be with respect to the lexical choices used to convey the semantics. For example, Figure 1 presents an example image with three captions ($\mathbf{P}_1, \mathbf{P}_2, \mathbf{N}$). Let's consider the first caption ($\mathbf{P}_1$) as our point of comparison. Semantically, $\mathbf{P}_1$ and $\mathbf{P}_2$ are equivalent, and, while $\mathbf{P}_1$ and the third caption ($\mathbf{N}$) are semantically opposite, they are syntactically and lexically similar. *Can LLMs resolve these peculiar differences between lexical similarity and semantic similarity?* In other words, do they understand the semantic relationships between the three sentences beyond the syntactic form?

In this work, we systematically evaluate language models' understanding of semantic and lexical differences between input text. We develop a benchmark dataset *VISLA*, **V**ariance and **I**nvariance to **S**emantic and **L**exical **A**lterations, to achieve this[1]. The intuition of VISLA is to disentangle the semantic and lexical similarities when interpreting the representational capabilities of a language model. VISLA achieves this by defining a set of three related captions for an image: $\mathbf{P}_1$, a caption of the image; $\mathbf{P}_2$, another caption, which is semantically equivalent to $\mathbf{P}_1$ but lexically different; and $\mathbf{N}$, an incorrect caption of the image which is lexically close to $\mathbf{P}_1$ but semantically

[*]Equal contribution [1]Dalhousie University, Canada [2]Vector Institute, Canada. Correspondence to: Sri Harsha Dumpala <sriharsha.d@dal.ca>.

[1]We will make this dataset publicly available upon acceptance of this work.

$\mathbf{P}_1$ There is a white horse pulling a trolley behind it.
$\mathbf{P}_2$ The trolley is being pulled by a white horse in front of it.
N There is a white horse pulling a trolley in front of it.

Figure 1. Figure shows an example from our VISLA benchmark. P_1 and P_2 are semantically equivalent but lexically different while N is semantically different than both P_1 and P_2 despite its lexical similarity with P_1. In our evaluations of state-of-the-art language models (consisting of 34 VLMs and 20 ULMs) on this example, we (surprisingly) find that none of them are able to successfully identify the semantically equivalent pair (P_1, P_2) from the semantically different pairs ((P_1, N), (P_2, N)).

opposite to both $\mathbf{P}_1$ and $\mathbf{P}_2$. This is also referred to as a *hard* negative caption (Hsieh et al., 2023). This triplet allows us to evaluate the compositional capability of VLMs and LLMs, while disentangling semantic and lexical similarities. Unlike the STS benchmark(s) (Cer et al., 2017) that evaluates the degree of semantic similarity between pairs of text snippets, VISLA is a 3-way semantic (in)equivalence task across varying levels of lexical shifts. Specifically, the two positives and a hard negative eliminate the trivial case of selecting between a positive and a negative as is usually done in previous benchmarks(e.g., (Hsieh et al., 2023)).

VISLA offers two benchmarking datasets: (a) generic, and (b) spatial. The generic dataset evaluates a model's ability to understand equivalent semantics with lexical variations, while the spatial dataset examines the ability of language models to correctly identify sentences describing similar spatial arrangements. In the VISLA triplets, semantically equivalent text pairs are visually represented with an image (refer to Figure 1). This design allows us to evaluate VLMs in both multimodal (vision-language) and unimodal (text-only) settings and compare their performance with LLMs, which we refer to as Unimodal Language Models (ULMs).

We consider an embedding-based methodology to evaluate language models using VISLA. The embeddings of a language model represent how it encodes semantic and lexical knowledge. We hypothesize that for an LM to correctly understand the relation between positive and negative sentences, it must encode the positive sentences closer to each other than either one to the hard negative sentence. Based on this, we perform an extensive evaluation of 34 VLMs and 20 ULMs in their ability to understand variance in semantic and lexical alteration. A few of the notable findings are summarized below:

- All text encoders—irrespective of their architecture, model size, training data size and optimization objective—struggle to separate out lexical and semantic variations.
- Spatial understanding in language models is highly sensitive to lexical information, and lexical overlap can divert models from capturing spatial semantics.
- Text encoders of vision LMs are more sensitive to semantic and lexical variations than unimodal text encoders.
- The pretraining of vision LMs that aligns text with semantic concepts of images showed better semantic understanding compared to CLIP.

Our contributions are as follows:

1. The VISLA task unifies image-to-text retrieval (VLMs) and text-to-text retrieval (VLMs, ULMs), enabling the evaluation and comparison of both VLMs and ULMs within a single benchmark.
2. The VISLA benchmark entails a 3-way semantic (in)equivalence task with varying lexical shifts, offering a more rigorous evaluation compared to benchmarks featuring only semantically distant text pairs. The hard negative caption (**N**) enables the assessment of semantic (in)variance of representation in the presence of lexical variance.
3. We perform a thorough evaluation of a large set of vision and unimodal language models, highlighting their strengths and weaknesses.

2. Related Work

VLMs and ULMs have achieved impressive results on a range of vision and language downstream tasks. These state-of-the-art VLMs and ULMs serve as foundation models for both multimodal applications, like image captioning (Li et al., 2023a), semantic segmentation of images (Ding et al., 2022; Liang et al., 2023), text-to-image generation (Ramesh et al., 2021; 2022; Saharia et al., 2022), and unimodal applications, like clustering (Wang et al., 2023b; 2022), reranking (Xiao et al., 2023), and retrieval (Li and Li, 2023). Their emergence as foundation models has motivated recent research to evaluate the strengths and weaknesses of these models. We summarize the findings from common benchmarks of VLMs and ULMs below.

Findings from the Existing Benchmark for VLMs: Thrush et al. (2022) evaluate VLMs through an image-text retrieval task and find that SOTA VLMs struggle to distinguish between texts containing the same words but ordered differently. Similarly, Yuksekgonul et al. (2023) evaluate VLMs in terms of their abilities to form object-attribute associations and highlight shortcomings of VLMs. Other studies with similar conclusions include (Zhao et al., 2022), (Ray et al., 2023) and (Wang et al., 2023a). Recent works have introduced benchmarks to evaluate different abilities of VLMs such as counter-intuitive reasoning (Zhou et al., 2023a), visual question answering (Xu et al., 2023), conceptual understanding (Schiappa et al., 2023), visio-linguistic reasoning (Chow et al., 2023), visual-spatial reasoning (Liu et al., 2023) and compositionality (Thrush et al., 2022). Kamath et al. (2023) demonstrate challenges in decoding salient aspects of input text encoded with CLIP and draw connections to the lack of compositionality in CLIP text embeddings. The task of evaluating compositionality in VLMs is the nearest neighbor to our work. Several datasets have been introduced to evaluate the compositionality of VLMs (Liu et al., 2023; Hsieh et al., 2023; Thrush et al., 2022; Zhao et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023; Ray et al., 2023; Wang et al., 2023a; Sahin et al., 2024). Most of the existing compositionality benchmarks formulate the evaluation task as image-text retrieval. Winoground (Thrush et al., 2022) is one of the earliest benchmarks to report the lack of compositional understanding in VLMs. Latest benchmarks encompassing different aspects of compositionality include VL-CheckList (Zhao et al., 2022), CREPE (Ma et al., 2023), Cola (Ray et al., 2023), and ARO (Yuksekgonul et al., 2023). Some benchmarks like Winoground have challenges beyond compositionality that include additional visual and textual reasoning (Diwan et al., 2022).

Findings from the Existing Benchmarks for ULMs: In the context of ULM text encoders, paraphrasing is the closest to our VISLA task. Paraphrasing is a well-studied problem in NLP. Several previous studies analyzed the ability of the language models to recognize paraphrasing in text. The Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) and Quora Question Pairs (QQP) (Iyer et al., 2017) are popular paraphrasing datasets (text-only without images) that are part of the GLUE (General Language Understanding Evaluation) (Wang et al., 2019) benchmark. The Semantic Textual Similarity (STS) benchmark (Cer et al., 2017) build from the STS shared tasks (Agirre et al., 2012; 2013; 2014; 2015; 2016) have pairs of text snippets with scores indicating the degree of semantic equivalence between them.

‘

Shortcoming of existing Benchmarks: Alper et al. (2023) find that the CLIP text encoder outperforms the ULMs in tasks that require implicit visual reasoning, while Chen et al. (2023a) find that ULMs perform better in terms of general language understanding. The ambiguities in the findings enforce the requirement of a more stringent benchmark with a precise diagnostic ability to discern the weaker encoder among multimodal encoders. Another research question of interest is understanding similarities and differences between ULMs (e.g., BERT (Devlin et al., 2019)) and VLMs (Alper et al., 2023; Chen et al., 2023a; Kamath et al., 2023). While the recently proposed ULM benchmark, Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) highlights the lack of generalization of ULMs on tasks involving text embeddings. They focus solely on text encoders and report an aggregate of different metrics across text-only datasets. Most VLM benchmarks are generated using rule-based algorithms (Ma et al., 2023; Yuksekgonul et al., 2023) and consist of only a pair of sentences (either semantically similar or dissimilar sentences). These similar pairs might not have strong semantic similarities, and the dissimilar pairs can have significant lexical differences, which does not represent a strict setting of evaluation. Moreover, we must finetune or linearly probe (Liu et al., 2023) these models to evaluate ULMs or VLMs text encoders using these datasets, which can require significant resources. None of the existing benchmarks systematically evaluates the resilience of model embeddings in the presence of lexical distractors (IIMURA, 2018; Taladngoen and Esteban, 2022), i.e., lexically similar but semantically different negative inputs.

3. VISLA Benchmark

We propose the VISLA benchmark to evaluate VLMs and ULMs in their ability to understand variance and invariance to semantic and lexical alterations in text. We leverage datasets (Hsieh et al., 2023; Liu et al., 2023) derived from MS-COCO (Lin et al., 2014), a large-scale dataset with images and their corresponding captions, and introduce two novel datasets. The process of acquiring the datasets for VISLA is described in Section 3.2 and Section 3.3. We further highlight how VISLA overcomes shortcomings of existing VLM and ULM benchmarks below.

3.1. Overcoming Shortcomings of Existing Benchmarks

Stricter Evaluation Setting: The VISLA benchmark comprises triplets of sentences where the first two sentences are semantically equivalent, while the third sentence is semantically opposite but remains lexically close. This setup represents a stricter and more challenging evaluation setting than text pairs that are both semantically and lexically distant.

Direct Comparison of ULMs and VLMs: The semantically equivalent text pairs in VISLA triplets are associated

with a semantically-consistent image. This allows us to evaluate VLMs in multimodal and unimodal settings and compare their performance with unimodal text encoders.

Mitigating Ambiguity: VISLA unifies evaluation of VLMs and ULMs to resolve ambiguous findings (Alper et al., 2023; Chen et al., 2023a) from previous benchmarks, allowing discernment of the weaker encoder among image and text encoders.

Evaluating Embedding Resilience to Lexical Distractors: The VISLA task contains hard negatives with high lexical overlap with positive captions. This allows us to evaluate the model's resilience to lexical distraction, i.e., its ability to recall positive captions in the presence of lexical distractors present in negative captions.

Off-the-shelf Evaluation: The VISLA triples and evaluation setting facilitates an off-of-the-shelf evaluation of VLMs and ULMs, unlike previous work (Liu et al., 2023), that may require further fine-tuning.

In addition, we follow guidelines as mentioned in Appendix A.1, to ensure the applicability of the VISLA task to both unimodal and VLMs.

3.2. Generic VISLA Benchmark

To create the generic VISLA benchmark, we build upon SUGARCREPE benchmark (Hsieh et al., 2023). SUGARCREPE leverages the recent advancements in conditional text generation using LLMs to generate hard negative captions, thereby overcoming the issues with procedurally generated captions. SUGARCREPE consists of (only) one positive and one hard negative caption for each image. The negative captions can range from contradictions in spatial arrangement to contradictions in actions or objects in the image. We expand on their methodology to further introduce an additional positive caption. The process of generating the additional positive caption consisted of two parts: 1) prompting. 2) automated and human validation.

Prompting: We followed an iterative prompting methodology to refine a final prompt that generates optimal second positive captions (P_2). Initially, we primed the generative model with a "role-play" prompt, utilizing the LLama 7b model (Touvron et al., 2023) for our generation process. Prior work, such as (Kong et al., 2023), demonstrated that "role-play" priming can enhance the reasoning abilities of LLMs. We primed the LLM to simulate the role of a "Data Generating AI", as described in Figure 2. The accuracy of LLMs in instruction following can be enhanced by employing explicit and itemized rules, a technique known as "Rules Prompting" (Zeng et al., 2023). Following the same strategy, we initially primed the model with the 'rules prompt' and subsequently with 'demonstrations' that adhere to these rules. The "rules instruction" and "demonstrations" are elaborated in Figure 4 in Appendix A.2.

> **Role-playing Prompt:** You are an instruction-following DataGenAI. Your expertise lies in accurately interpreting and generating datasets based on given instructions. Your responses are expected to be precise and limited to the required output.

Figure 2. Role playing prompt for "Data Generator AI".

Automated and Human Validation of Generated Captions: Using LLMs as evaluators for generated outputs has been extensively explored (Zheng et al., 2023; Zeng et al., 2023), providing a cost-effective alternative to human validation. We employed a distinct prompt for automatically verifying semantic consistency between original and generated prompts. The comprehensive prompt for utilizing LLM as an evaluator is detailed in Figure 5 in the Appendix A.3. The LLM was assigned to evaluate semantic consistency between positive caption pairs and generate a new caption if inconsistencies were detected. Subsequently, an expert human verified the outputs of the LLM evaluator, selecting the best positive captions among those generated in the prompting and automated evaluation steps, and making minor edits if required. We generated 973 data points comprising triplets (two positives and one negative) associated with images. Figure 6 in Appendix A.4 shows examples from the generic VISLA .

3.3. Spatial VISLA Benchmark

This dataset is a special case of generic VISLA benchmark that aims to have two positive captions conveying the same spatial relationships among the objects in the image (see Figure 7 in Appendix A.5 for examples from spatial VISLA dataset). The negative caption refers to a contradictory spatial arrangement of the objects considered in the positive captions. Below, we detail the dataset creation process that can be divided into three parts: 1) Positive caption (P_1, P_2) mining; 2) Negative caption (N) mining; and 3) Expert Validation and Annotation.

Positive Caption Mining: To create the spatial VISLA dataset, we used the Visual Spatial Reasoning (VSR) dataset (Liu et al., 2023), a subset of COCO (Lin et al., 2014). In VSR, each image has multiple captions, and the captions are accompanied by descriptive features such as subject, object, and the relationship between them, presented as separate fields. We made use of these descriptive features to select the potential P_2 and N captions. We utilized subject and object fields to identify the positive pair (P_1, P_2) for an image. For instance, given an image with caption P_1 = *'The cat is in*

the basket', where $subject = cat$ and $object = basket$, we extracted P_2 such that subject and object were switched, resulting in $subject = basket$ and $object = cat$, with caption P_2 = *'The basket contains the cat'*.

Negative Caption: We extracted an initial negative caption (N) by leveraging the observation that a different image containing the same object would likely describe a distinct spatial arrangement between them. Further, we identified the top three negative candidates for each data point by ranking all unique captions in the VSR (Liu et al., 2023) dataset using Jaccard similarity. These provided a set of negative captions candidates that human experts could select or edit.

Expert Validation and Annotation: To ensure that triplet (P_1, P_2 and N) selected above would adhere to the guidelines defined in Appendix A.1, we performed manual validation and correction of positives and negatives. Final triplets (P_1, P_2, N) were determined by reviewing each data point with input from at least one of four human experts. These experts also ensured that positive pairs (P_1, P_2) described the same spatial arrangement between objects, and that negatives described the contradictory spatial arrangement. Further, triplets that required image information were marked and modified appropriately. Utilizing the described heuristics to generate initial triplets related to the captions significantly reduced the manual effort needed to create triplets from scratch. We note that the majority of data points did require major or minor edits to meet the stringent constraints of a similar spatial arrangement of objects. Finally, we ended with 640 samples (Image, triplets) satisfying the defined guidelines. Figure 7 shows a sample image along with the corresponding triplet from the spatial VISLA benchmark.

4. Probing VLMs and ULMs using VISLA

We evaluate a comprehensive list of VLMs and ULMs on the VISLA task that differs in pre-training tasks, pre-training data size, and model size. This includes but is not limited to recent developments in VLMs that aim to improve compositionality of VLMs encoders (Li et al., 2022a; Cascante-Bonilla et al., 2023; Li et al., 2023b; Bugliarello et al., 2023; Ji et al., 2023; Singh et al., 2023) and ULMs (Xiao et al., 2023; Su et al., 2023; Wang et al., 2023b; 2022) that aims to improve the generalization across tasks involving text embeddings.

4.1. Experimental Setup

We use the text-to-text (T2T) retrieval task to evaluate the text encoders of unimodal models and VLMs on our benchmark. We use the image-to-text (I2T) retrieval task to evaluate the VLMs in a multi-modal setting. Table 1 illustrates the two retrieval settings for our benchmark, utilizing a tuple of I, P_1, P_2 and N for the T2T and I2T tasks. The evaluation process involves using triplets (P_1, P_2 and N) to assess text encoders and using the image as a query to evaluate vision encoders. The essential criterion is that the negative caption should be ranked lower than positive captions given either the Image query or Text query. Refer to Figure 3 for an illustration of the evaluation process. We report the accuracy of the model, assigning the last rank to the negative captions for both settings. Our evaluation scheme on VISLA task using the two settings, is detailed in Appendix B. These evaluations on the VISLA task provide insights into the performance of VLM and ULM encoders, described below.

Table 1. Different retrieval settings possible with our benchmark and its corresponding positive and negative candidates.

Retrieval Setting	Query	Positives	Negatives
Text-to-text (T2T)	$\{P_1/P_2\}$	$\{P_2/P_1\}$	$\{N\}$
Image-to-text (I2T)	I	$\{P_1, P_2\}$	$\{N\}$

Table 2. Comparison of different ULMs on the Generic and Spatial datasets. We report the Accuracy (%) of ranking the negative captions last, i.e., below the positive captions. We include the number of parameters in text encoders relative to BERT-base, i.e., 109.5 Million parameters. See Table 5 in Appendix C.2 for details.

Dataset / Model	# Params (BERT Scale)	Generic Acc(%)	Spatial Acc(%)
All-MiniLM-L6-v2	0.21	63.10	43.75
Bge-small-en-v1.5	0.3	74.92	47.97
All-MiniLM-L12-v2	0.3	69.06	47.50
GTE-small	0.3	70.30	48.13
Angle-BERT-base-uncased-nli-en-v1	1	74.51	51.56
BGE-base-en-v1.5	1	73.79	50.16
Sentence-T5-base	1.01	71.53	51.09
GTE-base	1	70.50	50.78
Clip-ViT-B-32-multilingual-v1	1.23	42.65	38.13
Clip-ViT-B-32	1.38	39.67	30.16
Instructor-large	3.07	72.66	**52.03**
Instructor-large (custom-ins)	3.07	75.54	**52.81**
UAE-Large-V1	3.06	74.92	49.84
GTE-large	3.06	71.94	49.06
All-RoBERTa-large-v1	3.25	72.46	47.03
Stsb-RoBERTa-large	3.25	74.00	50.78
LaBSE	4.31	37.82	44.53
Sentence-T5-xl	11.34	73.48	51.09
Angle-Llama-7b-nli-v2	62.28	**79.45**	**52.34**
E5-Mistral-7b-instruct	64.95	**79.65**	**52.50**

4.2. Can ULMs understand VISLA?

There have been recent developments in the quality of text embeddings fueled by progress in language model architectures and training objectives. We sampled 23 text encoders, covering various model sizes, architectures, and training

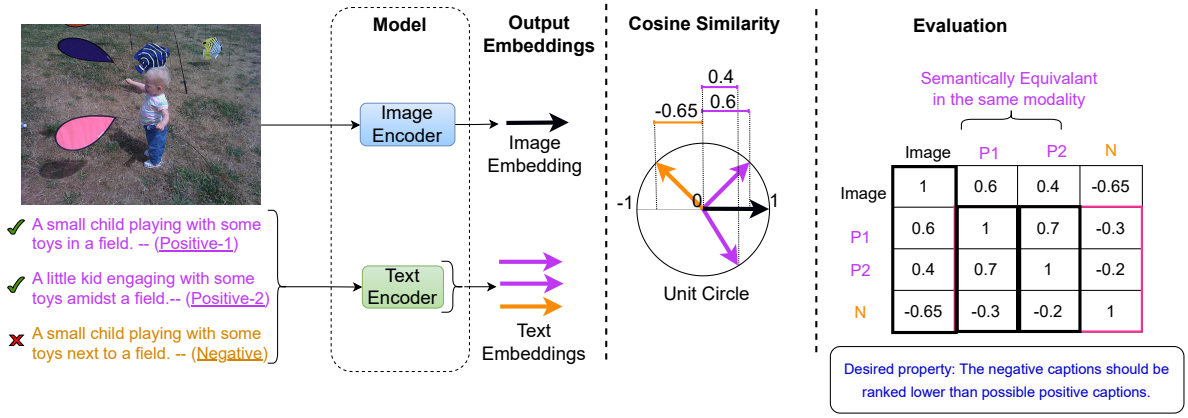


Figure 3. VISLA task Evaluation: Given an image M and a triplet of candidate captions $\{P_1, P_2, N\}$ of M, where P_1 and P_2 are semantically equivalent to each other (referred to as *positive captions* in text), we measure the accuracy of ranking the *negative caption* N below the positive captions for both the Image and Text Encoder.

objectives. For small-sized models, we inspect MiniLM (Reimers and Gurevych, 2019), GTE (Li et al., 2023c) and BGE (Xiao et al., 2023). These are trained on over 1 billion, 200 million, and 803 million sentence pairs, respectively. The results are presented in Table 2. In this category, BAAI General Embedding (BGE) (Xiao et al., 2023) outperforms others in generic VISLA, utilizing an improved contrastive loss objective in a multi-stage fashion. while in spatial VISLA , GTE-small (Li et al., 2023c) is marginally better than BGE, utilizing contrastive learning over a diverse mixture of datasets from multiple sources. MiniLM with 12 layers only marginally outperforms MiniLM with 6 layers for generic VISLA, and improves by 3.75% on spatial VISLA over the 6-layer variant. In medium-sized models, we observe no improvements in generic VISLA across all models and a modest increase of +(2.2%) and +(2.65%) on the Spatial VISLA for BGE and GTE, respectively. Suggesting that scale only has a minimal impact on encoding semantics. The BERT model trained with the recently proposed angle objective (Li and Li, 2023) outperforms other medium-sized models. Recently proposed to improve generalization across embedding tasks, the Instructor (Su et al., 2023) model, performs the best among the larger models, albeit with small improvements of 1% and 1.35% on generic and spatial VISLA. Even the 60 times larger, SOTA generative models like angle-llama-7b-nli-v2 (Li and Li, 2023) and e5-mistral-7b-instruct (Wang et al., 2023b; 2022) especially fine-tuned for emebedding tasks, peak at 52.50% ACC on the spatial VISLA and provide only an improvement of +(4.70%) and +(4.38%), over the small models in Generic and Spatial VISLA respectively. These results on the VISLA tasks indicate the difficulty of ULMs in resolving generic and spatial lexical alteration across all categories, irrespective of architecture, model size, training data size, and optimization objective. Further categorization of errors is presented in the next section.

Table 3. Comparison of ULMs on the Generic and Spatial VISLA datasets. P_1-N and P_2-N refer to the accuracy (%) of ranking positive caption 1 and positive caption 2 above the Negative caption, respectively. P_1 have more lexical overlap with N. Overall best values are in bold, and group-level best values are underlined.

Dataset	**Generic**		**Spatial**	
Model	P_1-N	P_2-N	P_1-N	P_2-N
All-MiniLM-L6-v2	92.29	64.65	<u>54.38</u>	46.56
Bge-small-en-v1.5	92.91	**77.29**	53.28	<u>52.81</u>
All-MiniLM-L12-v2	92.7	70.71	53.75	50.31
GTE-small	<u>93.53</u>	71.53	<u>54.69</u>	<u>52.19</u>
Angle-BERT-base-uncased-nli-en-v1	<u>93.73</u>	75.85	<u>55.94</u>	<u>58.28</u>
BGE-base-en-v1.5	92.6	<u>76.26</u>	54.38	55.78
Sentence-T5-base	<u>93.53</u>	73.07	<u>55</u>	<u>58.59</u>
GTE-base	<u>93.22</u>	71.84	<u>55.63</u>	54.22
Clip-ViT-B-32-multilingual-v1	79.14	44.6	52.97	45.47
Clip-ViT-B-32	78.93	41.93	52.66	35.78
Instructor-large	94.55	73.28	**56.72**	57.66
Instructor-large (custom-ins)	**95.27**	<u>76.05</u>	**56.09**	61.25
UAE-Large-V1	93.63	<u>76.46</u>	54.69	58.59
GTE-large	94.76	72.97	55.16	57.66
All-RoBERTa-large-v1	93.01	73.48	55	52.5
Stsb-RoBERTa-large	92.6	75.23	54.22	**66.88**
LaBSE	82.22	41.62	53.28	45.94
Sentence-T5-xl	94.76	74.72	55.31	63.75
Angle-Llama-7b-nli-v2	**95.68**	**80.58**	**56.41**	<u>61.41</u>
E5-Mistral-7b-instruct	**95.68**	**80.47**	55.31	60.16

4.3. Can lexical distractors overpower semantics?

We know positive captions exhibit substantial semantic alignment, while negative captions show minimal semantic alignment with their positive counterparts. In retrieval, a rudimentary scenario arises when the query and negative candidates differ both in semantics and lexicon. Our dataset is designed such that positive and negative instances deliberately showcase significant lexical overlap, while the positives show minimal lexical overlap between them. Both of these cases capture two challenging retrieval aspects in

Table 4. Comparison of VLMs performance when tested on the generic and spatial paraphrasing benchmarks. Performance reported in terms of Accuracy (%). VLM: Both vision and text encoder embeddings compared; Text: only text encoder embeddings compared. XVLM-ITR-COCO and XVLM-ITR-Flickr are finetuned on XVLM-16M models. T2T and I2T refer to text-to-text and image-to-text retrieval, respectively. Overall best values are in bold, and group-level best values are underlined.

	Generic		Spatial	
Model	T2T	I2T	T2T	I2T
CLIP-ViT-B/32	39.67	52.11	30.16	<u>44.69</u>
RoBERTa-ViT-B/32	<u>56.32</u>	<u>58.38</u>	<u>36.25</u>	37.66
ALIGN	44.50	50.56	34.53	35.16
ALIP	27.03	49.02	17.82	38.75
FLAVA	<u>57.35</u>	<u>59.40</u>	28.44	25.31
ALBEF	34.94	49.12	25.78	42.66
BLIP	51.89	54.10	39.38	<u>45.63</u>
BLIP2	48.61	51.19	<u>40.62</u>	41.09
ViLT	–	41.02	–	20.32
AltCLIP	<u>55.81</u>	<u>57.13</u>	<u>35.16</u>	<u>45.01</u>
SegCLIP	42.44	55.09	25.63	33.59
XVLM-4M	31.66	45.84	24.84	42.19
XVLM-16M	<u>49.64</u>	<u>58.79</u>	<u>31.41</u>	<u>50.31</u>
BLIP-ITM-COCO	–	61.36	–	33.59
ViLT-ITR-COCO	–	61.97	–	50.16
XVLM-ITR-COCO	**66.18**	**63.41**	**45.16**	**51.09**
XVLM-ITR-Flickr	61.25	62.89	39.69	45.16
NegCLIP	<u>54.16</u>	<u>52.41</u>	29.21	<u>34.84</u>
CLIP-SVLC	52.31	49.74	<u>30.94</u>	28.75
CyCLIP	31.03	38.23	12.50	31.41
BLIP-SGVL	–	27.85	–	33.75

terms of semantics. We illustrate this phenomenon in Figure 8 in Appendix C.1, using edit distance as a proxy for lexical overlap. For a given triplet, we reorder the positive captions so that the first positive caption, i.e., P1, exhibits a higher lexical overlap with the negative caption (N). This provides a controlled setting to inspect whether embeddings are more sensitive to lexical or semantic changes. Table 3 shows the results of this analysis. A notable observation is the evident separation in the model performances of P1-N and P2-N accuracies, i.e., the accuracy of the first caption ranking higher than the negative caption and the accuracy of the second positive caption ranking higher than the negative caption. Interestingly. We observe that P1-N accuracy is always higher than P2-N accuracy for the generic VISLA and not always higher than P2-N for the spatial VISLA . This suggests that spatial understanding in language models is highly sensitive to lexical information, and lexical overlap can divert models from capturing spatial semantics. In contrast, lexical overlap among the candidates does not hinder the correct recall of semantically equivalent options of a text query. Implying not all kinds of semantics are treated equally by the embeddings, especially spatial semantics.

4.4. How visual information effects VISLA in VLMs?

Analyzed models: We comprehensively evaluate a wide array of VLMs, which include: 1) Models trained with a contrastive learning objective such as CLIP-ViT-B/32 (Radford et al., 2021), RoBERTa-ViT-B/32 (Schuhmann et al., 2022), ALIGN (Jia et al., 2021) and ALIP (Yang et al., 2023). 2) Models trained by combining multiple objective functions, such as FLAVA (Singh et al., 2022), ALBEF (Li et al., 2021), BLIP (Li et al., 2022b) and BLIP-2 (Li et al., 2023a). 3) Models with a unified encoder for text and images, such as ViLT (Kim et al., 2021), and multi-lingual distilled models like AltCLIP (Chen et al., 2023b); 4) Models that align text with corresponding visual concepts in the image, such as SegCLIP (Luo et al., 2023), and XVLM (Zeng et al., 2022) - with two variants, XVLM-4M and XVLM-16M.

We also investigate several models that have been finetuned on downstream tasks of image-text retrieval, such as BLIP-ITM-COCO (Li et al., 2022b), ViLT-ITR-COCO (Kim et al., 2021) and XVLM-16M-ITR-COCO (Zeng et al., 2022). Specifically, BLIP, ViLT, and XVLM-16M models were trained for the ITM task using the COCO dataset. Additionally, XVLM-16M-ITR-Flickr (Zeng et al., 2022) denotes XVLM-16M models trained for the ITM task using the Flickr dataset. Moreover, we evaluate recent methods proposed to improve the compositionality of VLMs, including NegCLIP (Yuksekgonul et al., 2023), SVLC (Doveh et al., 2023), CyCLIP (Goel et al., 2022), and BLIP-SGVL (Herzig et al., 2023). These models differ in terms of model architecture, total number of parameters and embedding dimension and pretraining objectives and more details are in Table 6 of Appendix C.3.

Results: We evaluate the VLMs using the VISLA datasets in two distinct ways, T2T and I2T, as explained in Section 4.1. Table 4 provides a comparison between different VLMs. (See Tables 8 and 9in Appendix C.5 for detailed results)

<u>T2T task</u>: Among the models pretrained using contrastive loss, RoBERTa-Vit-B/32 performs better on both benchmarks. Models trained with multiple objective functions show better performance than those trained solely with a contrastive loss function. In particular, BLIP and BLIP-2 models, pretrained with contrastive, ITM, and image captioning objectives, achieve superior performance on both benchmarks. Additionally, text encoders of the models pretrained to align text with corresponding visual concepts in images perform better than CLIP-based text encoders. This indicates that the contrastive learning objective alone may not be sufficient for text encoders to learn the semantic relations between text and image. Interestingly, models proposed to improve compositionality, such as NegCLIP and CLIP-SVLC, achieve better performance than CLIP, underscoring the importance of compositionality for the VISLA task. Models fine-tuned on the downstream task of ITR,

particularly on the COCO dataset, achieve the best performance. Furthermore, for all VLMs, a significant drop in performance is observed for the spatial benchmark compared to the generic benchmark, suggesting that the text encoders of VLMs struggle with understanding spatial VISLA.

When compared with unimodal text encoders (see Table 2), the performance of all the VLMs falls behind on the VISLA task. This indicates that the text encoders of VLMs are more sensitive to the semantic and lexical variations compared to the unimodal text encoders.

I2T task: All VLMs achieve higher performance on the I2T task compared to the T2T task on both benchmarks. This demonstrates that the inclusion of visual information improves the performance of VLMs on the VISLA task. Similar to T2T task, models trained using multiple objective functions perform better than models trained using only contrastive loss on I2T task. Both models trained exclusively with contrastive loss and those trained using multiple objective functions show significant improvements in performance on the I2T task compared to the T2T task. The model pretrained to align text with semantic concepts of images performs better than CLIP, indicating that better semantic understanding is beneficial for the VISLA task. However, models trained to improve compositionality do not show consistent improvements in performance on the I2T task compared to the T2T task. Moreover, the performance of these models is comparable or lower than CLIP models. Particularly on the spatial benchmark, these models show a degradation in performance compared to CLIP.

4.5. Does model and data size effect VISLA ?

To address this question, we investigated various CLIP (Radford et al., 2021) variants trained on the WebImageText dataset, which comprises 400 million image-text pairs. These models encompass CNN-based architectures, such as RN50, RN101, RN50×4, RN50×16, and RN50×64, as well as transformer-based models like ViT-B/32 and ViT-L/14. Additionally, we examined CLIP-based models introduced by (Schuhmann et al., 2022) and (Gadre et al., 2023), pre-trained on extensive paired image-text datasets. Schuhmann et al. (2022) provided diverse CLIP variants, namely RoBERTa-ViT-B/32, ViT-H/14, ViT-g/14, xlm-roberta-base-ViT-B/32, and xlm-roberta-large-ViT-H/14, trained on a large image-text dataset called "LAION-5B," which consists of 5 billion image-text pairs. Similarly, Gadre et al. (2023) released two CLIP variants, namely Large:ViT-B/16 and xlarge:ViT-L/14, trained on the DataComp dataset, comprising 13 billion image-text pairs.

The performance of various CLIP variants on VISLA benchmarks is detailed in Table 7 (in Appendix C.4). For both I2T and T2T tasks, we observed no significant variations in performance among the different CLIP variants trained on a dataset of 400 million image-text pairs. In contrast, CLIP variants trained using the LAION dataset, comprising 2 billion image-text pairs, demonstrated superior performance on the Generic benchmark compared to those trained on the 400 million-sample dataset. However, increasing the LAION training data size from 2 billion to 5 billion did not result in performance improvement. Similar trends were noted in models trained on the DataComp dataset. Notably, all CLIP variants, regardless of the training data size, exhibited lower performance on the spatial benchmark compared to the Generic benchmark.

4.6. Understanding difficulties of VLMs on VISLA

To understand the low performance of VLMs on VISLA , we analyzed the semantically equivalent and semantically different pairs in the VISLA benchmark that confuse VLMs under the two evaluation settings – I2T and T2T.

Figure 9 (in Appendix C.6.1) highlights examples for which the VLMs (I2T task: both image and text as input) failed in identifying the correct captions (P_1, P_2) given the image (I). Specifically, the VLMs place the negative caption (N) closer to the image(I) than both the positive captions. Lexical alterations such as swapping the subject and the object and replacing words with their synonyms/antonyms can mislead the VLMs when tested in the multimodal setting.

Figure 10 (in Appendix C.6.2) provides the examples for which the VLMs succeed on the I2T task but fail to identify the semantically equivalent pairs when exclusively evaluating the text encoder of the VLMs. We observe that even lexical alterations, such as simple reordering of words using synonyms and antonyms, can fool the VLM text encoders from identifying the semantically equivalent pairs from the VISLA triples. These observations may suggest that the text encoders of VLMs struggle to differentiate semantics from syntactics. This is consistent with the observation in (Kamath et al., 2023)

5. Conclusion

In this work, we evaluate a comprehensive list of VLMs and unimodal language models on the Variance and Invariance to Semantic and Lexical Alterations (VISLA) task by introducing two new datasets. We show that unimodal text encoders have difficulties with VISLA task in general. Unimodal text encoders perform moderately on generic VISLA benchmark but show significant degradation in performance on spatial VISLA benchmark.

We also show that the text encoders of VLMs perform inferior to ULM text encoders. The performance of the VLMs on VISLA task is better under the multimodal setting when compared to the unimodal text-only setting. Interestingly, varying the model architecture or size of the VLMs will not

improve the performance on the VISLA benchmark. On the other hand, increasing the pre-training dataset size is shown to improve the performance of the VLMs on VISLA. Further qualitative analysis shows that the text encoders of VLMs struggle to understand variance and invariance of semantics to simple lexical alterations.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning in general and Language Models research in particular. We discuss several limitations of language models related to the separation of semantics of an input text from its syntactic and lexical form. In order to build trust-worthy Language Models, it is important to establish that the language models emphasize semantics contained in a sentence rather than the lexical form and syntactic style of the sentence. Our evaluation provides evidence of this problem through two curated datasets and can potentially be impactful for evaluating newer language models and/or inspiring novel solutions to this problem. There are many other potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

We thank the Canadian Institute for Advanced Research (CIFAR) for their support. Resources used in preparing this research were provided, in part, by NSERC, the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners.

References

E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre. Semeval-2012 task 6: a pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, page 385–393, USA, 2012. Association for Computational Linguistics.

E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo. *SEM 2013 shared task: Semantic textual similarity. In M. Diab, T. Baldwin, and M. Baroni, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-1004.

E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In P. Nakov and T. Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, Aug. 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2010. URL https://aclanthology.org/S14-2010.

E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In P. Nakov, T. Zesch, D. Cer, and D. Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2045. URL https://aclanthology.org/S15-2045.

E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, and T. Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1081. URL https://aclanthology.org/S16-1081.

M. Alper, M. Fiman, and H. Averbuch-Elor. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6788, 2023.

E. Bugliarello, A. Nematzadeh, and L. A. Hendricks. Weakly-supervised learning of visual relations in multimodal pretraining. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 3052–3071. Association for Computational Linguistics, 2023.

N. D. Cao, W. Aziz, and I. Titov. Editing factual knowledge in language models. In M. Moens, X. Huang, L. Specia, and S. W. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.522. URL https://doi.org/10.18653/v1/2021.emnlp-main.522.

P. Cascante-Bonilla, K. Shehada, J. S. Smith, S. Doveh, D. Kim, R. Panda, G. Varol, A. Oliva, V. Ordonez, R. Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023.

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

Z. Chen, G. Chen, S. Diao, X. Wan, and B. Wang. On the difference of bert-style and clip-style text encoders. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13710–13721, 2023a.

Z. Chen, G. Liu, B. Zhang, Q. Yang, and L. Wu. Altclip: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8666–8682. Association for Computational Linguistics, 2023b.

K. J. Chow, S. Tan, and M.-Y. Kan. Travlr: Now you see it, now you don't! a bimodal dataset for evaluating visio-linguistic reasoning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3314–3339, 2023.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

J. Ding, N. Xue, G. Xia, and D. Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 11573–11582. IEEE, 2022.

A. Diwan, L. Berry, E. Choi, D. Harwath, and K. Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 2236–2250. Association for Computational Linguistics, 2022.

B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

S. Doveh, A. Arbelle, S. Harary, E. Schwartz, R. Herzig, R. Giryes, R. Feris, R. Panda, S. Ullman, and L. Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 878–891. Association for Computational Linguistics, 2022.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

S. Goel, H. Bansal, S. Bhatia, R. Rossi, V. Vinay, and A. Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

R. Herzig, A. Mendelson, L. Karlinsky, A. Arbelle, R. Feris, T. Darrell, and A. Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 14077–14098. Association for Computational Linguistics, 2023.

C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

H. IIMURA. Distractor plausibility in a multiple-choice listening test. *JLTA Journal*, 21:65–81, 2018.

S. Iyer, N. Dandekar, and K. Csernai. First quora dataset release: Question pairs, 2017. URL https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs. Accessed: 2024-01-01.

Y. Ji, R. Tu, J. Jiang, W. Kong, C. Cai, W. Zhao, H. Wang, Y. Yang, and W. Liu. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6789–6798, 2023.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual

and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

A. Kamath, J. Hessel, and K.-W. Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, 2023.

W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, and X. Zhou. Better zero-shot reasoning with role-play prompting. *CoRR*, abs/2308.07702, 2023.

K. Krishna, Y. Song, M. Karpinska, J. F. Wieting, and M. Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=WbFhFvjjKj.

J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

J. Li, X. He, L. Wei, L. Qian, L. Zhu, L. Xie, Y. Zhuang, Q. Tian, and S. Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022a.

J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022b.

J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

J. Li, S. Tang, L. Zhu, W. Zhang, Y. Yang, T.-S. Chua, and F. Wu. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

X. Li and J. Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023c.

F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 7061–7070. IEEE, 2023.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

H. Luo, J. Bao, Y. Wu, X. He, and T. Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.

Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: massive text embedding benchmark. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics, 2023.

J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874. Association for Computational Linguistics, 2022.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

A. Ray, F. Radenovic, A. Dubey, B. A. Plummer, R. Krishna, and K. Saenko. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*, 2023.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.365. URL https://doi.org/10.18653/v1/2020.emnlp-main.365.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

U. Sahin, H. Li, Q. Khan, D. Cremers, and V. Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573, 2024.

M. C. Schiappa, M. Cogswell, A. Divakaran, and Y. S. Rawat. Probing conceptual understanding of large visual-language models. *arXiv preprint arXiv:2304.03659*, 2023.

C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

H. Singh, P. Zhang, Q. Wang, M. Wang, W. Xiong, J. Du, and Y. Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 869–893. Association for Computational Linguistics, 2023.

H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W. Yih, N. A. Smith, L. Zettlemoyer, and T. Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121. Association for Computational Linguistics, 2023.

U. Taladngoen and R. H. Esteban. Assumptions on plausible lexical distractors in the redesigned toeic question-response listening test. *LEARN Journal: Language Education and Acquisition Research Network*, 15(2):802–829, 2022.

T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

F. Wang, L. Ding, J. Rao, Y. Liu, L. Shen, and C. Ding. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *arXiv preprint arXiv:2308.12898*, 2023a.

L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023b.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

K. Yang, J. Deng, X. An, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023.

M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

Y. Zeng, X. Zhang, and H. Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR, 2022.

Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen. Evaluating large language models at evaluating instruction following. *CoRR*, abs/2310.07641, 2023.

T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023.

K. Zhou, E. Lai, W. B. A. Yeong, K. Mouratidis, and J. Jiang. ROME: evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10185–10197. Association for Computational Linguistics, 2023a.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=92gvk82DE-.

A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023. doi: 10.48550/ARXIV.2307.15043. URL https://doi.org/10.48550/arXiv.2307.15043.

A. VISLA Benchmark Generation

A.1. Dataset Guidelines

The main guidelines followed to create the benchmarks are:

- The lexical changes allowed to creation the three captions, include replacing words with synonyms and antonyms, reordering the words, etc. These lexical changes do not include adding more details about the image in the caption;
- Due to the lexical alterations, the three captions should not consist of any nonsensical and non-fluent errors;
- The three captions should be generated such that they do not need any visual, logical or commonsense reasoning to distinguish the semantically similar captions (C_{P1} and C_{P2}) from the semantically different caption (C_N) i.e., given only three captions without image, one should be able to distinguish C_{P1}, C_{P2} from C_N.
- To ensure fairness and avoid bias in dataset, we used gender neutral words such as 'person', 'individual', etc instead of using gender specific pronouns such as he, she, him , her etc.

A.2. Prompt for generic VISLA dataset

Rules Instruction: Given an input sentence describing an image caption, follow these steps:

1. Rephrase each provided sentence, focusing on preserving the original spatial relationship.
2. Pay careful attention to the positioning of objects or entities in relation to one another.
3. Ensure that the meaning remains consistent and that both the original and paraphrased sentences maintain logical coherence and grammatical correctness.

Demonstration: For example,
Input: Cat is under the table.
Paraphrase Idea: Rephrase the sentence to convey that the table is positioned above the cat.
Paraphrased: The table is above the cat.
Another example,
Input: The plane flies below the bright white clouds.
Paraphrase Idea: Ensure the spatial context is maintained by stating that the bright white clouds are situated above the flying plane.
Paraphrased: The plane flies below the bright white clouds.
Similarly,
Input: The third balcony is below the fourth balcony from the bottom.
Paraphrase Idea: Emphasize the consistent spatial arrangement while indicating that the fourth balcony is positioned above the third balcony from the bottom.
Paraphrased: The fourth balcony is above the third balcony from the bottom.

Remember to keep the meaning intact, and both the original and paraphrased sentences should be logically coherent and grammatically correct.

Lastly, for the final example:
Input: *[Original caption goes here]*
Paraphrase Idea: Focus on replicating the spatial arrangement while maintaining the original meaning of the sentence, correct grammar, same meaning.
Paraphrased: *[Your paraphrased sentence goes here]*

Figure 4. Rules Prompt used for priming LLM after role-playing instructions.

A.3. Validation prompt for Generic VISLA dataset

Figure 5 shows the comprehensive prompt used to validate the samples generated by priming the LLM. The outputs obtained from this prompt are further validated by a human expert. This reduces the manual effort required to create the VISLA benchmark.

Instruction: Given a pair of captions you job is to check if the second caption is consistent with the first caption. If it is consistent output the second caption as is, Otherwise rephrase the second caption to be consistent with the first sentence. We are especially interested in spatial consistency and spatial relationship of the objects with each other.
Demonstrations: examples,
Caption 1: A guy holding a skateboard is speaking into a microphone.
Caption 2: The guy holding the microphone is speaking into the skateboard.
isConsistent: No, you cannot speak into a skateboard.
newCaption: The guy is speaking into the microphone while holding a skateboard.

Caption 1: A family are playing frisbee on the beach.
Caption 2: The frisbee is being played on the beach by a family.
isConsistent: Yes, caption 2 is consistent as it is the same caption written in passive voice. new caption is the same as caption 2.
newCaption: A family are playing frisbee on the beach.

caption 1:A stop sign vandalized with an ”eating animals” sticker below the word ”stop.”
caption 2: The stop sign is below an ”eating animals” sticker.
isConsistent: The stop cannot be below and above the sticker at the same time.
newCaption: The word ”stop” sign is above an ”eating animals” sticker.

caption 1: There is a phone on top of a calculator.
caption 2: A calculator lies beneath the phone.
isConsistent:Yes, the sentences are semantically equivalent. new caption is same as caption 2.
newCaption: A calculator lies beneath the phone.

Now the same for the below caption only.
caption 1: *[Original caption goes here]*
caption 2: *[Generated caption goes here]*
isConsistent: *[Output Here]*

Figure 5. LLM Validation prompt to evaluate the generated caption.

A.4. Samples from generic dataset

Figure 6 shows examples from the generic VISLA dataset. Lexical alterations such as synonyms/antonyms of words, negations, re-ordering the words and adding non-content words were used to generate the semantically equivalent pair, and the semantically different negative caption. Most of the sentences in the dataset are generated by using combination of multiple lexical alterations.

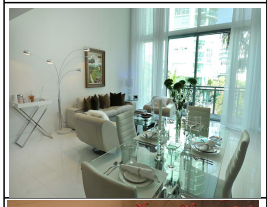
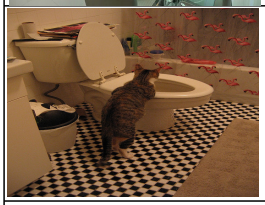
	A person pats an elephant as a couple of other persons watch. Couple of persons observe the elephant as the other person pats it. A person feeds an elephant as a couple of other persons watch.		Three people sitting on a bench facing towards the lake. Three individuals looking towards the lake while sitting on a bench Three people standing on a bench facing towards the lake.
	A white decorated living room with a set dining room table. A living room adorned in white decor featuring a arranged dining table. A white decorated living room without a set dining room table.		Three zebras located next to each other on a dirt hillside. Three zebras on a dirt hillside are situated near each other Three zebras located far from each other on a dirt hillside
	A cat has its front paws on a toilet seat. The cat's front paws are on the toilet seat. A cat has its front paws off a toilet seat.		Surfers stand with surfboards on a beach for the early morning sunrise. Surfers stand on a beach for the early morning sunrise while holding their surfboards. Surfers ride their surfboards on a beach for the early morning sunrise.
	A red truck sitting on a grassy field next to other trucks. The red truck is located next to other trucks on the grassy field. A red truck driving on a grassy field next to other trucks.		An elephant standing in a shaded cleaning in a wooded area. A shaded cleaning in a wooded area features an elephant standing in it. An elephant lying in a shaded clearing in a wooded area.
	A person riding a snowboard down a snow covered slope. An individual snowboarding down a snow-covered slope. A man is falling off a snowboard down a snow covered slope.		Two barefoot women holding game controllers in each hand. A couple of women grip game controllers with each hand while barefoot. Two barefoot women holding game controllers in one hand.

Figure 6. Some samples from the VISLA generic evaluation dataset.

A.5. Samples from spatial dataset

Figure 7 shows examples from the spatial VISLA dataset. This dataset mainly focus on the spatial arrangement of objects in the images. The semantically equivalent pair and the semantically different negative caption are generated by using different types of lexical alterations such as synonyms and antonyms of words, negations, re-ordering the words and swapping the subject and the object. Most of the sentences in the dataset are generated by using multiple lexical alterations.

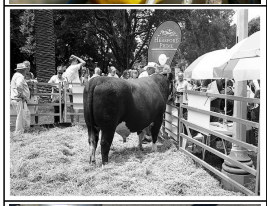
	The backpack is under the cat. The cat is on top of the backpack. The backpack is on top of the cat.		The pizza is inside the oven. The oven contains the pizza. The pizza is outside the oven.
	The bird is above the apple. The apple is below the bird. The bird is below the apple.		The horse and the person are not inside the truck. The horse and the person are outside the truck. The horse and the person are inside the truck.
	The dining table on the right side of the cow. The cow is left of the dining table. The dining table is left of the cow.		The person is standing in front of the houses. The houses are behind the standing person. The person is standing behind the houses.
	The teddy bear is in front of the fire hydrant. The fire hydrant is behind the teddy bear. The teddy bear is behind the fire hydrant.		The person is touching the sheep. The sheep is touching the person. The person is not touching the sheep.
	The teddy bear is next to the person. The person is alongside the teddy bear. The teddy bear is far from the person.		The donut and the coffee are in front of the potted plant. The potted plant is behind the donut and the coffee. The donut and the coffee are behind the potted plant.

Figure 7. Some samples from the VISLA spatial evaluation dataset.

B. Retrieval settings

We use two retrieval settings for evaluation in the VISLA task as described below,

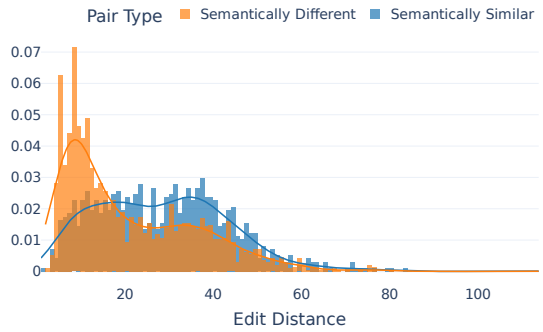
Text-to-Text Retrieval (T2T): We assess the unimodal and multimodal text encoder by providing the triplet caption set as input. For the T2T task, we extract text embeddings for the positive Caption 1 (E_{P1}), positive Caption 2 (E_{P2}), and negative Caption (E_N). We compute pairwise cosine similarity scores between Caption 1, Caption 2 and negative Caption ($S(E_{P1}, E_{P2})$, $S(E_{P1}, E_N)$, and $S(E_{P2}, E_N)$). The values of these cosine similarities determine the rank of the positive and the negative captions. We report the accuracy of the model, assigning the last rank to the negative captions.

Image-to-Text Retrieval (I2T): In this setting, we provide the image and the corresponding caption triplets as input to the Vision Language Models (VLMs). For the I2T task, we extract the image embedding (E_I), text embeddings for each positive Caption 1 (E_{P1}), positive Caption 2 (E_{P2}), and negative Caption (E_N). We compute cosine similarity scores between the image embedding and each text embedding ($S(E_I, E_{P1})$, $S(E_I, E_{P2})$, and $S(E_I, E_N)$). Similar to the T2T task, We report the accuracy of the model, assigning the last rank to the negative captions.

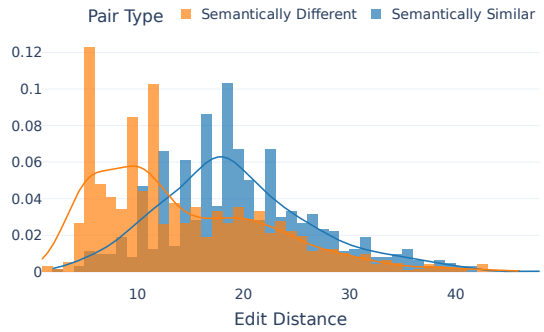
C. Additional Analysis

C.1. Lexical overlap of captions

Figure 8, demonstrates the intentional lexical overlap of semantically close caption pairs in VISLA . We employs edit distance as a measure for lexical overlap, showcasing the controlled setting used to investigate the sensitivity of embeddings to lexical versus semantic changes in language models.



(a) Generic VISLA dataset



(b) Spatial VISLA dataset

Figure 8. Distribution of edit distances between semantically similar pairs (i.e., among positive captions) and different pairs (i.e., between positive and negative captions) in the spatial VISLA dataset. Semantically similar sentences exhibit higher edit distances, indicating lexical differences. In contrast, semantically different sentences have lower edit distances, suggesting lexical similarities.

C.2. Detailed Results of ULMs on VISLA

Table 5. Comparison of ULMs on the Generic VISLA and the Spatial VISLA dataset. P1-N and P2-N refer to the accuracy (%) of ranking positive caption 1 and positive caption 2 above the Negative caption, respectively. P1- captions have more lexical overlap with the negative caption. Best performance within the same scale category is underlined, and across all models is bold-faced.We include the number of parameters in text encoders relative to BERT-base, i.e., 109.5 million parameters.

Dataset	**Dim**	**# Params**	**Generic**		**Spatial**	
Model		(BERT Scale)	**P1-N Acc(%)**	**P2-N Acc(%)**	**P1-N Acc(%)**	**P2-N Acc(%)**
All-MiniLM-L6-v2 (Wang et al., 2020)	384	0.21	92.29	64.65	54.38	46.56
BGE-small-en-v1.5 (Xiao et al., 2023)	384	0.3	92.91	**77.29**	53.28	52.81
All-MiniLM-L12-v2 (Wang et al., 2020)	384	0.3	92.7	70.71	53.75	50.31
GTE-small (Li et al., 2023c)	384	0.3	93.53	71.53	54.69	52.19
Angle-BERT-base-uncased-nli-en-v1 (Li and Li, 2023)	768	1	93.73	75.85	55.94	58.28
BGE-base-en-v1.5 (Xiao et al., 2023)	768	1	92.6	76.26	54.38	55.78
Sentence-T5-base (Ni et al., 2022)	768	1.01	93.53	73.07	55	58.59
GTE-base (Li et al., 2023c)	768	1	93.22	71.84	55.63	54.22
Clip-ViT-B-32-multilingual-v1 (Reimers and Gurevych, 2020)	512	1.23	79.14	44.6	52.97	45.47
Clip-ViT-B-32 (Radford et al., 2021)	512	1.38	78.93	41.93	52.66	35.78
Instructor-large (Su et al., 2023)	768	3.07	94.55	73.28	**56.72**	57.66
Instructor-large(custom-ins)(Su et al., 2023)	768	3.07	**95.27**	76.05	**56.09**	61.25
UAE-Large-V1 (Li and Li, 2023)	1024	3.06	93.63	76.46	54.69	58.59
GTE-large (Li et al., 2023c)	1024	3.06	94.76	72.97	55.16	57.66
All-RoBERTa-large-v1 (Reimers and Gurevych, 2019)	1024	3.25	93.01	73.48	55	52.5
Stsb-RoBERTa-large (Reimers and Gurevych, 2019)	1024	3.25	92.6	75.23	54.22	**66.88**
LaBSE (Feng et al., 2022)	768	4.31	82.22	41.62	53.28	45.94
Sentence-T5-xl (Ni et al., 2022)	768	11.34	94.76	74.72	55.31	63.75
Angle-Llama-7b-nli-v2 (Li and Li, 2023)	4096	62.28	**95.68**	**80.58**	**56.41**	61.41
E5-Mistral-7b-instruct (Wang et al., 2023b; 2022)	4096	64.95	**95.68**	**80.47**	55.31	60.16

C.3. Specification of VLMs evaluated on VISLA

We comprehensively evaluate a wide array of VLMs, which include:

1) Models trained with a contrastive learning objective such as CLIP-ViT-B/32 (Radford et al., 2021), RoBERTa-ViT-B/32 (Schuhmann et al., 2022), ALIGN (Jia et al., 2021) and ALIP (Yang et al., 2023). ALIGN and ALIP utilize noisy and synthetic captions, respectively.

2) Models trained by combining multiple objective functions, such as FLAVA (Singh et al., 2022): pretrained by combining contrastive, Image-text matching (ITM), masked image modeling (MIM) and masked language modeling (MLM) objectives; ALBEF (Li et al., 2021): which combines ITM and MLM; BLIP (Li et al., 2022b) and BLIP-2 (Li et al., 2023a): which combine contrastive, ITM and image captioning objectives.

3) Models with a unified encoder for text and images, such as ViLT (Kim et al., 2021), and multi-lingual distilled models like AltCLIP (Chen et al., 2023b)

4) Models that align text with corresponding visual concepts in the image, such as SegCLIP (Luo et al., 2023), and XVLM (Zeng et al., 2022) - with two variants, XVLM-4M and XVLM-16M.

We also investigate several models that have been finetuned on downstream tasks of image-text retrieval, such as BLIP-ITM-COCO (Li et al., 2022b), ViLT-ITR-COCO (Kim et al., 2021) and XVLM-16M-ITR-COCO (Zeng et al., 2022). Specifically, BLIP, ViLT, and XVLM-16M models were trained for the ITM task using the COCO dataset. Additionally, XVLM-16M-ITR-Flickr (Zeng et al., 2022) denotes XVLM-16M models trained for the ITM task using the Flickr dataset.

Moreover, we evaluate recent methods proposed to improve the compositionality of VLMs, including NegCLIP (Yuksekgonul et al., 2023), SVLC (Doveh et al., 2023), CyCLIP (Goel et al., 2022), and BLIP-SGVL (Herzig et al., 2023).

Table 6 provide further details about different VLMs.

Table 6. Details of the VLMs evaluated using the VISLA benchmarks. Pretraining Data type: R, N and S refer to Real, Noisy and Synthetic data types, respectively. Pretraining Objectives – ITC: image-text contrastive; ITM: image-text matching; MLM: masked language modeling; MMM: masked multimodal modeling; MIM: masked image modeling; IC: image captioning; IS: image segmentation using KL divergence; ITA: image-text alignment; CCL: Cycle-consistency loss; finetuning objectives – ITR: image-text retrieval; NL: Negative loss for text; SG: scene graph loss; PT, FT refer to pretraining and finetuning, respectively

Model	#Total Parameters	Embedding Dimension	Pretraining Data size	Pretraining Data Type	Pretraining Objectives	Finetuned
CLIP-ViT-B-32 2021	151M	512	400M	R	ITC	✗
RoBERTa-ViT-B-32 2022	212M	512	2B	R	ITC	✗
ALIGN 2021	490M	640	1.8B	R+N	ITC	✗
ALIP 2023	151M	512	15M	R+S	ITC	✗
FLAVA 2022	358M	768	70M	R	ITC, ITM, MLM MMM, MIM	✗
ALBEF 2021	210M	256	14M	R+N	ITC, ITM, MLM	✗
BLIP 2022b	225M	512	129M	R+S	ITC, ITM, IC	✗
BLIP2 2023a	1173M	256	129M	R+S	ITC, ITM, IC	✗
ViLT 2021	111M	768	10M	R	ITM, MLM	✗
AltCLIP 2023b	864M	768	42M	R	ITC	✗
SegCLIP 2023	151M	512	400M+4M	R	ITC, MIM, IS	✗
XVLM-4M 2022	216M	256	4M	R	ITC, ITM, MLM, ITA	✗
XVLM-16M 2022	216M	256	16M	R	ITC, ITM, MLM, ITA	✗
BLIP-ITM-COCO 2022b	223M	512	PT: 129M FT: 110K	R+S R	ITC, ITM, IC FT: ITM	✓
ViLT-ITR-COCO 2021	111M	768	PT: 10M FT: 110K	R	ITM, MLM FT: ITR	✓
XVLM-16M-COCO 2022	216M	256	PT:16M FT: 110K	R	ITC, ITM, MLM, ITA FT: ITR	✓
XVLM-16M-Flickr 2022	216M	256	PT: 16M FT: 30K	R	TC, ITM, MLM, ITA FT: ITR	✓
NegCLIP 2023	151M	512	PT: 400M FT:110K	R	ITC FT: ITM	✓
CLIP-SVLC 2023	151M	512	PT:400M FT:400M	R	ITC FT: ITC, NL	✓
BLIP-SGVL 2023	696M	768	PT: 129M FT:4M	R	ITC, ITM, IC FT: ITC, SG	✓
CyCLIP 2022	102M	1024	PT: 102M	R	ITC, CCL	✗

C.4. CLIP variants evaluation on VISLA

Table 7. Comparison between the performance of different variants of CLIP when tested on the generic and spatial VISLA benchmarks. Data, Model and Emb. refer to the pre-training dataset size and total number of parameters in the model (in Millions) and embedding dimension, respectively. Performance reported in terms of Accuracy (%)

	Pre-training	Pre-training	# Params	Embed.	Generic		Spatial	
Model	Dataset	Data size	Model	Dimen.	T2T	I2T	T2T	I2T
RN50	WebImageText	400M	102M	1024	39.77	53.55	33.28	45.31
RN101	WebImageText	400M	120M	512	40.29	52.21	31.09	43.9
CLIP-ViT-B/32	WebImageText	400M	151M	512	39.67	52.11	30.16	44.69
RN50x4	WebImageText	400M	178M	640	39.16	52.93	32.81	45.94
RN50x16	WebImageText	400M	291M	768	39.16	49.43	34.84	42.66
CLIP-ViT-L/14	WebImageText	400M	428M	768	39.57	50.15	27.19	42.5
RN50x64	WebImageText	400M	623M	1024	38.44	50.25	23.13	40.31
RoBERTa-ViT-B/32	LAION	2B	212M	512	56.32	**58.38**	**36.25**	37.66
ViT-H/14	LAION	2B	986M	1024	51.39	55.29	30.02	39.84
ViT-g/14	LAION	2B	1367M	1024	53.85	55.91	32.5	40.47
ViT-bigG/14	LAION	2B	2540M	1280	52.72	**58.38**	35.31	**40.48**
xlm-roberta-base-ViT-B/32	LAION	5B	366M	512	**57.86**	54.07	32.66	37.97
xlm-roberta-large-ViT-H/14	LAION	5B	1193M	1024	57.76	56.73	34.38	37.5
large:ViT-B/16	DataComp	1B	150M	512	43.06	49.85	22.03	33.91
xlarge:ViT-L/14	DataComp	13B	428M	768	49.54	53.65	27.97	38.13

C.5. Detailed Results of VLMs on VISLA

Detailed results of VLMs: In Table 8, we provide the detailed comparison of the performance of different VLMs on the generic VISLA dataset.

Table 8. Comparison of different multi-modal vision language models performance when tested on the Generic VISLA benchmark (consists of 973 samples). Data Size and Model size refer to the pre-training dataset size and total number of parameters in the model (in Millions), respectively. Performance reported in terms of Accuracy (%). P1-Ref: first positive caption compared to the negative caption; P2-Ref: second positive caption compared to the negative caption;

Model	Data	Model	Emb.	P1-N	P2-N	I2T	T2T
RN50 (Radford et al., 2021)	400M	102M	1024	70.09	64.75	53.55	39.77
RN101 (Radford et al., 2021)	400M	120M	512	67.01	65.26	52.21	40.29
CLIP-ViT-B-32 (Radford et al., 2021)	400M	151M	512	68.86	64.23	52.11	39.67
RN50x4 (Radford et al., 2021)	400M	178M	640	67.01	65.16	52.93	39.16
RN50x16 (Radford et al., 2021)	400M	291M	768	67.73	60.74	49.43	39.16
CLIP-ViT-L-14 (Radford et al., 2021)	400M	428M	768	63.52	64.54	50.15	39.57
RN50x64 (Radford et al., 2021)	400M	623M	1024	68.45	62.69	50.25	38.44
ALIGN (Jia et al., 2021)	1.8B	490M	640	67.21	62.38	50.56	44.50
ALBEF (Li et al., 2021)	14M	210M	256	76.46	54.88	49.12	34.94
SegCLIP (Luo et al., 2023)	400M	151M	512	62.58	76.98	55.09	42.44
FLAVA (Singh et al., 2022)	70M	358M	768	73.90	68.45	59.40	57.35
BLIP-Caption-COCO (Li et al., 2022b)	129M	225M	512	53.13	40.18	28.57	19.01
BLIP-ITM-COCO (Li et al., 2022b)	129M	223M	512	75.64	65.36	61.36	–
BLIP2 (Li et al., 2023a)	129M	1173M	256	62.28	64.43	51.19	48.61
XVLM-4M (Zeng et al., 2022)	4M	216M	256	77.29	51.34	45.84	31.66
XVLM-16M (Zeng et al., 2022)	16M	216M	256	78.62	64.54	58.79	49.64
XVLM-16M-COCO (Zeng et al., 2022)	16M	216M	256	80.98	67.73	63.41	66.18
XVLM-16M-Flickr (Zeng et al., 2022)	16M	216M	256	77.18	69.77	62.89	61.25
ViLT (Kim et al., 2021)	10M	111M	768	43.58	78.01	41.01	–
ViLT-ITR-COCO (Kim et al., 2021)	10M	111M	768	77.08	67.93	61.97	–
AltCLIP (Chen et al., 2023b)	42M	864M	768	72.15	65.16	57.13	55.81
ALIP (Yang et al., 2023)	15M	151M	512	74.10	55.71	49.02	27.03
RoBERTa-ViT-B-32 (Schuhmann et al., 2022)	2B	212M	512	72.25	68.24	58.38	56.32
ViT-H-14 (Schuhmann et al., 2022)	2B	986M	1024	70.30	63.72	55.29	51.39
ViT-g-14 (Schuhmann et al., 2022)	2B	1367M	1024	71.63	65.57	55.91	53.85
ViT-bigG-14 (Schuhmann et al., 2022)	2B	2540M	1280	74.92	65.16	58.38	52.72
xlm-roberta-base-ViT-B-32 (Schuhmann et al., 2022)	5B	366M	512	68.34	63.93	54.07	57.86
xlm-roberta-large-ViT-H-14 (Schuhmann et al., 2022)	5B	1193M	1024	70.81	64.34	56.73	57.76
large:ViT-B-16 (Gadre et al., 2023)	1B	150M	512	63.93	62.69	49.85	43.06
xlarge:ViT-L-14 (Gadre et al., 2023)	13B	428M	768	68.96	63.10	53.65	49.54
Models proposed to learn compositional and structural information							
NegCLIP (Yuksekgonul et al., 2023)	400M	151M	512	73.18	57.97	52.41	54.16
CLIP-SVLC (Doveh et al., 2023)	400M	151M	512	62.38	66.08	49.74	52.31
BLIP-SGVL (Herzig et al., 2023)	129M	696M	768	29.80	28.78	27.85	–
CyCLIP (Goel et al., 2022)	3M	102M	1024	54.16	56.22	38.23	31.03

Detailed results of VLMs: In Table 9, we provide the detailed comparison of the performance of different VLMs on the spatial VISLA dataset.

Table 9. Comparison of different multi-modal vision language models performance when tested on the spatial VISLA dataset (consists of 640 samples). Data Size and Model size refer to the pre-training dataset size and total number of parameters in the model (in Millions), respectively. Performance reported in terms of Accuracy (%). P1-Ref: first positive caption compared to the negative caption; P2-Ref: second positive caption compared to the negative caption; P1-P2-Ref: first and second positive caption compared to the negative caption; Text: only text encoder considered for analysis

Model	Data	Model	Emb.	P1-N	P2-N	I2T	T2T
RN50 (Radford et al., 2021)	400M	102M	1024	64.84	60.15	45.31	33.28
RN101 (Radford et al., 2021)	400M	120M	512	62.81	57.50	43.90	31.09
CLIP-ViT-B-32 (Radford et al., 2021)	400M	151M	512	63.12	60.63	44.69	30.16
RN50x4 (Radford et al., 2021)	400M	178M	640	63.13	61.56	45.94	32.81
RN50x16 (Radford et al., 2021)	400M	291M	768	59.84	58.44	42.66	34.84
CLIP-ViT-L-14 (Radford et al., 2021)	400M	428M	768	62.97	58.75	42.50	27.19
RN50x64 (Radford et al., 2021)	400M	623M	1024	59.22	56.25	40.31	23.13
ALIGN (Jia et al., 2021)	1.8B	490M	640	53.13	52.98	35.16	34.53
ALBEF (Li et al., 2021)	14M	210M	256	64.38	60.31	42.66	25.78
SegCLIP (Luo et al., 2023)	400M	151M	512	49.37	52.97	33.59	25.63
FLAVA (Singh et al., 2022)	70M	358M	768	43.43	41.25	25.31	28.44
BLIP-Caption-COCO (Li et al., 2022b)	129M	225M	512	47.50	48.44	33.62	31.25
BLIP-ITM-COCO (Li et al., 2022b)	129M	223M	512	46.40	41.56	33.59	–
BLIP2 (Li et al., 2023a)	129M	1173M	256	55.78	58.28	41.09	40.62
XVLM-4M (Zeng et al., 2022)	4M	216M	256	58.59	55.78	42.19	24.84
XVLM-16M (Zeng et al., 2022)	16M	216M	256	66.56	64.38	50.31	31.41
XVLM-16M-COCO (Zeng et al., 2022)	16M	216M	256	**67.65**	**65.93**	**51.09**	**45.16**
XVLM-16M-Flickr (Zeng et al., 2022)	16M	216M	256	64.53	60.31	45.16	39.69
ViLT (Kim et al., 2021)	10M	111M	768	34.06	48.91	20.32	–
ViLT-ITR-COCO (Kim et al., 2021)	10M	111M	768	69.06	62.97	50.16	–
AltCLIP (Chen et al., 2023b)	42M	864M	768	63.59	62.19	45.00	35.16
ALIP (Yang et al., 2023)	15M	151M	512	60.47	55.00	38.75	17.82
RoBERTa-ViT-B-32 (Schuhmann et al., 2022)	2B	212M	512	55.00	55.47	37.66	36.25
ViT-H-14 (Schuhmann et al., 2022)	2B	986M	1024	56.41	57.50	39.84	30.02
ViT-g-14 (Schuhmann et al., 2022)	2B	1367M	1024	57.50	60.31	40.47	32.50
ViT-bigG-14 (Schuhmann et al., 2022)	2B	2540M	1280	57.03	56.41	40.48	35.31
xlm-roberta-base-ViT-B-32 (Schuhmann et al., 2022)	5B	366M	512	57.97	54.84	37.97	32.66
xlm-roberta-large-ViT-H-14 (Schuhmann et al., 2022)	5B	1193M	1024	54.84	55.00	37.50	34.38
large:ViT-B-16 (Gadre et al., 2023)	1B	150M	512	56.41	50.31	33.91	22.03
xlarge:ViT-L-14 (Gadre et al., 2023)	13B	428M	768	55.63	55.94	38.13	27.97
Models proposed to learn compositional and structural information							
NegCLIP (Yuksekgonul et al., 2023)	400M	151M	512	57.34	54.38	34.84	29.21
CLIP-SVLC (Doveh et al., 2023)	400M	151M	512	59.06	49.22	28.75	30.94
BLIP-SGVL (Herzig et al., 2023)	129M	696M	768	42.03	40.94	33.75	–
CyCLIP (Goel et al., 2022)	3M	102M	1024	51.10	48.13	31.41	12.50

C.6. Qualitative Results

C.6.1. Examples that fail on both image-to-text I2T(✗) and text-to-text T2T(✗) task

	**P_1 Bunch of flowers sitting in vase filled with water and rocks on bottom.** **P_2 A vase filled with water and rocks on the bottom holds a bunch of flowers.** **N Bunch of flowers sitting in vase empty of water and rocks on bottom.**
	**P_1 An elephant standing in a shaded cleaning in a wooded area.** **P_2 A shaded cleaning in a wooded area features an elephant standing in it.** **N An elephant lying in a shaded clearing in a wooded area.**
	**P_1 An orange train engine moves down the track with one train car behind it.** **P_2 An orange train engine trailed by a single train car moves down the tracks.** **N An orange train engine moves down the track with no train cars behind it.**
	**P_1 The skier is leaning forward while jumping through the air.** **P_2 The skier jumps through the air with a lean forward.** **N The skier is leaning back while jumping through the air.**
	**P_1 A child places his hands on the head and neck of a sheep while another sheep looks at his face.** **P_2 A sheep looks at the face of the child who places his hands on another sheep's head and neck.** **N A child places his hands on the head and neck of a sheep while another sheep shies away from his face.**

Figure 9. Example from the generic VISLA that are misclassified by VLM when both image and text are provided as input, i.e., for I2T task. These examples show that for the VISLA , the issues faced by the text encoders of VLMs extend to the I2T task.

C.6.2. Examples that fail on text-to-text T2T(✗) and pass on image-to-text I2T(✓)

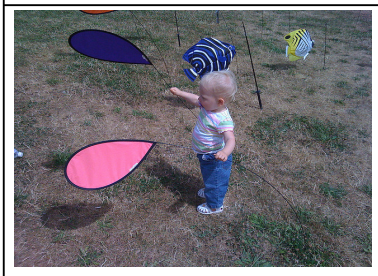
	**P_1 A brown dog laying on the ground with a metal bowl in front of him.** **P_2 A brown dog lie on the ground with a metal bowl placed before him.** **N A brown dog standing on the ground with a metal bowl in front of him.**
	**P_1 An empty clean kitchen with cabinetry, stove and dishwasher.** **P_2 An empty kitchen featuring cabinets, stove and a dishwasher is clean.** **N An empty clean kitchen without cabinetry, stove and dishwasher.**
	**P_1 People standing around in a waiting room with red floor and a flat screen TV.** **P_2 People are standing around in a red floored waiting room with a flat screen TV.** **N People sitting in a waiting room with red floor and a flat screen TV.**
	**P_1 A small child playing with some toys in a field.** **P_2 A little kid engaging with toys amidst a field.** **N A small child playing with some toys next to a field.**
	**P_1 A room full of colorful furniture and a tv.** **P_2 The room is filled with vibrant furniture and a tv.** **N A room devoid of colorful furniture and a tv.**

Figure 10. Example samples from the generic VISLA that are correctly recognized by the VLM when both image and text are provided as input but confused when only text input is provided (T2T task). These examples show that for the VISLA task, the text encoder of VLMs get confused to recognize semantically equivalent utterances even for simple lexical alterations such as negation, replacing words with synonyms, reordering of few words, etc.