

# Exploring a Unified Vision-Centric Contrastive Alternatives on Multi-Modal Web Documents

**Yiqi Lin<sup>1</sup> Alex Jinpeng Wang<sup>2</sup> Linjie Li<sup>3</sup> Zhengyuan Yang<sup>3</sup> Mike Zheng Shou<sup>1</sup>**

<sup>1</sup>Show Lab, National University of Singapore

<sup>2</sup>Central South University

<sup>3</sup>Microsoft

## Abstract

Contrastive vision-language models such as CLIP have demonstrated strong performance across a wide range of multimodal tasks by learning from aligned image-text pairs. However, their ability to handle complex, real-world web documents remains limited, particularly in scenarios where text and images are interleaved, loosely aligned, or embedded in visual form. To address these challenges, we propose Vision-Centric Contrastive Learning (VC<sup>2</sup>L), a unified framework that models text, images, and their combinations using a single vision transformer. VC<sup>2</sup>L operates entirely in pixel space by rendering all inputs, whether textual, visual, or combined, as images, thus eliminating the need for OCR, text tokenization, or modality fusion strategy. To capture complex cross-modal relationships in multimodal web documents, VC<sup>2</sup>L employs a snippet-level contrastive learning objective that aligns consecutive multimodal segments, leveraging the inherent coherence of documents without requiring explicitly paired image-text data. To assess the effectiveness of this approach, we introduce three retrieval benchmarks, AnyCIR, SeqCIR, and CSR, designed to evaluate cross-modal retrieval, fine-grained sequential understanding, and generalization to unseen data, respectively. Empirical results show that VC<sup>2</sup>L achieves competitive or superior performance compared to CLIP-style models on both the proposed benchmarks and established datasets such as M-BEIR and MTEB. These findings underscore the potential of multimodal web data as a valuable training resource for contrastive learning and illustrate the scalability of a unified, vision-centric approach for multimodal representation learning. Code and models are available at: <https://github.com/showlab/VC2L>.

## 1 Introduction

Learning vision-language correspondence from image-text pairs has significantly advanced multi-modal research, with the rise of contrastive learning methods like CLIP [1]. These models [1, 2, 3, 4, 5] align vision and language representations within a shared space and demonstrate strong zero-shot capabilities across a range of downstream tasks [6, 7, 8, 9, 10].

Despite their impressive performance, CLIP-style models face notable challenges when applied to real-world multimodal document [11, 12, 13] scenarios, e.g., retrieval, which often feature long-form content with interleaved text and images. Such scenarios reveal several key limitations in existing models. First, they struggle with interleaved multimodal inputs, where either the query or the retrieval target, or both, may contain combinations of text and images. Handling such inputs often requires additional post-processing or cross-modality fusion strategies [14]. Second, these models assume direct access to text, which is not always available in formats like scanned documents or image-based PDFs, where text is embedded as pixels and requires OCR for extraction. Finally,

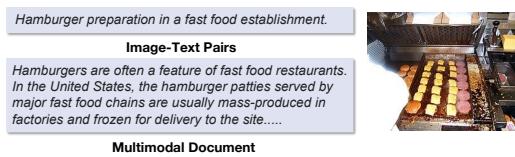


Figure 1: A comparison of image-caption pairs (Alt-Text) and multimodal documents (Wikipedia).

real-world documents are often long-form and loosely aligned across modalities, unlike datasets such as MS-COCO [15] or LAION [16], which provide clear correspondence between image-text pairs. In practice, documents frequently contain semantically related but unpaired elements, for instance, a paragraph may be followed by a relevant image without explicit correspondence or linkage, as shown in Figure 1. This setting differs significantly from the standard image-caption style data.

In this paper, we seek to explore the potential of *directly training CLIP on multi-modal interleaved documents to overcome these challenges*, given its foundational role in shaping vision-language learning. To address these challenges, we propose Vision-Centric Contrastive Learning (VC<sup>2</sup>L), a unified framework that processes all input modalities (text, images, and interleaved content) directly in pixel space. Inspired by CLIPPO [17], VC<sup>2</sup>L renders both textual and visual information as images and processes them with a single vision transformer. Input content is organized into a  $2 \times 2$  visual grid, which may contain image-only, text-only, or combined elements, as shown in Figure 2. This unified vision-centric approach eliminates the need for separate encoders, text tokenization, or OCR, and seamlessly accommodates diverse modality input forms.

Beyond input space and model unification, VC<sup>2</sup>L introduces a snippet-level contrastive learning strategy that leverages the natural coherence of document content. Rather than depending on explicitly aligned image-text pairs, our approach samples consecutive multimodal snippets from the same document and encourages their embeddings to be similar. Although these snippets are not strictly aligned, their sequential positioning often mirrors how humans interpret multimodal narratives, enabling a scalable and efficient solution for modeling interleaved real-world documents. Furthermore, we propose modality masking and text masking augmentation to diversify the contrastive target by randomly masking portions of the content within sampled multimodal snippets.

To evaluate the capacity of VC<sup>2</sup>L learn from multi-modal web documents, we design AnyCIR benchmark to evaluate the any-to-any modality information retrieval and SeqCIR benchmark to assess the fine-grained consecutive relationship modeling within documents by retrieving consecutive snippets sequentially. To evaluate the transferability of VC<sup>2</sup>L in real-world scenarios, we further design a zero-shot consecutive slide retrieval (CSR) benchmark, where slides are more complex image-text interleaved data. Our extensive experiments also show that VC<sup>2</sup>L can achieve superior zero-shot multi-modal information retrieval on M-BEIR [14] and text embedding learning on MTEB [18]. Additionally, we also investigate the impact of various contrast targets (image-caption, consecutive and non-consecutive snippets) and observe that joint image-text interleaved training can further improve language understanding in pixel space.

**Contributions:** 1). To the best of our knowledge, VC<sup>2</sup>L is the first CLIP-style framework trained directly on image-text interleaved web documents, which opens new opportunities for leveraging large-scale, loosely aligned multimodal content as training data. 2). VC<sup>2</sup>L is a single unified vision transformer operating in pixel space to handle text, images, and interleaved inputs, enabling effective multimodal understanding without OCR, tokenization, or modality-specific encoders. 3). To facilitate the evaluation of diverse modality understanding, we propose three consecutive information retrieval benchmarks, including AnyCIR, SeqCIR, and CSR. Moreover, our extensive experimental results show that VC<sup>2</sup>L achieves superior performance in our proposed benchmarks, the zero-shot multi-modal information retrieval benchmark M-BEIR, and the text embedding benchmark MTEB.

## 2 Related Work

### 2.1 Vision-Language Learning from Web Data

The pioneer work CLIP [1] establishes a breakthrough learning paradigm by applying contrastive learning on large-scale noisy image/alt-text paired data from the internet. Follow-up studies scale the image-text pairs data [16, 19] and the model design [3, 20, 4] to further improve the performance. More recently, with the rapid development of Multi-modal Large Language Models (MLLMs) [21, 22, 23], multi-modal web documents data, such as MMC4 [11] and OBELICS [12], have emerged as new sources of training data. These multi-modal documents typically consist of sequences of coherent text paragraphs interleaved with images. Several research [23, 24] demonstrate that joint training with image-text data and multi-modal web documents outperforms solely image-text pairs, which indicates the multi-modal documents contain useful vision-language correspondence from image-text pairs. Moreover, recent studies have explored advanced multi-modal embeddings across different text sources [14, 25], improved long-form caption handling [26, 27, 28] and leveraging

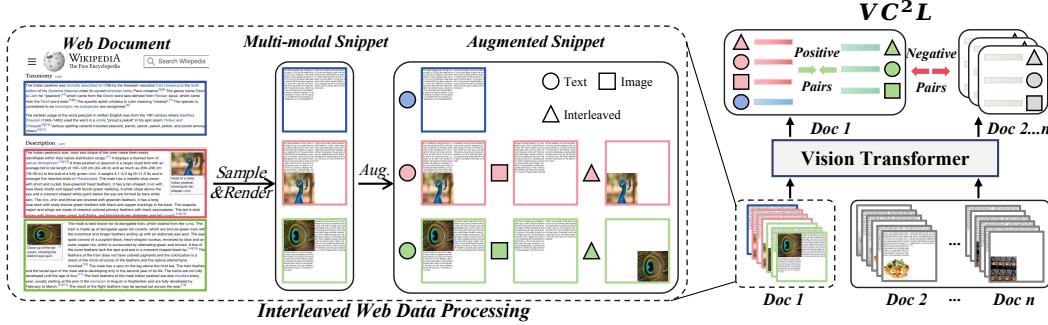


Figure 2:  $VC^2L$  explores an alternative vision-centric paradigm for unified vision-language modeling on interleaved web data. A single vision transformer is used to process any image-text modality from pixels and thereby natively learn a unified representation.

MLLMs [29, 30, 25, 31, 32, 33, 34, 35, 36, 37] to encode multi-modal information for question answering or retrieval. Differently, our goal is to offer a complementary perspective by exploring the potential of training a vision-centric CLIP model on multimodal web data, which presents new opportunities for a more versatile vision backbone in future MLLM pipelines.

## 2.2 Visual Representation for Language Modeling

Despite the impressive results achieved by text tokenization [38, 39] in language modeling [38, 40], text tokenization is vulnerable to text permutations [41], such as misspellings and has limited scalability to other languages [42]. To address these challenges, a line of work explores the tokenizer-free solution based on the visual representation of text. [43] uses glyph-vectors from Chinese character images to enhance the text representation. [41] proposed visual text representation as open-vocabularies to improve the robustness of machine translation. Recently, to close the gaps between the visual text representation and text tokenization, [42, 44, 45, 46, 47] further explore different pre-training strategies on visual text images, such as next patch prediction, next token prediction, and contrastive learning. In the vision-language domain, the most closely related work is CLIPPO [17]. CLIPPO utilizes rendered alt-text and image pairs to train the vision encoder using contrastive learning, the same as CLIP. In contrast,  $VC^2L$  marks the first attempt at exploration in the new source of training data, i.e., multimodal interleaved documents. Additionally, screenshot understanding [45, 48] is also closely related to visual text representation learning, which involves language modeling from documents [49], web pages [50], or UI images [51]. Despite directly learning text information from images, these screenshot language models can not handle omni-modality input.

## 3 Methodology

As shown in Fig. 2,  $VC^2L$  uses rendered consecutive snippets sampled from multi-modal web documents as training data. After data pre-processing and augmentation, each snippet in positive pairs can be either image-only, text-only, or an interleaved image-text rendered image. During training, the single vision model is optimized by contrastive loss on these consecutive data pairs.

### 3.1 Interleaved Web Data Processing

**Document Pre-processing.** Given a web document, our goal is to sample a pair of semantically relevant image-text snippets for training. Firstly, we split a document text into multiple text segments with a maximum of 1,100 characters in each segment. Then, we leverage the image assignment annotation provided in MMC4 dataset [11] to assign the image to its corresponding segments. Each interleaved snippet at least contains text but can be without images or assigned multiple images. For the multiple image cases, we only randomly sample one image for training.

**Data Augmentation.** Next, we apply two types of augmentations to obtain augmented snippets, i.e., *modality masking* and *text masking*. In modality masking, we only mask snippets with both text and image contents. During training, we apply modality masking with a masking rate of 40% on snippets to randomly drop one modality content. With modality masking, we are able to sample diverse training matching targets. For text masking, we randomly remove sentences from the

beginning or end of the text content in 40% of the snippets. Note that text masking is only applied to snippets containing more than four sentences. This augmentation enhances the model’s language understanding by preventing the model from overfitting recurring words.

**Multi-modal Snippet Rendering.** Given a multimodal snippet containing both image and text, we render its content into a  $2 \times 2$  grid. Each grid has a resolution of  $224 \times 224$  pixels. If the snippet includes an image, we resize it to fit the grid and place it in a randomly selected grid cell. For visual text rendering, we follow the approach in [17] using the GNU Unifont bitmap font. The long-form text can be rendered across multiple grids, starting from the top-left and proceeding left-to-right and top-to-bottom. Once one grid is fulfilled with either image or text content, the rendering process continues in the next available grid. More details are provided in the supplementary material.

### 3.2 Training Objectives

**Positive Pairs Sampling.** After data pre-processing, a document  $d_i$  is segmented as a serials of snippets, i.e.,  $\{s_i^n\}_{n=0}^N \in d_i$ . During training, we sample snippet pairs  $(s_i^q, s_i^k)$  from the same documents  $d_i$  as positive pairs, while the snippets from other documents are negative terms. We use consecutive snippets, i.e.,  $k = q + 1$ , to construct positive pairs as our default setting. To ablate the optimal training targets, we also investigate the sampling strategy of pairs with one-hop distance, i.e.,  $k = q + 2$ . To differentiate, we use **Omni**<sup>1</sup> to denote consecutive pairs only, and **Omni**+//++ to denote 20%/40% of pairs are sampled from one-hop distance pairs.

**Contrastive Learning.** Our training objective is contrastive loss [52] formulated as,

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(f_i^q \cdot f_i^k)/\tau}{\sum_{j=1}^N \exp(f_i^q \cdot f_j^k)/\tau}, \quad (1)$$

where  $(f_i^q, f_i^k)$  is the visual features extracted from sampled snippets  $(s_i^q, s_i^k)$  from the same document  $d_i$  and temperature  $\tau$  controls the sharpness of the logit distribution.

## 4 Consecutive Information Retrieval

To evaluate the consecutive information retrieval capabilities, we design two multi-modal snippet retrieval benchmarks based on OBELICS [12] and zero-shot slide retrieval based on Slideshare-1M [53]. Compared to the training dataset MMC4, the OBELICS preserves the original image text interleaved order, which is closer to real-world scenes. The slides in Slidershare-1M are naively interleaved multi-modal data with more complex interleaved forms.

**Any-to-Any Consecutive Information Retrieval (AnyCIR).** In this task, we aim to retrieve any modality consecutive information given any modality queries, as shown in Fig. 3a. The types of modality include interleaved (**IN**), Text only (**Tx**), and Image only (**Im**), resulting in 9 tasks in total with different combinations. The AnyCIR consists of 20,000 randomly sampled consecutive snippet pairs from distinct documents. Each snippet in the pair includes text and at least one image content. During inference, all the tasks share the same snippet pair source. For retrieval tasks with a single modality, we simply mask other modalities during rendering. We render images into a randomly chosen grid for both queries and candidates.

**Sequential Consecutive Information Retrieval (SeqCIR).** This task aims to evaluate the fine-grained consecutive information modeling capacity. For each query, the candidate pool consists of 26,433 snippets from 5,000 distinct documents. For each snippet, we use the full text and one randomly selected image if applicable. We use 2,524 snippets as the initial query set, which are the first snippets of the documents. For this task, we iteratively retrieve the next consecutive snippets and only successful retrieval queries are passed to the next iteration. For each iteration, we ignore the preceding snippets ahead of the query snippet in the documents. The Pass@K rate denotes the success rate of sequential retrieval at the  $k^{th}$  round, as shown in Fig. 3b. The SeqCIR is a very challenging task as the candidate pool of SeqCIR contains subsequent snippets from the same documents. It requires the model to accurately distinguish the most consecutive snippet.

**Zero-Shot Consecutive Slide Retrieval (CSR).** To better examine the transferability of VC<sup>2</sup>L under real-world scenario, we propose a benchmark of retrieving the most relevant slide. Specifically, we

---

<sup>1</sup>In this paper, Omni denotes the image, text, and image-text interleaved modality

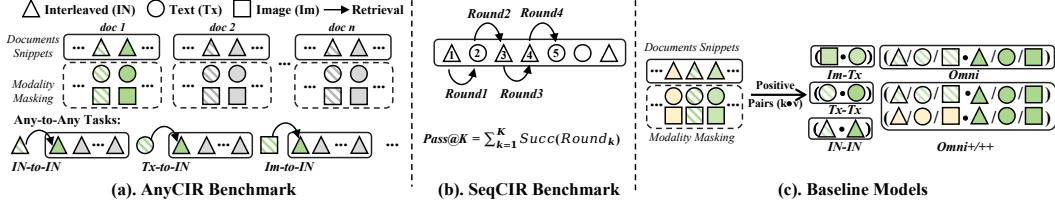


Figure 3: (a): In AnyCIR, we first sample consecutive snippet pairs from distinct documents and use the former snippet to retrieve the latter one. For each query, all the later snippets are candidates. (b): In SeqCIR, we sequentially retrieve the consecutive snippets in multiple rounds. For each query, all the snippets segmented from sampled documents are candidates while ignoring preceding snippets from the previous round. (c): The positive contrastive pair settings of different baseline models.

sample 28,016 pairs of consecutive slide images from Slideshare-1M [53]. Each pair is sampled from a distinct slide deck (more than 6 slides) after removing the first two slides. For evaluation, we use the former slide as a query and all the latter slides as candidates. Despite consecutive slides might share similar layouts or part of the content overlap, our experimental results show that it is still a challenging task even using these shortcuts instead of understanding the multimodal information.

## 5 Experiments

### 5.1 Experimental Setup.

**Data Variant Baselines.** To better understand the model capacity learned from interleaved data, we further construct different positive pair data as our baselines as illustrated in Fig. 3c. Our baselines include 1). Image-Text (**Im-Tx**) pairs sampled from a LAION subset; 2). Image-Text (**Im-Tx**) pairs from the same snippet of MMC4, where we use the MMC4 annotation to generate the pairs, i.e. the CLIP similarity assignment; 3). Text-Text (**Tx-Tx**) pairs by masking all the images in the snippets; 4). Interleaved-Interleaved (**IN-IN**) pairs by sampling from the snippets pairs containing both image and text content; 5). **Omni**<sub>224</sub> pairs first rendering in  $448 \times 448$  resolution then resize to  $224 \times 224$  resolution for fair comparison with original CLIP model; 6). **Omni+//+** denotes 20%/40% of pairs are sampled from one-hop pairs. All baselines use the same training setting.

**Implementation Details.** Our implementation is based on OpenCLIP [54]. In all experiments, we use ViT-B-16 [55] with an input resolution  $448 \times 448$ . We use a batch size of 1024 and a learning rate of 1e-4 for training 20 epochs. Our pretraining dataset uses the MMC4-core-fewer-face [11] subset, comprising 5 million documents with both images and text, totaling 17 million images. We use CLIP [1] checkpoint as our initialization due to the small scale of our training data. We include the vision encoder of CLIP [1], OpenCLIP [56], and CLIPPO [17] in the model size of ViT-B as our baseline. Note that these baselines are trained on different sources and scales of image-text pair data.

### 5.2 Consecutive Multi-Modal Retrieval

**Any-to-Any Consecutive Information Retrieval (AnyCIR).** In Table 1, we report 9 retrieval task results at Rank@1 metric. It can be observed that image-text interleaved data can help the model better understand visual text data. For example, Omni and IN-IN models achieve better results on the Tx-to-Tx retrieval task than the Tx-Tx baseline. Moreover, more diverse training data can boost the performance of omni-modality representation learning, as Omni achieves better performance on the IN-to-IN task compared to the IN-IN baseline. When training the model with non-consecutive samples, i.e., Omni+ or Omni++, the performance only slightly decreases, which indicates that the close snippets generally have consistent vision-language correspondence. Additionally, Omni<sub>224</sub> indicates that our performance gains are not only from the higher input resolution but also from our novel training data design. Interestingly, the CLIP vision encoder has stronger visual text understanding capacity over OpenCLIP, which is trained on a larger scale of datasets. When training on image-text pair data from LAION, the model performs poorly on the AnyCIR benchmark, indicating the large domain gap between image-caption and multi-modal document data.

**Sequential Consecutive Information Retrieval (SeqCIR).** Table 2 reports sequential consecutive snippets retrieval results in a total of four rounds. The best model only achieves a 3.7% success rate

Table 1: Any-to-Any Consecutive Information Retrieval benchmark on Rank@1 metric. The modalities include Image-Text Interleaved (**IN**), Text only (**Tx**), and Image only (**Im**). Gray results refer to the model input resolution as 224 and the default is 448.

Model	Data	IN-IN	IN-Tx	IN-Im	Tx-IN	Tx-Tx	Tx-Im	Im-IN	Im-Tx	Im-Im	Overall
CLIP-V[1]	WIT 400M[1]	24.10	6.18	5.27	14.23	11.47	1.02	11.60	0.93	12.45	9.69
OpenCLIP-V[54]	LAION 2B[16]	18.41	0.26	12.23	4.73	3.82	0.86	13.52	0.02	15.76	7.73
CLIPPO[17]	YFCC 100M[57]	10.17	0.01	9.99	0.00	0.01	0.01	6.31	0.02	11.79	4.25
VC <sup>2</sup> L (Omni <sub>224</sub> )	MMC4-core[11]	69.39	67.20	13.89	67.86	70.61	5.04	14.00	5.68	14.45	36.45
VC <sup>2</sup> L (Im-Tx)	LAION 40M[16]	25.64	15.23	11.89	21.21	26.40	5.72	15.07	5.36	16.20	15.86
VC <sup>2</sup> L (Im-Tx)	MMC4-core[11]	63.34	59.15	15.60	61.30	61.08	<b>12.34</b>	17.36	<b>12.31</b>	17.97	35.60
VC <sup>2</sup> L (Tx-Tx)	MMC4-core[11]	53.16	62.34	0.01	61.12	73.38	0.01	0.03	0.02	0.78	27.87
VC <sup>2</sup> L (IN-IN)	MMC4-core[11]	76.56	<b>74.85</b>	0.40	<b>74.19</b>	<b>74.81</b>	0.12	2.58	0.64	8.95	34.79
VC <sup>2</sup> L (Omni)	MMC4-core[11]	<b>78.27</b>	73.89	<b>22.10</b>	<b>74.19</b>	74.32	10.08	<b>22.00</b>	10.95	19.50	<b>42.81</b>
VC <sup>2</sup> L (Omni+)	MMC4-core[11]	77.94	73.68	21.87	73.73	73.68	10.06	21.76	10.70	19.29	42.52
VC <sup>2</sup> L (Omni++)	MMC4-core[11]	78.05	73.53	21.27	73.57	73.41	9.96	21.48	10.63	<b>19.55</b>	42.38

Table 2: Sequential Consecutive Information Retrieval. Pass@k denotes the retrieval success rate at  $k^{th}$  round. Gray results refer to the model input resolution as 224 and the default is 448.

Model	Data	Pass@1	Pass@2	Pass@3	Pass@4
CLIP-V[1]	WIT 400M[1]	11.69	1.51	0.24	0.04
OpenCLIP-V[54]	LAION 2B[16]	7.49	0.71	0.16	0.00
CLIPPO[17]	YFCC 100M[57]	3.86	0.36	0.09	0.00
VC <sup>2</sup> L (Omni <sub>224</sub> )	MMC4-core[11]	31.85	10.97	5.39	2.81
VC <sup>2</sup> L (Im-Tx)	LAION 40M[16]	13.00	1.90	0.32	0.04
VC <sup>2</sup> L (Im-Tx)	MMC4-core[11]	29.48	9.03	3.80	1.58
VC <sup>2</sup> L (Tx-Tx)	MMC4-core[11]	26.39	7.21	3.01	1.55
VC <sup>2</sup> L (IN-IN)	MMC4-core[11]	32.53	12.96	6.38	3.57
VC <sup>2</sup> L (Omni)	MMC4-core[11]	<b>34.43</b>	<b>13.07</b>	<b>6.78</b>	<b>3.76</b>
VC <sup>2</sup> L (Omni+)	MMC4-core[11]	33.28	12.60	6.50	3.68
VC <sup>2</sup> L (Omni++)	MMC4-core[11]	33.76	12.56	6.42	3.76

after four rounds, which indicates that these models still lack of capacity for fine-grained consecutive relation modeling. The results also draw the same observation as the AnyCIR benchmark that diverse training data helps omni-modality representation learning.

**Zero-Shot Consecutive Slide Retrieval (CSR).** As shown in Table 3, the Omni model achieves the best results with 44% rank@1 accuracy under zero-shot setting. It indicates that our learned interleaved representation is able to generalize to the complex interleaved data, i.e. slide. Moreover, the results demonstrate that the language understanding capacity of VC<sup>2</sup>L can be generalized beyond rendered text to various styles and font sizes. We also find that OpenCLIP is better than CLIP in CSR, which contrasts with previous benchmarks. One possible reason is that the OpenCLIP has been trained with slide data as shown in [58].

### 5.3 Traditional Multi-modal Information Retrieval

To investigate the ability of VC<sup>2</sup>L in traditional information retrieval tasks, we adopt zero-shot M-BEIR [14] for evaluation, which assembles 10 diverse datasets from multiple domains with 8

Table 3: Zero-Shot Consecutive Slides Retrieval. Gray results refer to the model input resolution as 224 and the default is 448.

Model	Data	R@1	R@5	R@10	Avg
CLIP-V[1]	WIT 400M[1]	34.60	45.10	49.29	43.00
OpenCLIP-V[54]	LAION 2B[16]	<b>38.08</b>	<b>48.33</b>	<b>52.27</b>	<b>46.23</b>
CLIPPO[17]	YFCC 100M[57]	26.42	34.31	37.30	32.68
VC <sup>2</sup> L (Omni <sub>224</sub> )	MMC4-core[11]	33.81	43.28	47.02	41.37
VC <sup>2</sup> L (Im-Tx)	LAION 40M[16]	26.21	33.13	35.85	31.73
VC <sup>2</sup> L (Im-Tx)	MMC4-core[11]	34.68	43.45	46.85	41.66
VC <sup>2</sup> L (Tx-Tx)	MMC4-core[11]	11.04	14.59	16.14	13.92
VC <sup>2</sup> L (IN-IN)	MMC4-core[11]	25.92	33.40	36.46	31.93
VC <sup>2</sup> L (Omni)	MMC4-core[11]	44.05	<b>55.55</b>	<b>59.74</b>	53.11
VC <sup>2</sup> L (Omni+)	MMC4-core[11]	<b>44.21</b>	55.54	59.68	<b>53.14</b>
VC <sup>2</sup> L (Omni++)	MMC4-core[11]	43.74	55.16	59.29	52.73

Table 4: Zero-shot results on M-BEIR<sub>union</sub> (Recall@5). Im-Tx<sub>la</sub> denotes training on LAION data.

Task	Dataset	CLIP <sub>B</sub> [1]	CLIP <sub>L</sub> [1]	SigLIP[4]	BLIP[3]	BLIP2[21]	Im-Tx <sub>la</sub>	Im-Tx	Tx-Tx	IN-IN	Omni	Omni+	Omni++
1. $q_t \rightarrow c_i$	VisualNews	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.2	0.2	0.2
	MSCOCO	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	Fashion200K	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
2. $q_t \rightarrow c_t$	WebQA	32.5	32.1	34.0	38.1	35.2	35.9	47.3	41.0	46.0	46.2	48.5	49.3
3. $q_t \rightarrow (c_i, c_t)$	EDIS	3.0	6.7	1.1	0.0	0.0	1.7	2.3	4.4	11.4	10.6	11.5	12.3
	WebQA	0.8	5.5	2.1	0.0	0.0	1.2	6.8	24.0	40.7	27.4	29.1	29.5
4. $q_i \rightarrow c_t$	VisualNews	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.3	0.2
	MSCOCO	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.3	0.3	0.3
	Fashion200K	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5. $q_t \rightarrow c_t$	NIGHTS	27.1	25.3	28.7	25.1	24.0	28.0	27.1	0.2	15.7	25.0	24.3	25.5
6. $(q_i, q_t) \rightarrow c_t$	OVEN	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.1	0.6	0.6	1.0
	InfoSeek	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.2	0.2	0.4
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	1.0	4.4	4.8	2.2	3.9	6.8	2.7	0.0	0.5	3.8	4.2	3.5
	CIRR	1.6	5.4	7.1	7.4	6.2	7.4	3.1	0.0	0.2	5.5	5.9	5.7
8. $(q_i, q_t) \rightarrow c_i$	OVEN	1.0	24.5	27.2	10.1	13.8	14.5	2.2	0.0	0.1	5.8	6.1	4.8
	InfoSeek	0.6	22.1	24.3	7.9	11.4	11.1	1.7	0.0	0.2	4.2	4.6	3.1
- Average		4.2	7.9	8.1	5.7	5.9	6.7	5.9	4.3	7.2	8.1	<b>8.5</b>	<b>8.5</b>

Table 5: Mass Text Embedding Benchmark. The rows in Cyan refer to the text encoder directly processing the text input. Gray results refer to input resolution as 224, and the default is 448.

	Class.	Clust.	PairClass.	Ranker.	Retr.	STS	Summ.	Avg.
Num. Datasets	12	11	3	4	15	10	1	56
Glove[59]	57.29	27.73	70.92	43.29	21.62	61.85	28.87	41.97
Komninos[60]	57.65	26.57	72.94	44.75	21.22	62.47	30.49	42.06
BERT[38]	61.66	30.12	56.33	43.44	10.59	54.36	29.82	38.33
SimCSE-BERT-unsup[61]	62.5	29.04	70.33	46.47	20.29	74.33	31.15	45.45
CLIP-T[1]	60.17	32.7	75.4	46	14.76	65.7	30.29	42.9
OpenCLIP-T[54]	59.2	36.61	72.43	47.91	28.05	70.43	26.57	47.76
CLIP-V[1]	55.76	31.64	63.85	45.12	14.51	62.55	26.81	40.34
OpenCLIP-V[54]	49.4	23.85	56.55	42.05	11.75	54.6	28.57	34.71
VC <sup>2</sup> L (Im-Tx/LAION)	49.04	27.67	67.34	43.67	16.49	65.26	29.74	39.27
VC <sup>2</sup> L (Im-Tx)	52.46	34.48	70.67	47.19	19.58	65.27	30.64	42.62
VC <sup>2</sup> L (Tx-Tx)	51.12	33.26	70.62	46.56	17.89	65.51	26.72	41.56
VC <sup>2</sup> L (IN-IN)	53.83	35.13	73.27	48.03	20.59	68.48	29.31	44.06
VC <sup>2</sup> L (Omni)	53.69	36.75	72.34	48.10	21.93	67.18	28.44	44.41
VC <sup>2</sup> L (Omni+)	53.25	36.95	72.50	48.34	23.07	67.62	27.91	<b>44.76</b>
VC <sup>2</sup> L (Omni++)	52.95	36.99	71.99	48.29	22.27	67.58	27.79	44.45

distinct multi-modal retrieval tasks. In our setting, we render all modality information (image and text) into a single image for all the queries and candidates without using instructions. As we find out the balance of the modality information is critical to this task, we pad all the text input to 800 chars by repeating them. We provide the ablation study results on supply materials.

Table 4 shows the zero-shot union candidate pool results of VC<sup>2</sup>L and baselines, including CLIP<sub>B</sub>(ViT-B), CLIP<sub>L</sub>(ViT-L), SigLIP [4], BLIP [3] and BLIP2 [21]. VC<sup>2</sup>L using single vision encoder outperforms the models with separate text encoder under the zero-shot setting, e.g. SigLIP. Also, it can be seen that the models trained on interleaved data generally are good at WebQA [62] while performing poorly on InfoSeek [63] compared to the CLIP-style model. It indicates that the interleaved data and image-caption data empower the model with different capacities.

#### 5.4 Text Embedding Benchmark

To evaluate the language understanding capability, we use MTEB [18] English subset, which comprises 7 different tasks in a total of 56 datasets. During inference, we render all text into images and use the pooled representation as the text embedding. We can observe that VC<sup>2</sup>L achieve competitive performance against most of unsupervised baselines, including Glove [59], Komninos [60], BERT [38] and SimCSE [61], which are trained on a large language corpus. When training with one-hop pair samples as the alignment target, our model achieves better performance. Similar to the aforementioned findings, the MTEB benchmark shows that the multi-modal data helps the model to better learn language representation from pixels. We also provide the results of the text(-T) and vision(-V) encoder performance of CLIP and OpenCLIP, where the vision encoder input is

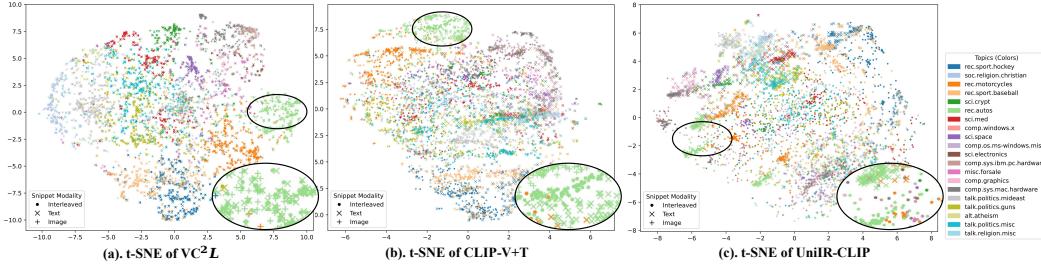


Figure 4: t-SNE visualization of interleaved, text and image snippets embedding on OBELICS.

Table 6: Ablation experiments on AnyCIR benchmark. The Avg denotes 9 tasks average performance.

(a) Model initialization.						(b) Image Rendering Positions.				
Init	Model	IN-IN	Tx-Tx	Im-Im	Avg	Position	Im-IN	Im-Tx	Im-Im	Avg
✓	IN-IN	65.85	64.55	6.46	29.60	grid-0	22.07	10.88	<b>19.53</b>	42.84
	IN-IN	<b>76.56</b>	<b>74.81</b>	<b>8.95</b>	<b>34.79</b>	grid-1	<b>22.18</b>	<b>11.03</b>	19.50	<b>42.88</b>
✓	Omni	62.30	61.22	12.18	30.42	grid-2	22.01	10.91	19.51	42.84
	Omni	<b>78.27</b>	<b>74.32</b>	<b>19.50</b>	<b>42.81</b>	grid-3	<b>22.18</b>	<b>11.03</b>	19.43	42.82
(c) Modality Masking.						(d) Text Masking.				
Ratio	IN-IN	Tx-Tx	Im-Im	Avg	Ratio	IN-IN	Tx-Tx	Im-Im	Avg	
0.0	76.56	<b>74.81</b>	8.95	34.79	0.0	77.41	72.39	19.30	41.98	
0.2	76.22	71.63	<b>19.50</b>	41.74	0.2	<b>78.34</b>	74.26	19.27	42.71	
0.4	77.41	72.39	19.30	<b>41.98</b>	0.4	78.27	<b>74.32</b>	19.50	<b>42.81</b>	
0.6	77.60	73.29	18.74	41.75	0.6	77.70	73.56	19.48	42.48	
0.8	<b>78.00</b>	73.96	17.06	40.80	0.8	77.85	73.32	<b>19.58</b>	42.42	
1.0	76.56	74.26	8.71	34.70	1.0	77.41	72.60	19.08	41.96	
(e) Non-Consecutive Pair Sampling.						Ratio	IN-IN	IN-Tx	IN-Im	Avg
0						0	<b>78.27</b>	<b>74.32</b>	19.50	<b>42.81</b>
0.1						0.1	78.04	73.53	<b>19.74</b>	42.54
0.2 (+)						0.2 (+)	77.94	73.68	19.29	42.52
0.3						0.3	78.13	73.65	19.31	42.44
0.4 (++)						0.4 (++)	78.05	73.41	19.55	42.38
0.5						0.5	77.95	73.54	19.29	42.31

rendered text at 224 resolution size. Interestingly, the text encoder of OpenCLIP outperforms all the unsupervised baselines while its vision encoder poorly understands the visual text information.

### 5.5 Discussion: Benefits of Unified Pixels Space

VC<sup>2</sup>L provide a more general-purpose vision-centric encoder that can seamlessly understand the image, visual text, and their relationship. Unifying everything into pixels can reduce specialized design in separate encoder counterparts (e.g. CLIP), resulting in a much lower computational cost compared to forwarding text inputs through an additional text encoder or extracting text through OCR models. Moreover, our approach supports a maximum text input length of 1,100 characters ( $\approx 275$  tokens) in a fixed cost, while the text input of CLIP is limited to 77 tokens.

**Embedding Space.** In Fig. 4, we visualize the distribution of interleaved, image and text embeddings from the same snippets of three models, including VC<sup>2</sup>L, CLIP-V+T with averaging features, and UniIR-CLIP [14]. The labels of the snippet are predicted by topic model [64] trained on 20NewsGroups [65]. It can be observed that our model can learn useful representations that are aligned with linguistic semantics, as snippets on similar topics are close to each other. Compared to the separate encoder baselines, VC<sup>2</sup>L learn a more unified omni-modality representation, which indicates that unifying in pixel space can further reduce the modality discrepancy.

## 6 Ablation Study and Visualization

**Effect of Model Initialization.** As shown in Table 6a, we observed that the CLIP initialization is important for VC<sup>2</sup>L. Note that our training data only contains 5 million documents with around 17 million images, which is relatively small compared to WIT-400M. The scale-up experiments are left for future study due to the computational constraints and limited data scale.

**Importance of Image Rendering Positions.** In Table 6b, we ablate the effect of the image rendering position in grids as text content uses a fixed rendering order. We rendered all the image content into the same grid positions for queries, while the candidates still use random positions. The results indicate that VC<sup>2</sup>L learns a robust representation against different rendered grid positions.

**Modality Masking and Text Masking Ratio Selection.** In Table 6c, we investigate the modality masking ratio of training data. It can be observed that modality masking is crucial for image-to-image

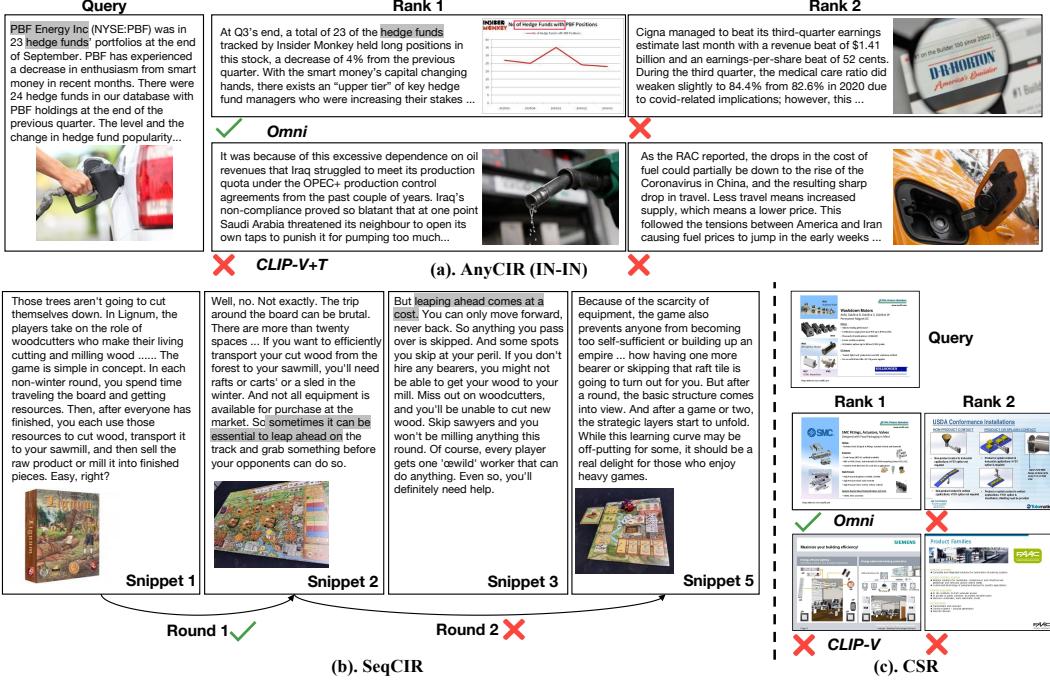


Figure 5: Visualization of retrieval results on AnyCIR, SeqCIR, and CSR benchmarks.

retrieval ability learning. In our setting, the best masking ratio is 40% and the larger ratio will drop the performance. Table 6d reports the results of applying different text masking ratios during training. We find that randomly dropping sentences in the text can improve language understanding capacity. One possible reason is that the longer text has more redundant information.

**Non-Consecutive Pair Sampling.** In Table 6e, we compare models using different ratios of one-hot consecutive pair for training. Generally, more consecutive pairs achieve higher performance on the AnyCIR benchmark as these data are more aligned with AnyCIR tasks. The one-hop consecutive pairs only slightly degrade the performance, which indicates model can learn useful representation from the non-consecutive snippets with a weaker connection.

**Retrieval Results Visualization.** As shown in Fig. 5(a) VC<sup>2</sup>L understands the loosely vision-language correspondence correctly while CLIP-V+T(feature averaging) is dominated by the image feature in AnyCIR IN-to-IN task. In Fig. 5(b), it can be observed that SeqCIR is a very challenging task as it requires the model to capture the precise connection between the consecutive snippets from omni-modality input. Lastly, Fig. 5(c) indicates that despite being trained on rendered data, VC<sup>2</sup>L can effectively generalize to real-world complex layouts with different font sizes and styles.

## 7 Conclusion and Limitations

We introduce VC<sup>2</sup>L, a unified vision-centric framework that renders interleaved multimodal content directly in pixel space, enabling a simple yet effective contrastive learning approach without relying on modality-specific components. By leveraging the natural coherence in multimodal documents and applying snippet-level contrastive learning with masking-based augmentation, VC<sup>2</sup>L learns robust representations from loosely aligned, real-world multimodal web documents. Our benchmarks validate that this vision-centric approach generalizes well across diverse retrieval scenarios and datasets. We hope that VC<sup>2</sup>L serves as a stepping stone for exploring multi-modal documents as valuable training data in the vision-language research community.

Although VC<sup>2</sup>L can process any modality input using a single model from pixels, its efficiency and scalability are limited by its fixed input size. Future work on designing a dynamic input strategy or new architecture could significantly enhance the performance and unlock more vision-centric applications for multi-modal web data understanding.

## References

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [1](#), [2](#), [5](#), [6](#), [7](#)
- [2] Chao Jia, Yafei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. [1](#)
- [3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. [1](#), [2](#), [7](#)
- [4] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. [1](#), [2](#), [7](#)
- [5] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. [1](#)
- [6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. [1](#)
- [7] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. [1](#)
- [8] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. [1](#)
- [9] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022. [1](#)
- [10] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. [1](#)
- [11] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024. [1](#), [2](#), [3](#), [5](#), [6](#), [13](#)
- [12] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. [1](#), [2](#), [4](#), [14](#)
- [13] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024. [1](#)
- [14] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023. [1](#), [2](#), [6](#), [8](#), [14](#)
- [15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. [2](#)
- [16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell 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. [2](#), [6](#)

- [17] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11017, 2023. [2](#), [3](#), [4](#), [5](#), [6](#)
- [18] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023. [2](#), [7](#), [14](#)
- [19] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [20] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. [2](#)
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [2](#), [7](#)
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. [2](#)
- [23] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. [2](#)
- [24] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mml1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. [2](#)
- [25] Young Kyun Jang, Junmo Kang, Yong Jae Lee, and Donghyun Kim. Mate: Meet at the embedding–connecting images with long texts. *arXiv preprint arXiv:2407.09541*, 2024. [2](#), [3](#)
- [26] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024. [2](#)
- [27] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*, 2024. [2](#)
- [28] Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. *arXiv preprint arXiv:2412.04616*, 2024. [2](#)
- [29] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhui Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024. [3](#)
- [30] Yujie Lu, Xiujun Li, Tsu-Jui Fu, Miguel Eckstein, and William Yang Wang. From text to pixel: Advancing long-context understanding in mllms. *arXiv preprint arXiv:2405.14213*, 2024. [3](#)
- [31] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. [3](#)
- [32] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. [3](#)
- [33] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. [3](#)
- [34] Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. Leveraging visual tokens for extended text contexts in multi-modal learning. *Advances in Neural Information Processing Systems*, 37:14325–14348, 2024. [3](#)
- [35] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*, 2024. [3](#)
- [36] Zhiheng Lyu, Xueguang Ma, and Wenhui Chen. Pixelworld: Towards perceiving everything as pixels. *arXiv preprint arXiv:2501.19339*, 2025. [3](#)

- [37] Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. *arXiv preprint arXiv:2504.17432*, 2025. 3
- [38] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 7
- [39] Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 3
- [40] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [41] Elizabeth Salesky, David Etter, and Matt Post. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, 2021. 3
- [42] Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. *arXiv preprint arXiv:2207.06991*, 2022. 3
- [43] Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [44] Chenghao Xiao, Zhuoxu Huang, Danlu Chen, G Thomas Hudson, Yizhi Li, Haoran Duan, Chenghua Lin, Jie Fu, Jungong Han, and Noura Al Moubayed. Pixel sentence representation learning. *arXiv preprint arXiv:2402.08183*, 2024. 3
- [45] Tianyu Gao, Zirui Wang, Adithya Bhaskar, and Danqi Chen. Improving language understanding from screenshots. *arXiv preprint arXiv:2402.14073*, 2024. 3
- [46] Yekun Chai, Qingyi Liu, Jingwu Xiao, Shuohuan Wang, Yu Sun, and Hua Wu. Dual modalities of text: Visual and textual generative pre-training. *arXiv preprint arXiv:2404.10710*, 2024. 3
- [47] Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025. 3
- [48] Ze Liu, Zhengyang Liang, Junjie Zhou, Zheng Liu, and Defu Lian. Any information is just worth one single screenshot: Unifying search with visualized information retrieval. *arXiv preprint arXiv:2502.11431*, 2025. 3
- [49] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, JinYeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 3
- [50] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 3
- [51] Gang Li and Yang Li. Spotlight: Mobile ui understanding using vision-language models with a focus. *arXiv preprint arXiv:2209.14927*, 2022. 3
- [52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [53] André Araujo, Jason Chaves, Haricharan Lakshman, Roland Angst, and Bernd Girod. Large-scale query-by-image video retrieval using bloom filters. *arXiv preprint arXiv:1604.07939*, 2016. 4, 5, 14
- [54] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5, 6, 7
- [55] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [56] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5

- [57] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. [6](#)
- [58] Yiqi Lin, Conghui He, Alex Jinpeng Wang, Bin Wang, Weijia Li, and Mike Zheng Shou. Parrot captions teach clip to spot text. *arXiv preprint arXiv:2312.14232*, 2023. [6](#)
- [59] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. [7](#)
- [60] Alexandros Komninos and Suresh Manandhar. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1490–1500, 2016. [7](#)
- [61] T Gao, X Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021. [7](#)
- [62] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. [7](#)
- [63] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. [7](#)
- [64] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. [8](#)
- [65] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995. [8](#)

## A Broader Impact

This work presents VC<sup>2</sup>L, a vision-centric contrastive learning framework designed to improve retrieval performance from complex, interleaved multimodal documents. By rendering both text and images into a unified pixel space and learning from consecutive web document snippets, VC<sup>2</sup>L enables efficient and scalable retrieval ability across image, text, and image-text interleaved modality inputs.

**Positive Impacts:** VC<sup>2</sup>L has the potential to significantly enhance multimodal retrieval systems by supporting retrieval across heterogeneous content formats without complex preprocessing pipelines like OCR. This makes it particularly valuable for use in digital libraries, enterprise knowledge bases, and educational platforms, where documents often contain visual and textual information. The model’s simplicity and efficiency may also lower barriers to entry for organizations with limited computational resources.

**Negative Impacts:** Improved retrieval capabilities may also carry risks. For example, the malicious user could exploit the retrieval ability to mine sensitive information from publicly available documents. Additionally, the reliance on pixel-based representations could reduce interpretability and obscure how retrieval decisions are made, potentially reinforcing hidden biases in the data.

**Mitigation Strategies:** To address these concerns, we recommend incorporating content filtering, user access controls, and explainability features into any deployed retrieval systems based on VC<sup>2</sup>L. Ensuring that training data is diverse and ethically sourced is also critical to minimizing biases. Finally, robust monitoring procedures should be in place to detect and respond to misuse.

In summary, VC<sup>2</sup>L offers a novel and practical solution for multimodal retrieval from complex document sources. However, responsible deployment and oversight are essential to mitigate potential risks and ensure its positive societal impact.

## B More Implementation Details

**Data Pre-processing.** Given a document, we chunked the document into several snippets in a sliding window strategy based on text sequence. For MMC4 [11], the document text is stored in a list of

sentences. To create snippets, we merge consecutive sentences until their combined length reaches 1100 characters or less. Then we use the image-text assignment provided by MMC4 to assign each image to the corresponding snippet. For OBELICS [12], we first split the text content based on the newline character and then use the same sliding window strategy to generate text snippets. Differently, OBELICS organizes the documents as an image-text interleaved sequence, where the image position is extracted from the original HTML files. In both AnyCIR and SeqCIR, we assign each image to the closest preceding text snippet, while images appearing at the beginning of the document are assigned to the first text snippet.

**Training Data Details.** During training, to maintain optimal text length, we apply text masking augmentation only to snippets containing more than four sentences and exceeding 250 characters. Empirically, we found that a maximum text length of 768 characters during training led to better performance. During testing, the model can handle up to 1,100 characters without any degradation in performance. Therefore, we set the maximum training text length to 768 characters and 1,100 characters for the testing setting including AnyCIR, SeqCIR and MTEB [18] benchmark. After initialization from the CLIP pre-trained checkpoint, the positional embedding is randomly initiated for  $448 \times 448$  input size. For each training batch, the data modalities are mixed from image, text, and image-text interleaved without specialized balance.

## C Additional Experiment Analysis

Table 7 presents the complete results of the AnyCIR benchmark (in total of 9 tasks) used in the ablation study, including model initialization, image rendering positions, modality masking ratio, text masking ratio and consecutive pair sampling. Moreover, we further provide the analysis of the text padding technique used in the M-BEIR [14] task. Table 8 shows the ablation study on text padding to exceed a certain length by repeating it and its impact on the performance of the M-BEIR task. Note that the number of words of the query in this sub-task (image-text pair retrieval image-text pair) is often less than 10 words. The results suggest that the short text information might be surpassed in the image-text interleaved representation for OmniContrast in such cases.

## D Visualization

**Training Data.** In Figure. 6, we showcase some rendered snippet samples used for training from the MMC4 datasets in a batch. We can observe that the model is trained for matching various target, i.e., interleaved to image, interleaved to interleaved, text to text and image to image. Note that the samples are rendered after applied with modality mask and text mask augmentations.

**Benchmark Samples.** We further present more examples of our proposed consecutive information retrieval AnyCIR (Figure. 7), SeqCIR (Figure. 8) and CSR benchmark (Figure. 9). In Figure 7, we visualize the consecutive pairs sampled from a document AnyCIR. It can be observed that the vision-language correspondence of these pairs is loose compared to the image-text caption data. In Figure. 8, we visualize a full sequence of multi-round retrieval in SeqCIR, which is very challenging because the consecutive snippets within the same documents share high relevance. In Figure. 9, we showcase some consecutive slides sampled from the slide decks [53]. Compare to the training data, the slide text are more short but with various layout and font size, which are out-of-domain data for OmniContrast. Note that some slides share the same template, requiring models not only to focus on visual context but also on language content.

Table 7: Full results of ablation study in AnyCIR.

Settings		IN-IN	IN-Tx	IN-Im	Tx-IN	Tx-Tx	Tx-Im	Im-IN	Im-Tx	Im-Im	Overall
-	IN-IN	65.85	64.26	0.10	63.84	64.55	0.05	1.10	0.19	6.46	29.60
	Init ✓	76.56	74.85	0.40	74.19	74.81	0.12	2.58	0.64	8.95	34.79
	- Omni	62.30	59.29	8.52	59.11	61.22	1.47	8.23	1.49	12.18	30.42
Image Rendering Positions	Init ✓	78.27	73.89	22.10	74.19	74.32	10.08	22.00	10.95	19.50	42.81
	grid-0	78.17	73.96	22.15	74.38	74.32	10.12	22.07	10.88	19.53	42.84
	grid-1	78.26	74.05	22.07	74.38	74.32	10.12	22.18	11.03	19.50	42.88
	grid-2	78.31	74.01	22.00	74.38	74.32	10.12	22.01	10.91	19.51	42.84
Modality Masking Ratio	grid-3	78.18	73.78	22.04	74.38	74.32	10.12	22.18	11.03	19.43	42.83
	0.0	76.56	74.85	0.40	74.19	74.81	0.12	2.58	0.64	8.95	34.79
	0.2	76.22	71.47	21.94	71.44	71.63	10.67	21.56	11.25	19.50	41.74
	0.4	77.41	72.06	21.72	72.74	72.39	9.71	21.78	10.72	19.30	41.98
	0.6	77.60	73.35	20.72	72.90	73.29	9.02	20.70	9.47	18.74	41.75
	0.8	78.00	74.32	17.38	73.93	73.96	6.89	17.96	7.69	17.06	40.80
Text Masking Ratio	1.0	76.56	74.49	0.54	74.07	74.26	0.26	2.78	0.65	8.71	34.70
	0.0	77.41	72.06	21.72	72.74	72.39	9.71	21.78	10.72	19.30	41.98
	0.2	78.34	73.96	21.85	74.25	74.26	10.16	21.46	10.89	19.27	42.71
	0.4	78.27	73.89	22.10	74.19	74.32	10.08	22.00	10.95	19.50	42.81
	0.6	77.70	73.44	21.94	73.42	73.56	10.11	21.88	10.77	19.48	42.48
	0.8	77.85	73.20	21.86	73.20	73.32	10.11	22.01	10.64	19.58	42.42
	1.0	77.41	72.38	21.60	72.66	72.60	9.67	21.64	10.61	19.08	41.96
Consecutive Pair Sampling	0.0	78.27	73.89	22.10	74.19	74.32	10.08	22.00	10.95	19.50	42.81
	0.1	78.04	73.27	21.88	73.66	73.53	9.90	21.96	10.94	19.74	42.54
	0.2	77.94	73.68	21.87	73.73	73.68	10.06	21.76	10.70	19.29	42.52
	0.3	78.13	73.46	21.46	73.76	73.65	9.98	21.51	10.68	19.31	42.44
	0.4	78.05	73.53	21.27	73.57	73.41	9.96	21.48	10.63	19.55	42.38
	0.5	77.95	73.50	21.29	73.37	73.54	9.80	21.59	10.47	19.29	42.31

Table 8: Ablation study of text padding length on M-BEIR benchmark.

Task	Dataset	Text Padding Length				
		-	100	400	800	1000
$(q_i, q_t) \rightarrow (c_i, c_t)$	oven_task8	0.26	0.65	4.37	5.77	5.21
	infoseek_task8	0.09	0.33	3.01	4.21	4.05

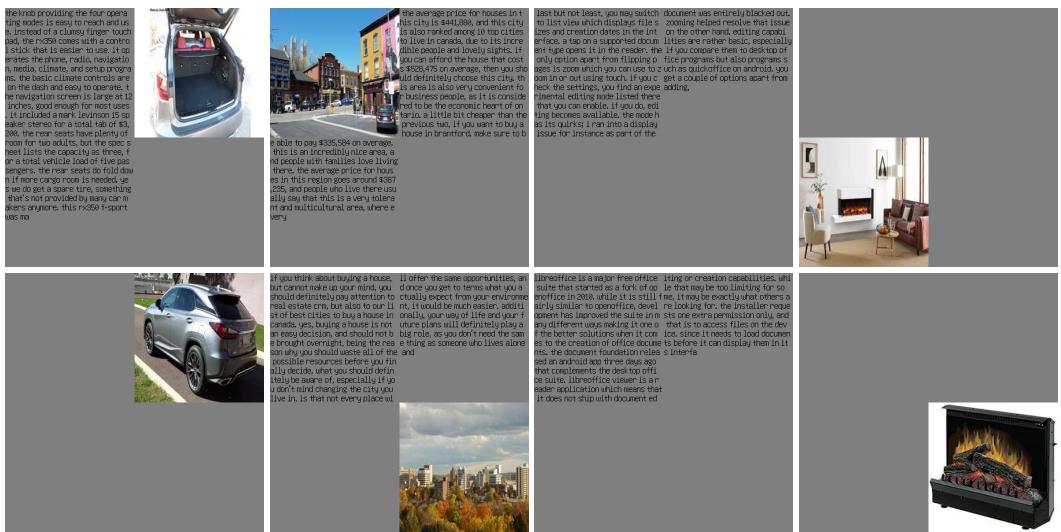


Figure 6: Rendered image-text snippets from a training batch. Each column represents the positive pairs.

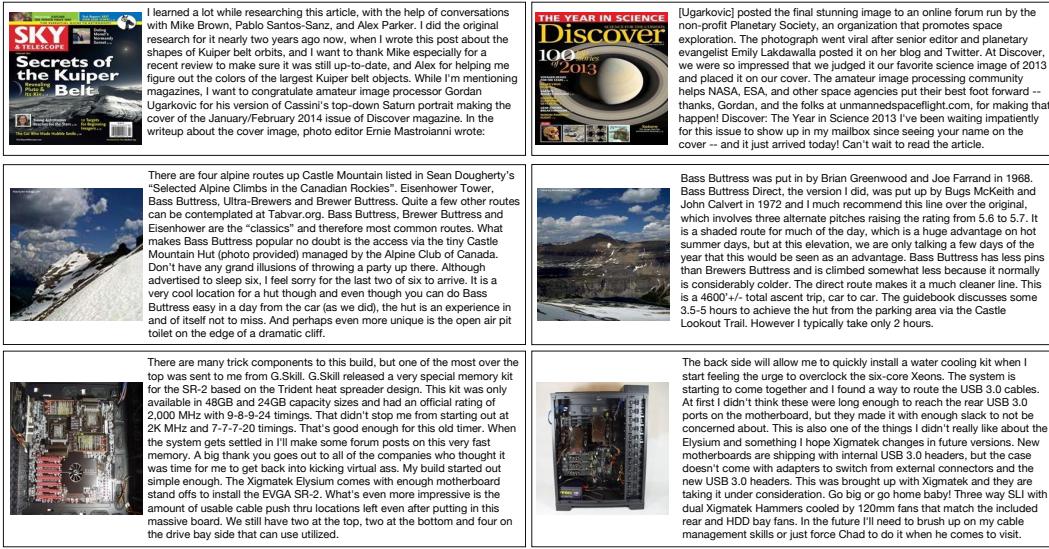


Figure 7: Visualization samples in AnyCIR benchmark. Each row represents the consecutive pairs.

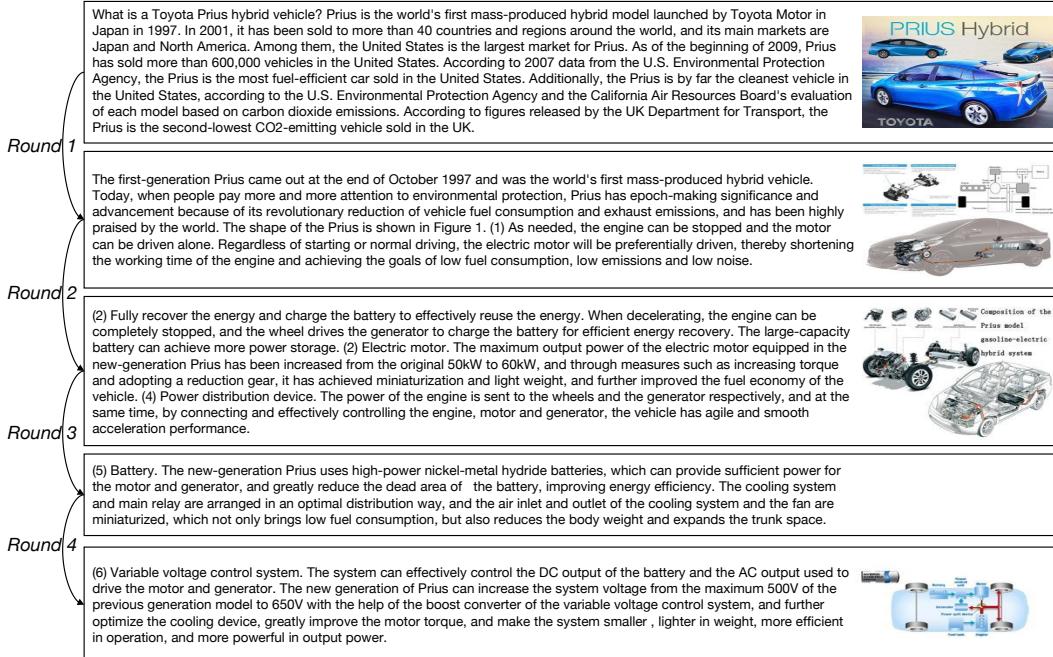


Figure 8: Visualization sample in SeqCIR benchmark.



Figure 9: Visualization samples in CSR benchmark. Each column represents the consecutive pairs.