Visually-Enhanced Phrase Understanding

Tsu-Yuan Hsu[*] **Chen-An Li**[*] **Chao-Wei Huang** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{b08201047,b08902123,f07922069}@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Large-scale vision-language pre-training has exhibited strong performance in various visual and textual understanding tasks. Recently, the textual encoders of multi-modal pre-trained models have been shown to generate high-quality textual representations, which often outperform models that are purely text-based, such as BERT. In this study, our objective is to utilize both textual and visual encoders of multi-modal pre-trained models to enhance language understanding tasks. We achieve this by generating an image associated with a textual prompt, thus enriching the representation of a phrase for downstream tasks. Results from experiments conducted on four benchmark datasets demonstrate that our proposed method, which leverages visually-enhanced text representations, significantly improves performance in the entity clustering task.[1]

1 Introduction

Recent advances in vision-language pre-training have seen the successful alignment of visual and linguistic inputs through the implementation of cross-modal pre-training objectives, such as language modeling and contrastive learning (Lu et al., 2019; Radford et al., 2021). These pre-trained models have shown impressive performance on downstream vision-language tasks, validating their cross-modal capabilities (Su et al., 2019).

While most previous studies focused on multi-modal tasks, researchers have shown that pre-trained cross-modal encoders are equally proficient at uni-modal language understanding, matching the performance of pre-trained text encoders. Lu et al. (2022) were the pioneers in utilizing machine abstract imagination from pre-trained cross-modal encoders, demonstrating improvement on general NLU tasks. Yan et al. (2022) established that the text encoder of CLIP (Radford et al., 2021) surpasses models designed for producing phrase representations, including Phrase-BERT (Wang et al., 2021) and UCTopic (Li et al., 2022a). They hypothesized that the visual supervision during pre-training empowers CLIP to produce visually-grounded phrase representations, beneficial for language-only tasks. Such a phenomenon aligns with neuroscience studies, demonstrating that visual and linguistic semantic representations are coordinated in the human brain (Popham et al., 2021).

Despite the strong performance of the previous method, it only utilized the text encoder of a cross-modal pre-trained model. In contrast, our study aims to exploit its multi-modal representation capacity, incorporating both text and image encoders. We introduce a **visually-enhanced phrase understanding** framework to exploit multiple modalities for uni-modal tasks. Our framework comprises a text-to-image generator and a text-image cross-modal encoder. We employ a text-to-image generator to produce visual cues for a textual candidate. Subsequently, the generated image and the textual prompt are processed by the cross-modal encoder to create visually-enhanced phrase embeddings. Unlike Lu et al. (2022), our method does not require supervised data for downstream tasks, making it more scalable. Our approach also differs from VOKEN (Tan and Bansal, 2020), as they generated visual cues in tokens and processed the signal solely on the language side, whereas we employ representations directly from different modalities. Therefore, our model can capture more abstract concepts from images, enhancing generalizability.

We evaluate our approach on four benchmark phrase understanding datasets. The experiments demonstrate that our proposed visual enhancement significantly outperforms all text-only baselines, demonstrating that abstract visual concepts can provide complementary cues for text understanding.

[*] Equal contribution

[1] Source code: https://github.com/MiuLab/VisualLU

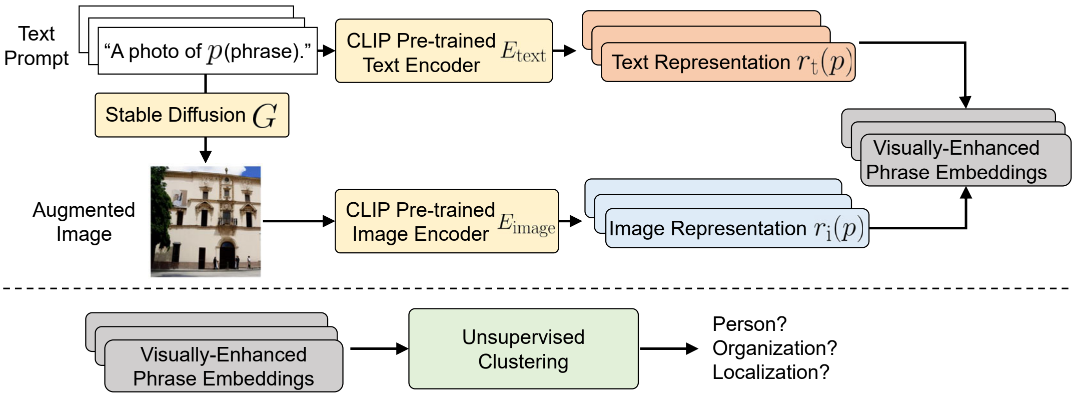

Text Prompt
"A photo of p(phrase)."
CLIP Pre-trained Text Encoder E_{text}
Text Representation $r_{\text{t}}(p)$
Stable Diffusion G
Visually-Enhanced Phrase Embeddings
Augmented Image
CLIP Pre-trained Image Encoder E_{image}
Image Representation $r_{\text{i}}(p)$
Visually-Enhanced Phrase Embeddings
Unsupervised Clustering
Person?
Organization?
Localization?

Figure 1: Illustration of the proposed framework.

2 Method

Our proposed method is illustrated in Figure 1, where we first generate images associated with phrases using a text-to-image diffusion model. Following this, we utilize pre-trained text and image encoders to construct visually-enhanced phrase embeddings for downstream understanding tasks.

2.1 Text-To-Image Model

Recently, text-to-image models have attracted significant interest. Among these, diffusion models have played an important role in text-to-image generation, showing impressive performance. To more effectively generate visual cues associated with texts, this study adopts stable diffusion (Rombach et al., 2022) as our image generation model.

During the training phase, an image autoencoder is trained using an extensive image database. A time-conditional U-Net (Long et al., 2015) forms the core of the diffusion model, learning to denoise image latent representations incrementally.

In the sampling procedure, we first obtain a text prompt and derive a text embedding from the text encoder. Subsequently, we use Gaussian noise as the latent representation, and progressively denoise the latent representation via the diffusion model and a scheduler algorithm. Ultimately, an image is generated by reconstructing the latent representation through the image decoder.

2.2 CLIP (Contrastive Language-Image Pretraining)

CLIP (Radford et al., 2021) is a large-scale vision-language pre-training model using contrastive learning, which achieves remarkable performance in zero-shot image classification tasks. Given a batch of data D, CLIP jointly trains an image encoder and a text encoder to maximize the similarities of $|D|$ paired text-image representations while minimizing the similarities of other $(|D|^2 - |D|)$ unpaired text-image representations. Given the weak alignment between texts and images, this study employs the pre-trained CLIP text encoder E_{text} and image encoder E_{image} to extract meaningful cues from different modalities. Our experiments focus on showing that the pre-trained CLIP encoders provide superior visual enhancement for texts, compared to separately pre-trained text and image encoders.

2.3 Visually-Enhanced Multimodal Representation

Given a text sequence with an entity candidate phrase p, we design our text prompt as "A photo of <p>", a proven effective default template that delivers robust zero-shot classification performance (Radford et al., 2021). As depicted in Figure 1, we initially use the text prompt to generate a text-associated image with the text-to-image model G. Following this, we employ the pre-trained text and image encoders of CLIP to extract corresponding representations $r_i(p)$ and $r_t(p)$ as follows.

$$\begin{aligned} r_t(p) &= E_{text}(\text{“A photo of } p\text{”}) \\ r_i(p) &= E_{image}(G(\text{“A photo of } p\text{”})) \end{aligned}$$

Lastly, we concatenate the two embeddings originating from different modalities to create visually-enhanced phrase embeddings, which potentially capture richer and more comprehensive information and thus benefit downstream tasks.

		CoNLL2003		BC5CDR		W-NUT 2017		MIT-Movie		Average	
		ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Baselines	BERT-base	.394	.021	.711	.201	.252	.026	.589	.014	.486	.065
	BERT-large	.415	.020	.551	.005	.318	.025	.680	.013	.490	.016
	RoBERTa-base	.633	.362	.519	.001	.425	.211	.697	.227	.568	.200
	RoBERTa-large	.601	.241	.744	.294	.379	.057	.541	.005	.566	.149
	LUKE-base	.653	.281	.519	.006	.301	.199	.843	.343	.570	.207
	LUKE-large	.688	.348	.756	.340	.324	.208	.734	.271	.625	.292
	Phrase-BERT (2021)	.619	.339	.597	.061	.423	.246	.914	.559	.638	.301
	UCTopic (2022a)	.682	.335	.933	.677	.287	.140	.807	.307	.677	.365
	+ Contextual Prompt	.759	.425	**.946**	**.710**	.391	.387	.601	.107	.674	.407
	CLIP Text (2022)	.728	.392	.521	.003	**.464**	.320	.784	.358	.624	.268
	+ Contextual Prompt	.743	**.460**	.831	.430	.420	.260	.773	.340	.692	.373
Ours	Proposed Image	.738	.414	.734	.197	.432	.293	.895	.525	.698	.357
	Proposed Text-Image	**.775**	.457	.800	.325	.446	**.338**	**.937**	**.647**	**.740**	**.442**

Table 1: Entity clustering results on four datasets. Proposed Image uses image representation. Proposed Text-Image uses both text and image representations. The best scores are marked in bold and the second-best ones are underlined.

3 Experiments

To evaluate whether our visually-enhanced phrase embeddings provide improved semantic cues, we conduct a series of experiments focused on entity clustering, as our primary task is to categorize entity candidates with similar concepts only based on phrase representations in an unsupervised fashion.

3.1 Setup

Our experiments are conducted on four diverse datasets, each with annotated entities from various domains:

- CoNLL2003 (Sang and De Meulder, 2003) comprises 20,744 sentences, incorporating four types of entities: persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC).
- BC5CDR (Li et al., 2016) is formed from 1,500 PubMed articles and contains chemical and disease entities.
- W-NUT 2017 (Derczynski et al., 2017) is collected from public platforms, including YouTube and Twitter, with a focus on identifying previously unseen entities in emerging discussions. It includes six types of entities.
- MIT-Movie (Liu et al., 2013) contains 12,218 sentences featuring title and person entities.

Following previous research (Xu et al., 2017; Li et al., 2022b; Yan et al., 2022), we implement K-means clustering on the cross-modal representations to perform unsupervised phrase understanding tasks. In this setup, the number of clusters is set to the number of classes present in the dataset. The Hungarian algorithm (Papadimitriou and Steiglitz, 1998) is employed to optimally allocate each cluster to a class.

To evaluate the quality of the representations and compare them fairly with the previous work, we employ accuracy (ACC) and normalized mutual information (NMI) as our evaluation metrics. The results reported are averages over five separate clustering runs. For our proposed image and text-image approaches, we conduct runs over three seeds for diffusion models to generate images.

3.2 Baselines

We position our model in comparison to various language models and phrase understanding models to validate the effectiveness of our cross-modal framework. The used representations are the same as described in the prior work.

- **BERT/RoBERTa** are well-established pretrained language models (Devlin et al., 2019; Liu et al., 2019) capable of distilling intrinsic patterns from input texts into meaningful representations.[2]
- **LUKE** (Yamada et al., 2020) enhances RoBERTa by introducing entity embeddings to the input, as well as an entity-aware attention mechanism.[3]
- **Phrase-BERT** (Wang et al., 2021) refines BERT using a contrastive objective to gen-

[2] We take the embedding of [CLS] with further processing as phrase representations.

[3] We take the last layer of Transformer as our phrase-associated representations.

Proposed	CoNLL2003		BC5CDR		W-NUT 2017		MIT-Movie	
	ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
Image	.738±.025	.414±.028	.734±.033	.197±.069	.432±.024	.293±.035	.895±.034	.525±.056
Text-Image	.775±.009	.457±.016	.800±.031	.325±.070	.446±.015	.338±.015	.937±.001	.647±.013

Table 2: Entity clustering results with three diffusion model runs.

Text Encoder	Image Encoder	CoNLL2003		BC5CDR		W-NUT 2017		MIT-Movie		Average	
		ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
RoBERTa-base	-	.633	.362	.519	.001	.425	.211	.697	.227	.568	.200
-	ViT-B/32	.629	.343	.668	.109	.380	.238	.895	.523	.643	.303
RoBERTa-base	ViT-B/32	.656	.361	.668	.109	.386	.237	.894	.521	.651	.307
CLIP Text	-	.728	.392	.521	.003	**.464**	.320	.784	.358	.624	.268
-	CLIP ViT-B/32	.749	.423	.757	.197	.426	.279	.928	.600	.710	.375
CLIP Text	CLIP ViT-B/32	**.771**	**.451**	**.844**	**.406**	.434	**.332**	**.935**	**.641**	**.746**	**.458**

Table 3: Comparison of the separately pre-trained encoders and CLIP over one diffusion model run. CLIP ViT-B/32 is the image encoder of CLIP where the architecture is the same as ViT-B/32. Best results are marked in bold.

erate more powerful phrase representations.[4]

- **UCTopic** (Li et al., 2022a) employs an unsupervised contrastive learning strategy, with LUKE serving as the foundational model, to create robust and context-aware embeddings.[5]
- **CLIP Text** (Yan et al., 2022) leverages the text encoder of CLIP for understanding.[6]

3.3 Results

The evaluation results are presented in Table 1. Our proposed visually-enhanced representations outperform all baselines on the CoNLL2003 and MIT-Movie datasets, while achieving competitive performance on the BC5CDR and W-NUT 2017 datasets. Moreover, solely utilizing image representations encoded from generated images yields a higher average ACC than all the baselines. This suggests that the visual signal offers valuable cues for enhanced phrase understanding. Hence, we conclude that integrating different modalities can effectively augment phrase representations. For a more granular understanding, we provide detailed scores across multiple turns in Table 2. The lower standard deviation of our proposed text-image approach indicates superior stability.

3.4 Analysis of Different Encoders

To further investigate whether the CLIP encoders, pre-trained jointly, are more effective for visual enhancement, we compare them with image and text encoders that have been pre-trained individually. Table 3 presents the experimental results, where we substitute the text and image encoders of CLIP with RoBERTa-base and ViT-B/32 respectively. We notice that phrase representations augmented by ViT-B/32 outperform textual representations, which suggests the richness of information drawn from multiple modalities. It is evident that CLIP encoders surpass individually pre-trained encoders, implying that text and image encoders, when pre-trained together, can more effectively enrich phrase representations by integrating text and image at the representation level.

3.5 Contextual Prompt

Previous work (Yan et al., 2022) demonstrated that enriching phrase candidates with a large pre-trained language model can yield more domain-specific keywords for textual prompts. Specifically, given a phrase p, the prompt "p is a `[MASK]`" is fed into a language model, which in turn returns the top K predictions $\{m_1, m_2, \ldots, m_K\}$ for the `[MASK]` token. Subsequently, we formulate the contextual prompt as "A photo of p, a $m_1, m_2, \ldots, m_K$." In this paper, we set K to 3 for the contextual prompts. Table 1 shows that the addition of such contextual prompts enhances the performance of text-only baselines.

We further probe into whether a contextual prompt can boost our performance and present the results in Table 4. Our observation is that utilizing contextual prompts for text embeddings yields comparable performance, indicating that our visual cues already encompass the domain-specific signal. We hypothesize that generating images from

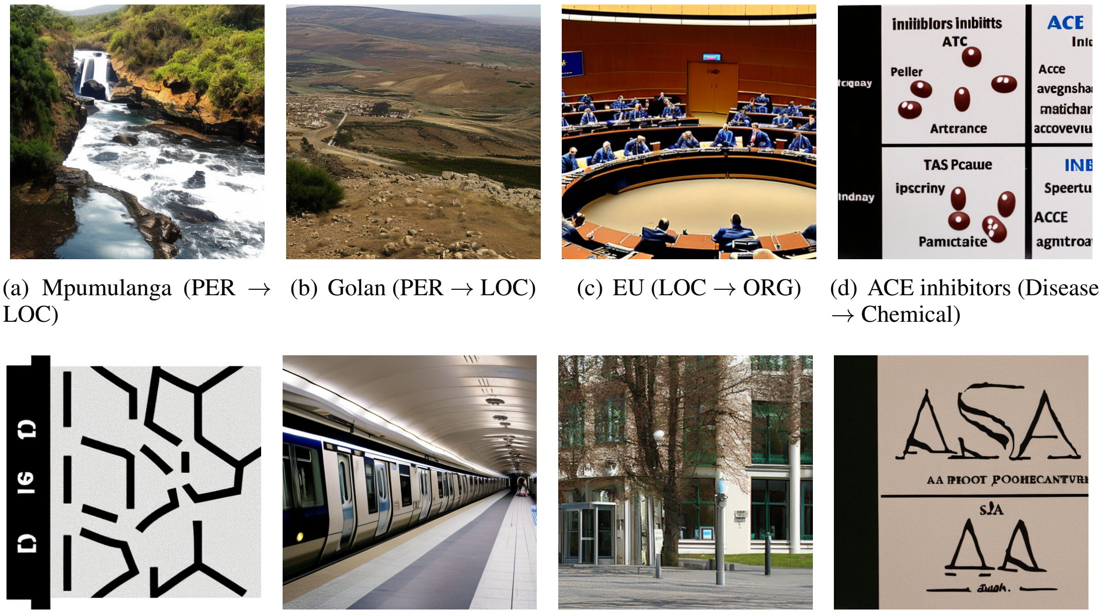
[4] We take the average of the last layer in Transformer as our phrase-associated representations.

[5] We take the pooling of the entity-associated vectors based on `https://github.com/JiachengLi1995/UCTopic`.

[6] We take `[EOT]` of the last Transformer layer's output as phrase representations.

Approach	Text Input	Image Input	CoNLL2003		BC5CDR		W-NUT 2017		MIT-Movie		Average	
			ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Proposed Text-Image	*Vanilla*	$G(\textit{Vanilla})$	.771	.451	.844	.406	.434	.332	.935	.641	.746	.458
Proposed Text-Image	*Contextual*	$G(\textit{Vanilla})$	.766	.445	.853	.429	.424	.308	.937	.643	.745	.456
Proposed Text-Image	*Contextual*	$G(\textit{Contextual})$	.742	.406	.872	.503	.409	.236	.888	.487	.728	.408
CLIP Text	*Contextual*	-	.743	.460	.831	.430	.420	.260	.773	.340	.692	.373

Table 4: The utility of contextual prompt. *Vanilla*: “A photo of p.”; *Contextual*: “A photo of p, a m_1, m_2, m_3.” (p is the entity and m_1, m_2, m_3 are the keywords of p.)



(a) Mpumulanga (PER → LOC)

(b) Golan (PER → LOC)

(c) EU (LOC → ORG)

(d) ACE inhibitors (Disease → Chemical)

(e) p-choloroaniline (Disease → Chemical)

(f) The Metro (corporation → location)

(g) BAYERISCHE VEREINSBANK (ORG → LOC)

(h) SA (group → group)

Figure 2: Our generated images with the associated phrases.

contextual prompts may introduce more noise, resulting in difficulty encoding effective visual representations for phrase understanding. Notably, our baseline setting already achieves significantly improved performance compared with earlier work utilizing additional keywords, demonstrating the informativeness of our cross-modal representations.

3.6 Qualitative Analysis

To further examine how our visual cues enhance text understanding, we present several generated images along with their understanding results in Figure 2. Previous work, CLIP Text, incorrectly classifies “Mpumulanga” and “Golan” as PER (persons). However, with the visual cues generated in our model, shown in Figure 2(a-b), we can correctly classify them as LOC (locations). The images generated by our model, displayed in Figure 2(c-f), further enrich the phrase representations and better understand the concepts. This demonstrates the effectiveness of our multi-modal framework.

However, there are cases where the generated image may lead to incorrect categorization, as is the case with “BAYERISCHE VEREINSBANK” in Figure 2(g). The image misled the categorization process, changing the cluster from the correct classification (ORG, or organization) to an incorrect one (LOC, or location). Figure 2(h) displays an instance where the generated image does not provide useful visual information for an unusual entity, and the incorrect classification (group) persists. Therefore, there is still room for enhancement in future work.

4 Conclusion

This work presents a multi-modal framework that leverages a text-to-image model to bridge between language and visual modalities for enhancing text comprehension. The model effectively transforms text inputs into coherent images, enriching phrase representations by merging outputs from different modalities. Experimental results show our framework surpassing robust phrase understanding models across diverse domains.

Limitations

Due to the maximum input length constraint of both the CLIP text encoder and the text-to-image model, we are unable to process long texts. We are interested in exploring alternative prompt configurations to circumvent this limitation. Our methodology is readily extendable to these settings, making it an intriguing area of study.

Ethics Statement

Our approach leverages a pre-trained text-to-image model to visually enhance representations. However, the text-to-image model may carry over biases and improper content from its training data. This necessitates additional analyses to safeguard against any undue influence of these biases on our method.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the Young Scholar Fellowship Program by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 111-2628-E-002-016.

References

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiacheng Li, Jingbo Shang, and Julian McAuley. 2022a. UCTopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6169.

Jiacheng Li, Jingbo Shang, and Julian McAuley. 2022b. UCTopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6169, Dublin, Ireland. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass. 2013. Query understanding enhanced by hierarchical parsing structures. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 72–77. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4392–4402.

Christos H Papadimitriou and Kenneth Steiglitz. 1998. *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference* *on Computer Vision and Pattern Recognition*, pages 10684–10695.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved phrase embeddings from bert with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10837–10851.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

An Yan, Jiacheng Li, Wanrong Zhu, Yujie Lu, William Yang Wang, and Julian McAuley. 2022. CLIP also understands text: Prompting clip for phrase understanding. *arXiv preprint arXiv:2210.05836*.

A Datasets

- **CoNLL2003**: This dataset comprises 20,744 sentences with four distinct types of entities - persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC). Our experiments utilize 30,027 entities that are labeled as PER, ORG, or LOC.
- **BC5CDR**: This dataset features 1,500 PubMed articles that are populated with chemical and disease entities, adding up to a total of 28,354 entities.
- **W-NUT 2017**: This dataset is an accumulation of data collected from public platforms like YouTube and Twitter, with a focus on distinguishing previously unseen entities within emerging discussions. It includes six types of entities: person, location, group, corporation, creative_work, and product. The dataset contains a total of 3,890 entities.
- **MIT-Movie**: This dataset includes 12,218 sentences populated with title and person entities, accounting for a total of 9,920 entities.

B Implementation Details

In our work, we use the Huggingface models to generate all the representations:

- **BERT/RoBERTa**: We take `pooler_output` as the representations, where `pooler_output` is the classification token after processing through a linear layer and an activation function.[7] The linear layer weights are learned by next sentence prediction during pre-training.
- **LUKE**: `entity_last_hidden_states` is used as the representation, which is the last hidden states of the input entity.[8]
- **Phrase-BERT**: Phrase representations can be easily acquired by calling `model.encode()`.[9]
- **UCTopic**: We obtain the phrase representations with the released source code.[10]
- **CLIP**: `pooler_output` is taken as the representation for both the text encoder[11] and the image encoder.[12]

C Pre-trained Models

For the pre-trained CLIP model, we adopt the version ViT-B/32, which consists of a ViT-B/32 image encoder and a 12-layer Transformer text encoder.

[7] https://huggingface.co/docs/transformers/main_classes/output#transformers.modeling_outputs.BaseModelOutputWithPooling.pooler_output

[8] https://huggingface.co/docs/transformers/model_doc/luke#transformers.LukeModel

[9] https://huggingface.co/whaleloops/phrase-bert

[10] https://github.com/JiachengLi1995/UCTopic/blob/main/clustering.py#L43

[11] https://huggingface.co/docs/transformers/model_doc/clip#transformers.CLIPTextModel

[12] https://huggingface.co/docs/transformers/model_doc/clip#transformers.CLIPVisionModel

Approach	Inference steps	**CoNLL2003**		**BC5CDR**		**W-NUT 2017**		**MIT-Movie**		**Average**	
		ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
Proposed Image	10	.746	.420	.722	.153	.446	.308	.932	.615	.712	.374
	30	.745	.420	.759	.198	.435	.285	.929	.604	.717	.377
	50	.749	.423	.757	.197	.426	.279	.928	.600	.715	.375
Proposed Text-Image	10	.783	.474	.824	.348	.427	.315	.940	.652	.744	.447
	30	.771	.450	.849	.430	.443	.341	.937	.646	.750	.467
	50	.772	.451	.844	.406	.434	.332	.935	.641	.746	.458

Table 5: Comparison on different inference steps of stable diffusion. The reported numbers are run over one Stable Diffusion seed.

For the text-to-image diffusion model, we use stable diffusion v2-base[13] trained on the subset of LAION-5B (Schuhmann et al., 2022) in our experiments.

D Inference Details

We conduct our experiments on single V100 GPU.

- Generation time of stable diffusion v2-base with respect to inference steps is elaborated in Appendix E.
- Each clustering experiment takes no more than 10 minutes to run.

D.1 Licenses

- BERT (Apache License Version 2.0)
- RoBERTa (MIT License)
- LUKE (Apache License Version 2.0)
- Phrase-BERT (T License)
- UCTopic (MIT License)
- CLIP (MIT License)
- vit-base-patch32-224-in21k (Apache License Version 2.0)
- stable-diffusion-2 (CreativeML Open RAIL++-M License)

E Efficiency vs. Efficacy

Results over different inference steps of stable diffusion v2-base are shown in Table 5. It took 0.84 seconds per image for inference step 10, 2.02 seconds per image for inference step 30, and 3.24 seconds per image for inference step 50. The balance between efficiency and efficacy depends on application usage.

[13] https://huggingface.co/stabilityai/stable-diffusion-2-base