# KB-VLP: Knowledge Based Vision and Language Pretraining

Kezhen Chen*, Qiuyuan Huang+, Yonatan Bisk$, Daniel McDuff+, Jianfeng Gao+
Northwestern University*, Microsoft Research+, Carnegie Mellon University$

ICML

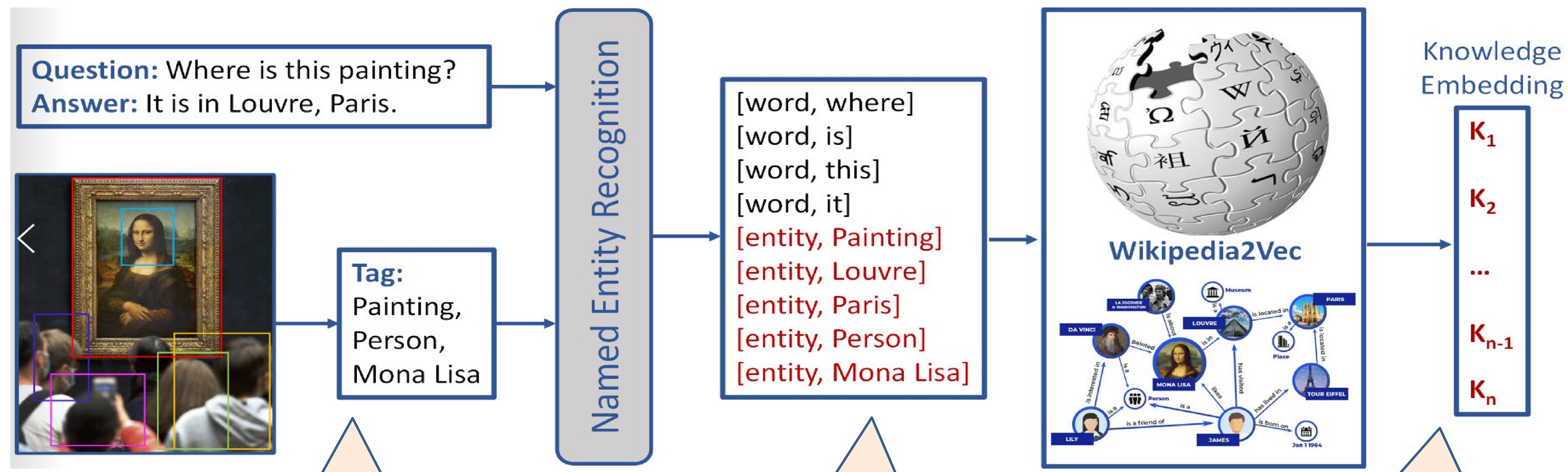
## Motivation

- Vision-Language Pretraining (VLP) has received increasing attention
- Existing models ignore external knowledge
- Models should consider both
  - multiple modalities
  - rich structural information in knowledge
- Knowledge embeddings in pretraining improve
  - Standard VL tasks
  - Commonsense tasks

## Contributions

- Knowledge-based self-supervised pretraining
- Use Wikidata to get external knowledge
- Experiments and analysis demonstrate the effectiveness of our approach.

## Extracting Knowledge

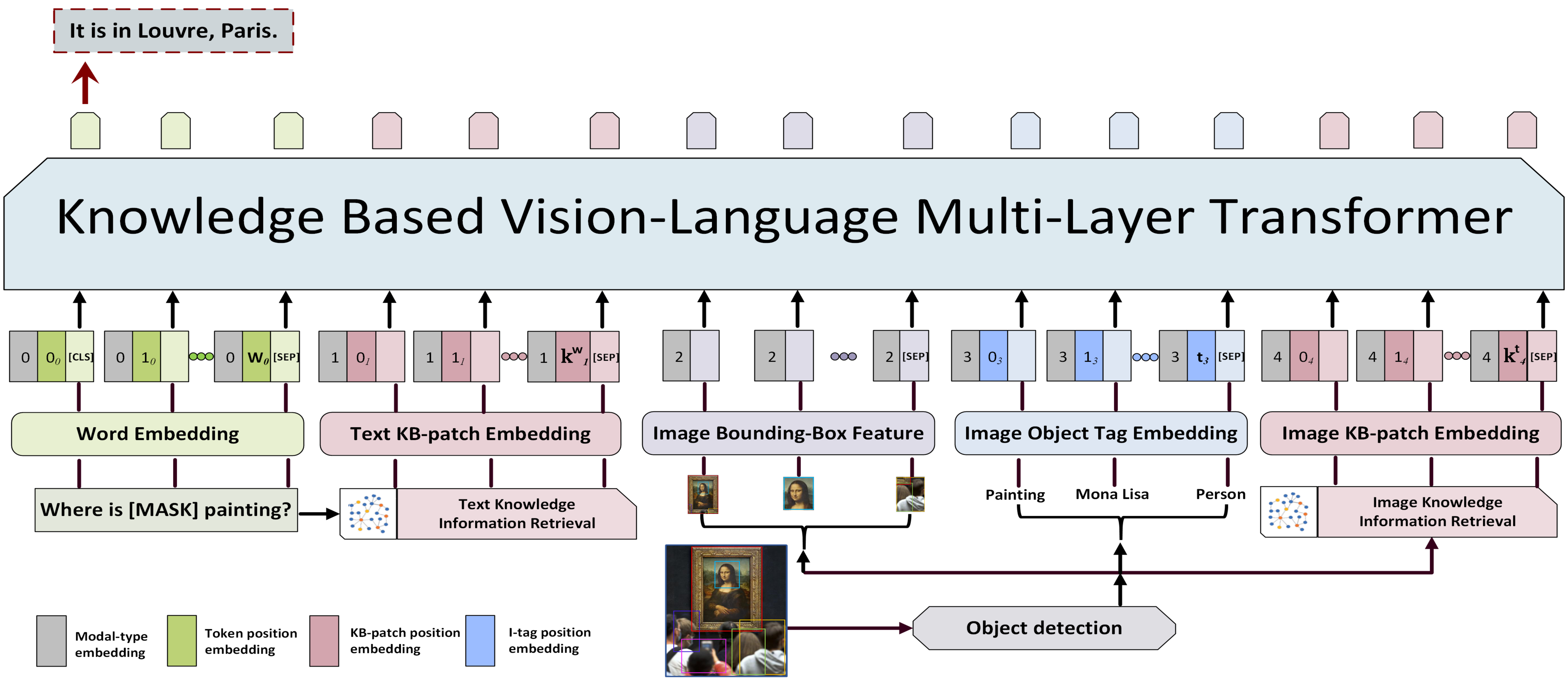



Named Entity Recognition performed on text and image tags to detect entities in Wikidata KB

Text tokens are separated into a set of entities, and a set of English words.

Wikipedia2Vec is used to generate a pretrained embedding vector for each entity and word

## Bitmap to Structured Representations




***Pretraining Approach***

Input: (Text, Image Tags, Image Bounding-Box Features, KB-patch Embeddings (Text KB-patch and Image KB-patch))

KB-patches are generated via knowledge extraction on text and image tags

Sequence-Level Loss

- One of the elements in input tuple is replaced by the element of a random document
- We use a four-way contrastive loss, asking the model to predict which element is replaced

Token-Level Loss

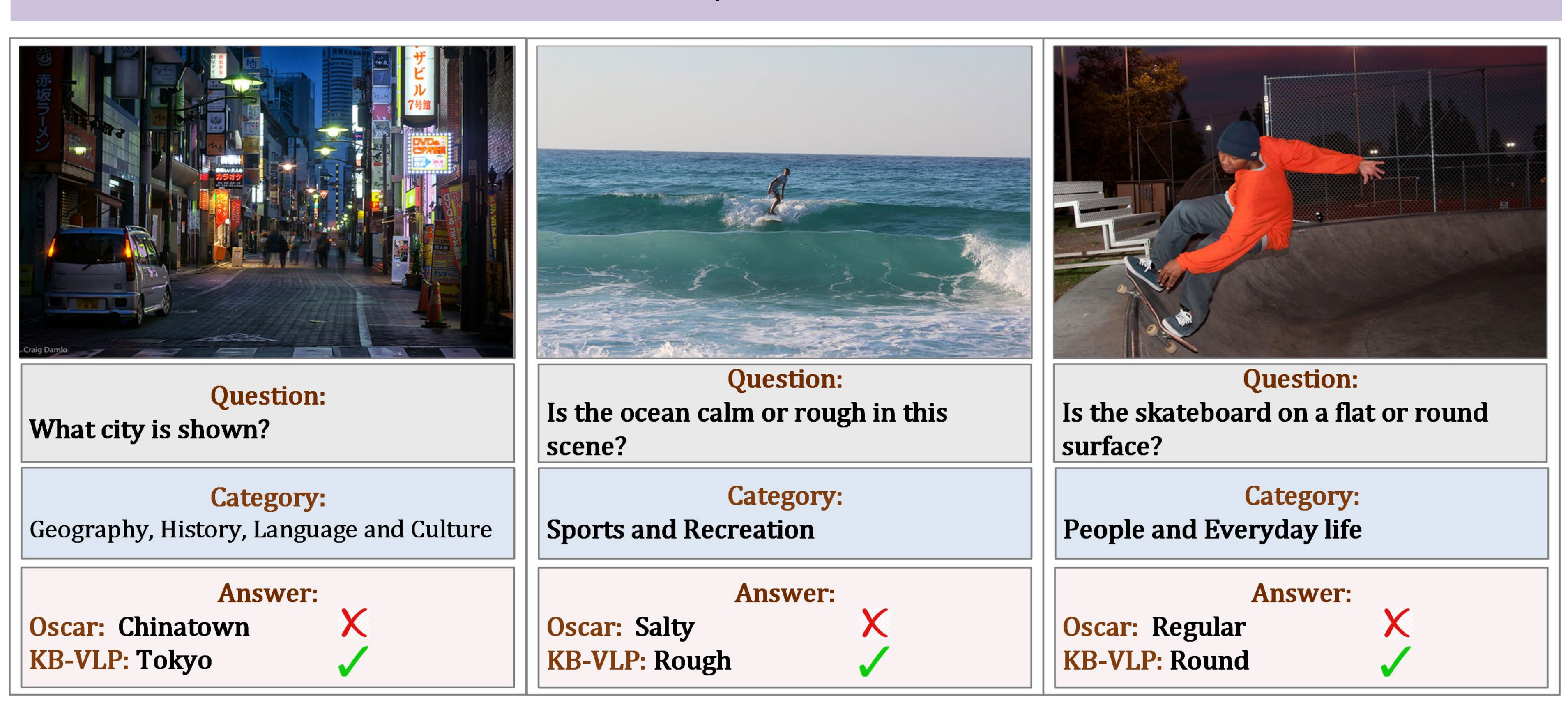
- Text elements: Masked Token Loss of BERT
- For knowledge embeddings, each embedding has a chance to be replaced by a random embedding:
  - On text knowledge, the model predicts whether the embedding is the original one or replaced
  - On image knowledge, the model predicts the original entities from a subset

## Experiments

| Model | VQA Dev | VQA Test-std | $NLVR^2$ Dev | $NLVR^2$ Test-P | GQA Dev | GQA Test-std | OK-VQA R@1 | OK-VQA R@5 | OK-VQA R@10 | OK-VQA ACC-full |
|---|---|---|---|---|---|---|---|---|---|---|
| NSM (Drew A. Hudson, 2019) | – | – | – | – | – | **63.17** | – | – | – | – |
| ViLBERT (Lu et al., 2019a) | 70.63 | 70.92 | – | – | – | – | – | – | – | – |
| VL-BERT (Su et al., 2020) | 70.50 | 70.83 | – | – | – | – | – | – | – | – |
| VisualBERT (Li et al., 2019) | 70.80 | 71.00 | 67.40 | 67.00 | – | – | – | – | – | – |
| LXMERT (Tan & Bansal, 2019) | 72.42 | 72.54 | 74.90 | 74.50 | 60.00 | 60.33 | – | – | – | – |
| 12-in-1 (Lu et al., 2019b) | 73.15 | – | – | 78.87 | – | 60.65 | – | – | – | – |
| UNITER-B (Chen et al., 2019) | 72.27 | 72.46 | 77.14 | 77.87 | – | – | – | – | – | – |
| Oscar-B (Li et al., 2020b) | 73.16 | 73.44 | 78.07 | 78.36 | 61.19 | 61.58 | 34.50 | 63.95 | 73.47 | 30.07 |
| **KB-VLP** (ours) | **73.63** | **73.89** | **78.23** | **78.44** | **62.40** | 62.57 | **41.10** | **72.05** | **82.28** | **33.41** |

- KB-VLP is finetuned on four tasks – VQA, NLVR2, GQA, and OK-VQA
- On VQA, NLVR2 and GQA, KB-VLP outperforms baseline VLP models
- On OK-VQA, KB-VLP has significant improvements than Oscar
- The results show:
  - Using knowledge in pretraining improves standard VL tasks
  - Using knowledge in pretraining enhances commonsense tasks

## Experiments




Three examples from OK-VQA that KB-VLP model generates correct answer, but Oscar does not. Comparing the generated answers from KB-VLP and Oscar indicates that Oscar model is limited to visual detection and KB-VLP has stronger reasoning and understanding ability.