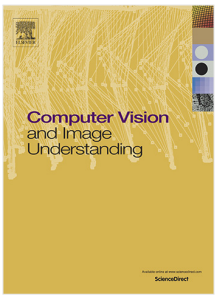
Journal Pre-proof



Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval

Mustafa Shukor, Nicolas Thome, Matthieu Cord

PII: S1077-3142(24)00152-8
DOI: https://doi.org/10.1016/j.cviu.2024.104071
Reference: YCVIU 104071

To appear in: *Computer Vision and Image Understanding*

Received date : 14 July 2023
Revised date : 26 May 2024
Accepted date : 3 July 2024

Please cite this article as: M. Shukor, N. Thome and M. Cord, Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval. *Computer Vision and Image Understanding* (2024), doi: https://doi.org/10.1016/j.cviu.2024.104071.

Revised manuscript (clean version)

ELSEVIER

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval

Mustafa Shukor[a,**], Nicolas Thome[a], Matthieu Cord[a,b]

[a]*Sorbonne University, ISIR, Paris, France*
[b]*Valeo.ai, Paris, France*

ABSTRACT

Vision-Language Pretraining (VLP) and Foundation models have been the go-to recipe for achieving SoTA performance on general benchmarks. However, leveraging these powerful techniques for more complex vision-language tasks, such as cooking applications, with more structured input data, is still little investigated. In this work, we propose to leverage these techniques for structured-text based computational cuisine tasks. Our strategy, dubbed VLPCook, first transforms existing image-text pairs to image and structured-text pairs. This allows to pretrain our VLPCook model using VLP objectives adapted to the strutured data of the resulting datasets, then finetuning it on downstream computational cooking tasks. During finetuning, we also enrich the visual encoder, leveraging pretrained foundation models (*e.g.* CLIP) to provide local and global textual context. VLPCook outperforms current SoTA by a significant margin (+3.3 Recall@1 absolute improvement) on the task of Cross-Modal Food Retrieval on the large Recipe1M dataset. We conduct further experiments on VLP to validate their importance, especially on the Recipe1M+ dataset. Finally, we validate the generalization of the approach to other tasks (*i.e*, Food Recognition) and domains with structured text such as the Medical domain on the ROCO dataset. The code will be made publicly available.

1. Introduction

Vision-Language Pretraining (VLP) (Chen et al., 2020; Su et al., 2019; Li et al., 2021a) has become the general recipe to attain SoTA results on downstream tasks, with the key success is learning a shared latent space where all modalities are aligned. This paradigm helps to overcome the human labor associated with designing a task or domain customized approaches, and pushes towards more simplification, by unifying the model, training objective and input/output format (Wang et al., 2022b,a). As going large scale is an important ingredient to push the performance limits, we have witnessed recently a lot of work going in this direction, leading to what so-called foundation models (Alayrac et al., 2022; Yu et al., 2022; Radford et al., 2021; Chen et al., 2022).

However, these approaches are still evaluated on general benchmarks (*e.g.*, VQA (Antol et al., 2015), Image-Text Retrieval (Plummer et al., 2015)), to the detriment of more complex albeit important tasks. These benchmarks highly resemble the pretraining data, in terms of image distribution, text format, length and structure. Similarly, existing Foundation models have shown great transfer capabilities to several downstream tasks (Liu et al., 2023; Gao et al., 2024; Azad et al., 2023; Celaj et al., 2023), however, it is still also unclear how they perform beyond common tasks. The key stumbling block to leverage VLP and Foundation models for such domains, is the complex input that is hard to digest. In particular the tasks involving images with associated text that goes beyond simple image caption, to richer, longer and structured text.

In this work, we question how to leverage VLP and existing Foundation models for tasks requiring structured text. As image-text alignment has proven to be successful for multimodal tasks, we focus on Image-Text Retrieval being one of the best benchmarks to evaluate such alignment. To validate the proposed approach, we consider the traditional task of on Cross-Modal Food Retrieval (Salvador et al., 2017), aiming at bridging the gap between VLP and Computational Cooking.

Computational Cooking or Food applications (Martinel et al., 2015; Ofli et al., 2017; Salvador et al., 2017) are one of the important applications that fit very well in this marginalized list,

**Corresponding author:
e-mail: mustafa.shukor@sorbonne-universite.fr (Mustafa Shukor)

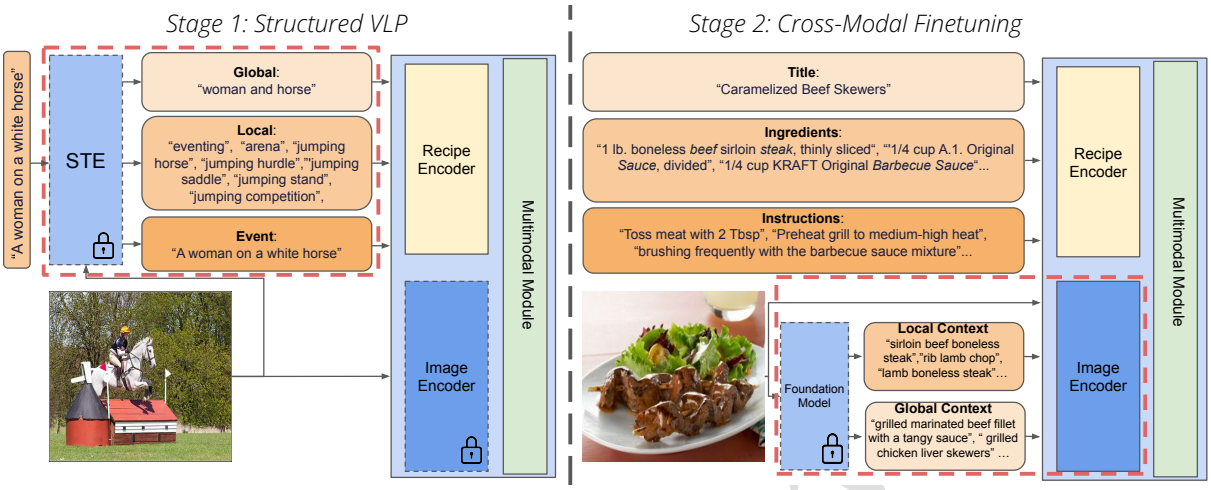


Fig. 1: VLPCook framework with 2 sequential stages. Stage 1 (left) or VSLP (Sec. 3.1): the Structured Text Extraction (STE) module transforms the caption to a structured recipe-like input that is used to pretrain the model on a large corpus of structured text and images. Stage 2 (right) or Cross-Modal Finetuning (Sec. 3.2): we leverage existing foundation models to enrich the vision encoder with local and global textual context. Main contributions are highlighted in red. The lock symbol means the model is frozen.

with no existing work to bridge the gap with VLP. In particular, Cross-Modal Food Retrieval (Salvador et al., 2017, 2021; Shukor et al., 2022b) which is the current main benchmark to assess the model performance on computational cooking. The images are of different food plates with high inter and low intra category similarity. The text, consists of the corresponding recipe that is composed of 3 entities; title (global description), ingredients (local descriptions, objects or entities that might be seen or not) and instructions (events that we generally see only their effects or final results).

As the main hurdle to enable VLP for food models is the input data, we choose to adapt the input data to be compatible, structurally and semantically, to fit in these models. In addition, we exploit existing large scale Vision-Language Models (VLMs), to guide the vision encoder with structured context. This guidance is through region-level or local context (e.g. ingredients), and image-level or global context (*e.g.* titles). Our approach, dubbed VLPCook, consists of 2 stages; (1) Vision and Structured-Language Pretraining (VSLP) of the model on the created structured text, then (2) Cross-Modal Finetuning guided by foundation models. The approach is illustrated in Fig. 1.

Our main contributions can be summarized as follows: **a)** We propose a new approach for transforming existing datasets of image-text pairs to datasets of image and structured-text pairs, and show that VLP on such datasets gives significant improvement. **b)** We propose a new model that leverages existing pretrained foundation models to inject structured local and global textual context to guide the visual encoder.

To validate the work, we conduct an extensive experimental study on the challenging task of Cross-Modal Food Retrieval, which leads to the following interesting outcomes: **a)** VLPCook outperforms significantly other SoTA on the Recipe1M dataset, with absolute improvement of +3 and +3.3 of R@1 on the 1k and 10k setups respectively. **b)** The first work showing the effectiveness of VLP in the cooking context, after experimenting with different kinds of existing food approaches. **c)** Despite what was reported (Marin et al., 2019) on the poor generalization from Recipe1M+ to Recipe1M, we show that pretraining on this large dataset can unlock its potential, and lead to large improvement of +2.4 R@1 on Recipe1M test set. **d)** Contrary to recent findings showing that foundation models can attain SoTA on standard benchmarks (*e.g.* VQA v2, COCO retrieval), we show that finetuning these models lag significantly behind SoTA on the underlying task of Cross-Modal Food Retrieval. **e)** We validate the generalization of the work to other tasks (*i.e.*, Food Recognition) and domains, such as the Medical domain, showing significant improvement over baselines.

2. Related Work

Vision and Language Pretraining (VLP). Vision and Language Pretraining (VLP) (Chen et al., 2020; Su et al., 2019) aims at learning vision-language representation by pretraining on datasets of images and texts ((Sharma et al., 2018; Schuhmann et al., 2021; Radford et al., 2021)). The model is then evaluated on several downstream tasks such as VQA (Antol et al., 2015), and image-text retrieval (Plummer et al., 2015). This line of research has shown promising success in the last few years, leading to state of art (SoTA) results (Li et al., 2021a; Dou et al., 2022; Li et al., 2022a) compared to task-customised models, and providing modular encoders that are seamlessly used in a variety of ways. Besides several other improvements, the major ones have been either in the architectural design, or the pretraining objectives. On the model side, we have models with separate vision and language encoders (*e.g.*, CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021)), that are fast at inference but requires large datasets to train, and heavy fusion models which use a cross modal interaction module (Dou et al., 2021; Li et al., 2021a; Shukor et al., 2022a) and achieve SoTA results while training on reasonably sized datasets. On the learning side, the

main training objectives can be categorised into contrastive (ITC (Radford et al., 2021), ITM (Chen et al., 2020)) and masked predictions (MLM (Devlin et al., 2018), MIM (Shukor et al., 2022a; Dou et al., 2022)). The models that work best are those that combine several objectives.

Leveraging Foundation Models. Foundation models (Radford et al., 2021; Singh et al., 2021; Alayrac et al., 2022; Zhang et al., 2022) are general models that can be adapted to many unimodal and multimodal tasks. In spite of being successful, due to the need for huge resources to train these models from scratch, researchers and practitioners have leveraged them, without the burden of retraining; such as initialization and finetuning (Shukor et al., 2022b; Shen et al., 2022), as frozen modules (Shukor et al., 2023; Ramesh et al., 2022; Couairon et al., 2022), enriching the input (Sara et al., 2022) and extracting visual concepts (Shukor et al., 2022a). In our work, we leverage existing pretrained foundation models to extract different aspects of textual contexts to enrich the visual representation.

Food Applications and Learning from Sructured Data. Many work have been proposed in the recent years for food tasks, such as food categorization (Bossard et al., 2014), calorie estimation (Myers et al., 2015), image generation (Zhu and Ngo, 2020) and cross modal retrieval (Salvador et al., 2017). Since the inception of large scale food datasets such as Recipe1M (Salvador et al., 2017) followed by Recipe1M+ (Marin et al., 2019) the task of cross-modal retrieval have gained a lot of attention (Li et al., 2024; Song et al., 2023; Huang et al., 2023; Salvador et al., 2021; Shukor et al., 2022b). In terms of performance and architectural designs, cross modal food retrieval work can be divided into transformer-based (Salvador et al., 2021; Guerrero et al., 2021; Shukor et al., 2022b; Papadopoulos et al., 2022) or transformer-free (Salvador et al., 2017; Carvalho et al., 2018; Fain et al., 2019) approaches, with a significant improvements of the former. Specifically, on the vision side, ViT is used as an image encoder, and on the recipe side, standard (Guerrero et al., 2021) or hierarchical transformers (Salvador et al., 2021; Shukor et al., 2022b) are adopted. In terms of training objectives, almost all approaches use triplet loss (Weinberger et al., 2005; Ding et al., 2015) in addition to some regularization such as semantic triplet (Carvalho et al., 2018; Shukor et al., 2022b), embedding classification (Salvador et al., 2017), adversarial losses (Wang et al., 2019) and multimodal regularization with image-text matching objective (Shukor et al., 2022b). In addition to food applications, learning from structured texts and images has been investigated in several domains and tasks, such as Medical applications (Pelka et al., 2018), News applications (Biten et al., 2019), Multimedia Event extraction (Li et al., 2020b,a) and Situation Recognition (Suhail and Sigal, 2019; Cooray et al., 2020). In the context of VLP, few work have been recently proposed (Li et al., 2022b,c), however, they do not consider the case of structured text as input during test and focus on learning a structural representations.

3. VLPCook

Overview: We introduce VLPCook, the first work trying to bridge the gap between VLP and the Computational Cooking domain. VLPCook proposes a novel pretraining pipeline that tackles the issues of complex cooking inputs, and a finetuning framework that leverages this pretraining and foundation models for cooking tasks, such as the task of Cross-Modal Food Retrieval. VLPCook consists in 2 stages: (1) Vision and Structured-Language Pretraining (VSLP in Sec. 3.1); to perform VLP relevant to complex cooking recipes, we transform the image captions (in existing image-text pairs datasets) to structured text, and form new datasets of image and structured text pairs. This allows us to benefit from a large-scale VLP adapted to the specificity of cooking datasets. (2) Cross-Modal Finetuning (Sec. 3.2); on the downstream cooking task, where we leverage existing foundation models, without any retraining, to contextualize the visual encoder with local and global textual context. The approach is illustrated in Fig. 1. As our goal is to leverage VLP and foundation models and show their benefits for the cooking domain, we decide to build our approach on top of recent SoTA food models and keep as much as possible the same model architecture/finetuning objectives.

Background on VLP: VLP consists of pretraining Vision-Language models on large datasets of image-text pairs, then finetuning on several multimodal downstream tasks. Several pretraining objectives are used in VLP. Here we focus only on 2 of them; Image-Text Contrastive (ITC) and Image-Text Matching (ITM):

ITC: several ITC losses have been proposed, such as InfoNCE (Oord et al., 2018) and triplet loss (Ding et al., 2015; Weinberger et al., 2005). In this work, we use a triplet loss on top of the unimodal encoders. On one hand, we pull the image embedding to be close to the corresponding recipe embedding, and vice versa, and on the other hand, we push far away the embeddings of different recipes. ITC is used to globally align both modalities, which is important for tasks such as cross-modal retrieval.

ITM: is a binary classification loss to train the model to predict matched image-text pairs (Chen et al., 2020). This loss is applied on top of the multimodal module (*e.g.*, transformer decoder) and aims to learn more fine-grained interaction between modalities.

3.1. Vision and Structured-Language Pretraining (VSLP)

Existing VLP approaches use image captions; a sentence describing generally the image. However, image captions are not directly aligned with some domains such as Food applications. Specifically, image-captions generally contain one sentence describing globally the image, while recipes are longer (> 200 words), with a richer description, including global (title), local (ingredients), and structured (hierarchical) information.

Here we focus on computational cooking tasks that require such complex text input. The text or the recipe consists of different elements, forming a hierarchical structure; global information about the image (*e.g.*, title), local information (*e.g.* ingredients) and the interaction between different entities (*e.g.* instructions). The text is long (*e.g.* more than 10 ingredients/instructions) and rich, as it contains very specific details (*e.g.* ingredients name and quantity). Recent food models have dedicated recipe encoders (Salvador et al., 2021; Shukor et al., 2022b) to exploit such structure. They use several stages of transformers: one for each ingredient/instruction (T), another for the list of ingredients/instructions (HT), and the last stage with

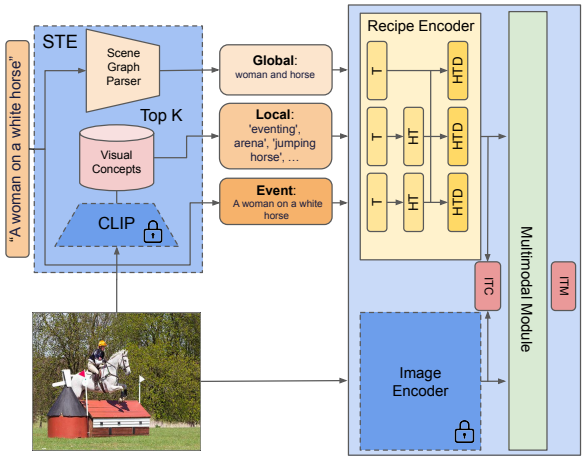


Fig. 2: Illustration of our VSLP (Stage 1 of VLPCook). To enable VLP for food models, image-text pairs are transformed to image and structured-text pairs, that are compatible with hierarchical recipe encoders. The Structured Text Extraction (STE) module generates 3 entities; (a) global description ("title") using SGP, local descriptions ("ingredients") using CLIP-based retrieval, and the "event" ("instructions") which can be simply the caption. During VLP, we optimize ITC and ITM losses and keep the vision encoder frozen.

transformer decoders (HTD) that take the tokens of one entity as query and the tokens of other ones as keys and values (Fig.2).

To bridge this gap between VLP and the food domain, we propose first to create datasets of structured image-text pairs, then use them to pretrain food models. This stage is illustrated in Fig. 2.

From Image Captions to Structured Text (Recipe-fying the captions): we propose a new approach to transform existing image captions, in existing datasets of image and text pairs, to richer and structured text. Transforming existing datasets helps us to leverage large scale ones, which is cheaper than creating large scale datasets of image-recipe pairs from scratch. We make the analogy between the obtained text and recipes and detail the process in the following:

Global information (Title): as the caption describes globally the image, we use it to extract the title. However, it may also include some unnecessary details to be considered for the title, as well as noise (especially for datasets scraped from internet). We filter out the caption and keep the main elements, we extract only the objects using Scene Graph Parsing (SGP) (Schuster et al., 2015) techniques and assemble them with a simple "and" (*e.g.*, title: Woman and Piano and stage).

Local information (Ingredients): here, local entities or objects in the image should be included. As captions usually does not contain many details, we leverage additional sources of information to extract all relevant, seen or unseen, objects in the image. To this end, we use existing foundation models, without retraining them, as they enjoy good generalization capabilities on different domains and tasks, to retrieve the closest entities. Specifically, these entities are retrieved from a database that contains all objects extracted from the captions of several image-text datasets (*e.g.* COCO, SBU). To get the local entities of an image, the image is fed to a CLIP visual encoder (Radford et al., 2021), then a cosine similarity is applied to compute the distance between the image and all textual embeddings of local entities, to select the closest k ones.

Event (instructions): To describe the event, we consider the caption. Even though the caption might describe only one event in which some of the objects participate, we found that using additional captions does not help significantly.

Note that, this approach can be leveraged in a straightforward way to other domains with structured text, such as Medical applications.

VLP with Structured Text: Once we create datasets of images and structured-text pairs, we can feed such data to the hierarchical text encoder and pretrain our model (Fig. 2) using standard VLP objectives. We use both ITC and ITM objectives. For text-to-image ITC loss (similarly for the image-to-text ITC), the triplet loss is fed with the text (t) and image (v) embeddings:

$$l(t_a, v_p, v_n, \alpha) = [d(t_a, v_p) + \alpha - d(t_a, v_n)]_+, \quad (1)$$
$$t = \mathcal{E}_t(G, L, E), \quad v = \mathcal{E}_v(I),$$

where t_a, v_p and v_n are the anchor, positive and negative embeddings respectively, α is the margin and $d(\cdot,\cdot)$ is a distance function. The image embedding is obtained after processing the image (I) with the image encoder $\mathcal{E}_v$. The text embedding is obtained after processing the structured text, with the extracted local (L), global (G) and event (E) elements. Specifically, $\mathcal{E}_t$ first encodes each entity independently using transformer encoders, then exploits their interactions with cross attention (Shukor et al., 2022b). We then compute ITC loss ($\mathcal{L}_{itc}$) by summing the triplet losses over the batch and weight the loss by the inverse of number of active triplet as done in Adamine (Carvalho et al., 2018). All examples in the batch are considered negatives, except the images that correspond to the recipe and vice-versa. The ITM loss can be written as:

$$\mathcal{L}_{itm} = -\mathbb{E}_{T,I\sim D}[y \log(s(T,I)) + (1-y)\log(1-s(T,I))], \quad (2)$$

where y is the label (*i.e.*, 1 for matching pairs and 0 otherwise) and D is the set of structured text ($T = \{L, G, E\}$) and image (I) pairs, and $s()$ is the score on top of the multimodal module. The total loss becomes:

$$\mathcal{L} = \mathcal{L}_{itc} + \lambda \mathcal{L}_{itm} \quad (3)$$

On the image side, to ease the pretraining, and leverage the initial visual representation, we follow LiT (Zhai et al., 2022) and keep the vision encoder frozen, we also find that this gives better results. We use a general vocabulary (used in BERT) and change the embedding layer during this stage.

3.2. *Leveraging Foundation Models for Structured Downstream Tasks*

We propose to leverage foundation models (CLIP (Radford et al., 2021)), without any retraining, for cross modal food retrieval. The approach is based on injecting local and global textual contexts in the image encoder, to enrich the visual representation and steer it towards the textual embedding space. This

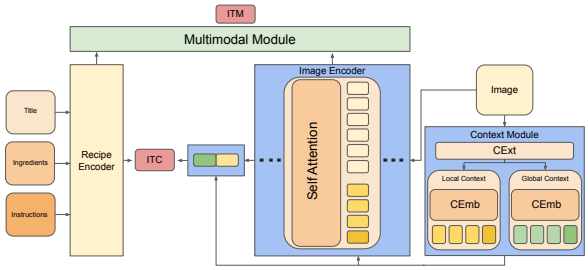


Fig. 3: Illustration of our contextualized vision encoder (stage 2 of VLPCook). The ViT is contextualized by the context module, which extracts local and global context (CExt), then project them using a light-weight module (CEmb) to obtain the context tokens. Local context tokens are concatenated to the image tokens at the input of the ViT, and the global context token (CLS token) is concatenated at the output.

context inherits the features and biases in the pretrained CLIP, which excels in general cross-modal retrieval tasks. We adopt a vision transformer (ViT (Dosovitskiy et al., 2021)) on the image side. We elaborate first on how we contextualize the ViT, then we detail the finetuning step. The model is illustrated in Fig. 3.

Contextualized Visual Representation: We inject different types of contexts during the image encoding; global and local. For global context, we inject different titles, while for local one, we inject different ingredients. The titles and ingredients are extracted from the image using our CLIP-based retrieval approach (Sec. 3.1). During training, we inject different titles, ingredients and different combination of them for each batch to add more variability and some regularization during training.

To obtain the context tokens, we concatenate all context elements (all titles for global context or all ingredients for local one) to form one sentence that is embedded using the Context Embedding (CEmb) module (Fig. 3). CEmb consists of a light-weight text encoder and a linear projection layer to project the textual tokens to the space of the visual tokens. We inject the local context early, in the input of the ViT (concatenation to the image tokens), and the global one, later in its output (concatenation of CLS token before the linear projection), where we have higher abstraction level and more global representation. The forward pass of the contextualized ViT can be expressed as follows:

$$x = ViT(Concat(i_1, .., i_k, c^l_1, .., c^l_p)) \quad (4)$$
$$x = F(Concat(x_{cls}, c^g_{cls}))$$

Where i_j, c^l_j and c^g_j are the tokens of the image (k tokens), local context (p tokens) and global context respectively. The cls means the class token and F is a linear layer.

This is different from other food approaches that add only global information (food category or class) later by concatenating it to the visual embedding (Xie et al., 2021a) or other approaches that concatenate object tags (OSCAR (Li et al., 2020c)) or visual concepts (ViCHA (Shukor et al., 2022a)) only at the input, without any distinction between local and global contexts. Our approach is also inspired by prompt tuning techniques (Lester et al., 2021; Lu et al., 2022) where a couple of learnable tokens are concatenated before the main text to adapt the frozen model to a given task.

Finetuning: We finetune the model on cross-modal food retrieval. During this stage, we inject the local and global contexts (Sec 3.2). The model consists of a ViT, hierarchical recipe encoder and a mulitmodal module (Shukor et al., 2022b), mainly we train the model using Adamine triplet loss (Carvalho et al., 2018) with incremental margin, in addition to the ITM loss as a multimodal regularization at the output of the mulimodal module. During test, we only use the unimodal encoders for fast retrieval. The context is injected also during test.

Implementation details: the model consists of hierarchical transformer encoders and decoders on the recipe side, a ViT-B/16 on the image side and a multimodal module. For VLP, we start by pretraining (with frozen ViT) with learning rate (lr) of 1e-5 and total batch size of 200 on 4 GPUs (50 per GPU) for 30 epochs. In the second finetuning stage on Recipe1M, we follow the implementation details of other work (Shukor et al., 2022b). We associate each image to 5 titles and 15 ingredients. During training, we sample only 2 titles and 4 ingredients randomly in each batch. The context is embedded by the first 2 layers of the BERT (Devlin et al., 2018) encoder, followed by linear projection (more details in the appendix).

4. Experiments

Datasets and metrics: We use several datasets; such as Recipe1M (Salvador et al., 2017) where each example consists of a recipe (title, ingredients, instructions) and image pair. Recipe1M+ (Marin et al., 2019) that is an extension of Recipe1M with 13M images and 1M recipe, and Image and Structured Text pairs (IST), which is our dataset constructed with the STE module from 3 public datasets; COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017) and SBU (Ordonez et al., 2011) to form a total of 2M pairs including around 1M different images. We follow previous works and report the recall @1/5/10 in addition to RSUM which is the sum of the 3 recalls.

	10k					
	image-to-recipe			recipe-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Adamine (Carvalho et al., 2018)	14.8	34.6	46.1	14.9	35.3	45.2
R2GAN (Zhu et al., 2019)	13.5	33.5	44.9	14.2	35.0	46.8
MCEN (Fu et al., 2020)	20.3	43.3	54.4	21.4	44.3	55.2
ACME (Wang et al., 2019)	22.9	46.8	57.9	24.4	47.9	59.0
SN (Zan et al., 2020)	22.1	45.9	56.9	23.4	47.3	57.9
IMHF (Li et al., 2021b)	23.4	48.2	58.4	24.9	48.3	59..4
Wang et. al (Wang et al., 2021a)	23.4	48.8	60.1	24.6	50.0	61.0
SCAN (Wang et al., 2021b)	23.7	49.3	60.6	25.3	50.6	61.6
HF-ICMA (Li et al., 2021c)	24.0	51.6	65.4	25.6	54.8	67.3
MSJE (Xie et al., 2021b)	25.6	52.1	63.8	26.2	52.5	64.1
SEJE (Xie et al., 2021c)	26.9	54.0	65.6	27.2	54.4	66.1
M-SIA (Li et al., 2021d)	29.2	55.0	66.2	30.3	55.6	66.5
DaC (Fain et al., 2019)	30.0	56.5	67.0	-	-	
X-MRS (Guerrero et al., 2021)	32.9	60.6	71.2	33.0	60.4	70.7
H-T (ViT) (Salvador et al., 2021)	33.5	62.1	72.8	33.7	62.2	72.7
T-Food (ViT) (Shukor et al., 2022b)	40.0	67.0	75.9	41.0	67.3	75.9
T-Food (CLIP-ViT) (Shukor et al., 2022b)	43.4	70.7	79.7	44.6	71.2	79.7
VLPCook	45.3	72.4	80.8	46.4	73.1	80.9
VLPCook (R1M+)	**46.7**	**73.3**	**83.3**	**47.8**	**74.1**	**81.8**

Table 1: Comparison with other work. Recall@k (↑) is reported on the Recipe1M test set. Our approaches (VLPCook) significantly outperform all existing work. Best metrics are in bold, and next best metrics are underlined.

4.1. Foundation Models in the Cooking Context.

Best SoTA results on general benchmarks are currently obtained by finetuning foundation models, however, here we show that for tasks requiring more complex input, such as food retrieval, this paradigm lags significantly behind existing food models. To this end, we finetune on Recipe1M for cross-modal retrieval, considering 2 kinds of approaches; light fusion (CLIP) and heavy fusion (ALBEF) approaches.

CLIP (Radford et al., 2021): Is trained contrastively on 400M of image-text pairs and consists of a ViT-Base/16 as image encoder and a transformer as text encoder.

ALBEF (Li et al., 2021a): Is trained using ITC, ITM and MLM losses on 14M images and their corresponding text. It consists of a ViT-Base/16 on the image side, a BERT on the text side, in addition to a multimodal decoder.

For both models, we change the word embedding layer, the vocabulary, and maximum number of textual tokens to 300. We train for 120 epochs with the two losses; Adamine triplet with incremental margin, semantic regularization, and ITM (for ALBEF). We use Adam optimizer and learning rate of 1e-5 (for CLIP ViT we use lr of 1e-6) and a total batch size of 80 and 56 for CLIP and ALBEF respectively. Tab. 2 shows that CLIP and ALBEF give reasonable performance and outperform most of the baselines (Tab. 1). However, and contrary to other general benchmarks, their performance is still below SoTA food models.

	image-to-recipe			recipe-to-image		
Model	R@1	R@5	R@10	R@1	R@5	R@10
X-MRS (Guerrero et al., 2021)	64.0	88.3	92.6	63.9	87.6	92.6
H-T (ViT) (Salvador et al., 2021)	64.2	89.1	93.4	64.5	89.3	93.8
T-Food (Shukor et al., 2022b)	68.2	87.9	91.3	68.3	87.8	91.5
CLIP	63.5	85.4	90.0	64.1	85.8	90.1
ALBEF	61.0	84.7	89.9	61.9	84.6	89.8

Table 2: Finetuning foundation models on Recipe1M (1k setup).

4.2. VLPCook Results

Results on Recipe1M. Tab. 1 shows that VLPCook significantly outperforms current SoTA (+1.9 R@1) on the challenging 10k setup. Importantly, the gap between VLPCook pretrained on Recipe1M+ and SoTA is even bigger (+3.4 R@1 on 10k). We also show some qualitative results in Fig. 4. We can notice the superiority of VLPCook compared to the current SoTA (Tfood CLIP-ViT). Specifically, in the first example, VLPCook correctly retrieves the right image. In the second example, our approach retrieves semantically similar images (Lasagna), while for TFood, there are totally different plates (*e.g.* rice, pasta).

	image-to-recipe			recipe-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Marin et al. Marin et al. (2019)	17.0	38.0	48.0	17.0	42.0	54.0
T-Food *	44.3	75.0	83.60	45.0	75.5	83.9
T-Food (CLIP-ViT) *	**46.5**	**76.8**	**85.4**	46.8	77.0	85.2
VLPCook*	45.2	75.9	84.0	**47.3**	**77.6**	**85.3**

Table 3: Comparison with other work. Recall@k (↑) is reported on the Recipe1M+ test set (1k setup). Best metrics are in bold. VLPCook* here is without VLP. *: we retrain these models on Recipe1M+.

Results on Recipe1M+. in Tab. 3, we show the first finetuning results on Recipe1M+ with interesting scores (more details in the appendix). Due to the large dataset size, we report the results of VLPCook without VLP (only with the context module). The scores are almost multiplied by 3 compared to the baseline (Marin et al., 2019). Moreover, we retrain the SoTA T-Food models on this dataset and show significant improvment compared to T-Fodd and comparable scores to T-Food (CLIP-ViT). This reveal that our context module is more beneficial for lower data regime (*e.g.*, on Recipe1M dataset) The low scores on this challenging dataset makes it interesting to devise more complex approaches in the future.

	image-to-recipe			recipe-to-image		
Model	R@1	R@5	R@10	R@1	R@5	R@10
VLPCook	73.6	90.5	93.3	74.7	90.7	93.2
w/o VSLP	72.3	90.6	93.4	73.6	90.8	93.5
w/o VSLP and CLIP-ViT	69.7	88.6	91.9	70.7	88.8	92.1
w/o VSLP and CLIP-ViT and Context	68.2	87.9	91.3	68.3	87.8	91.5

Table 4: Ablation Study. Both VSLP and Context module bring significant improvement. Results on Recipe1M test set (1k setup).

	image-to-recipe			recipe-to-image		
Model	R@1	R@5	R@10	R@1	R@5	R@10
Baseline (B)	68.2	87.9	91.3	68.3	87.8	91.5
B + VLP (w/o strcuture)	67.2	87.3	91.0	67.5	87.5	91.1
B + VSLP (Unfreeze Vis. Enc.)	67.6	87.3	91.3	67.6	87.2	90.9
B + VSLP (w/ VinVL tags)	68.8	88.3	91.8	69.9	88.3	91.7
B + VSLP (ours)	69.5	88.0	91.4	69.7	88.1	91.5

Table 5: Ablation study on VSLP. Different variants of VSLP. Results on Recipe1M test set (1k setup). Baseline corresponds to VLPCook using ViT and without VSLP and the Context module.

4.3. Ablation Study of VLPCook

We report the scores on the 1k setup of Recipe1M test set:

VLPCook (Sec. 3): In Tab. 4, we show the effect of our contributions, mainly VLP and Context injection. We can notice that each one brings significant improvement compared to the baseline, as well as the combination of them. In addition, we show different design choices for VLP in Tab.5. We can notice that pretraining with structured text is better than traditional VLP on plain text. Moreover, freezing the visual encoder and using additional vinvl tags bring additional improvements.

Local and Global Context (Sec. 3.2): In Tab. 6, we do an ablation on the type and the position of the injected context. We notice that using only the ingredients (Ing) or titles (Ttl) (lines 2 and 3 Tab. 6) outperforms the baseline (line 1) without any context. Moreover, using both contexts is always better, regardless of their position. We also show that the best configuration is by injecting the ingredients at the input to the visual encoder and the titles at the output (line 5).

VSLP on the Recipe1M+ Dataset Recipe1M+ is the largest dataset for food applications, however, to the best of our knowledge, there is no work, besides the work that introduced this dataset (Marin et al., 2019), that consider it for cross-modal food retrieval. This might be due to, in addition to computation resources needed, the poor generalization from Recipe1M+ to Recipe1M as shown by the authors (Marin et al., 2019). Here

Fig. 4: Recipe-to-image comparison on the Recipe1M test set, 1k setup. TFood (first and third rows) *vs.* our VLPCook (second and fourth rows). The image in green is the ground truth, followed by the top 4 retrieved images in order. One can notice that our VLPCook approach better captures some finegrained details (type of meat) and most of the retrieved images are semantically similar.

	Context		Position		RSUM	RSUM	RSUM
	Ing	ttl	Input	Output	1K	10K	
1	✗	✗			495.00	367.10	862.10
2	✓		✓		500.54	371.43	871.97
3		✓		✓	498.61	372.16	870.77
4	✓	✓	✓(ttl&Ing)		500.86	374.68	875.54
5 (ours)	✓	✓	✓(Ing)	✓(ttl)	501.75	374.30	**876.05**
6	✓	✓	✓(ttl)	✓(Ing)	501.79	372.44	874.23

Table 6: Ablation study on the context and injection position. Local context (Ing) is better injected in the input of the ViT, and global one (ttl) in the output.

we try to leverage this dataset, and assess its benefit during pretraining. We pretrain several variants, for 30 epochs on all the recipes of Recipe1M+ (after excluding those in the validation and test set of Recipe1M) following the same implementation details as Sec. 3 (except training using only 2 GPUs), and then finetune these models on Recipe1M. The results of Tab. 7 show that Recipe1M+ is more effective than our IST, however, the latter contains only 1M images compared to 13M in the former, and the images and recipes are in the same distribution of those during finetuning. To fairly compare with IST, we also pretrain on Recipe1M+ by keeping only 10% of the images (*i.e* 1.3 images in average per recipe). Interestingly, we can notice from Tab. 7 that pretraining on IST leads to better results.

4.4. Further Experiments

Food Recognition. Retrieval task is one of the best setups to evaluate cross-modal alignment, on the other hand, there is an established consensus in the community that cross-modal

Model	VSLP	image-to-recipe			recipe-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
VLPCook w/o CLIP-ViT	IST	69.8	89.2	92.7	70.9	89.6	92.7
	R1M+	71.0	89.3	92.7	71.9	89.6	92.7
VLPCook	IST	73.6	90.5	93.3	74.7	90.7	93.2
	R1M+	74.9	91.4	93.7	75.6	91.2	93.6
VLPCook	R1M+ (1.3M Im.)	73.4	90.7	93.2	73.8	90.8	93.1

Table 7: VSLP on our IST dataset vs on Recipe1M+ (R1M+).

alignment significantly helps solving multimodal downstream tasks. To echo this finding, we test the benefit of VLP for Food Recognition on Food101 (Bossard et al., 2014) and the large ISIA Food500 (Min et al., 2020). We compare SoTA food models to our VLPCooK pre-trained with VSLP, following the linear probe setup on top of frozen ViTs. Table .8 below shows very good results, e.g. we have a significant improvement in accuracy for Food Recognition. This shows the ability of our approach to generalize to other food tasks.

Food Recognition	ImageNet (ViT)	H-T (ViT)	VLPCook (ViT)
Food101	80.99	84.44	**89.14**
ISIA Food500	52.34	57.562	**60.30**

Table 8: Linear regression classification on the test sets of Food101 and ISIA Food500. Backbone (ViT) kept frozen.

Beyond Computational Cooking: Medical Domain Our approach can be seamlessly adapted to other domains. To support that, we consider the task of structured medical retrieval. We experiment with Text-Image Retrieval for medical databases. We use the large scale ROCO dataset (Pelka et al., 2018) that consists of 81k radiology images and "reports" pairs, where the report contains a caption, keywords, Unified Medical Language

Systems Concept Unique Identifiers (CUIs) and Semantic Types. We consider the list of keywords and Semantic Types as "ingredients", the caption as "instruction" and we extract the title from the caption (Sec.3.1). Table 9, shows that our VSLP (VSLP) lead to additional ~4 points of R@1 with respect to our baseline (VLPCook). This shows the broader impact of our approach and its benefits for domains and tasks requiring structured textual input.

Method	PT	image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
VLPCook	∅	14.53	38.20	51.71	15.08	39.03	51.83
VLPCook	VSLP	**18.44**	**42.78**	**55.90**	**17.95**	**42.51**	**55.06**

Table 9: Our VSLP on ROCO Image-Text Medical Retrieval dataset.

5. Conclusion

In this work, we show the benefits of VSLP for Computational Cooking. We also, successfully leverage pretrained foundation models, to enrich the vision encoder with structured context. These contributions led to a new SoTA for Cross-Modal Food Retrieval. We show that this approach has a broader impact and can be adopted for other computational cooking applications or more general multimodal tasks, especially, those with complex input, such as Medical databases.

Acknowledgments. This work was supported by ANR grant VISA DEEP (ANR-20-CHIA-0022), and HPC resources of IDRIS 2022-[AD011013415] made by GENCI.

References

Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 .

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, pp. 2425–2433.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D., 2023. Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689 .

Biten, A.F., Gomez, L., Rusinol, M., Karatzas, D., 2019. Good news, everyone! context driven entity-aware captioning for news images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12466–12475.

Bossard, L., Guillaumin, M., Van Gool, L., 2014. Food-101 – mining discriminative components with random forests, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham. pp. 446–461.

Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M., 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 35–44.

Celaj, A., Gao, A.J., Lau, T.T., Holgersen, E.M., Lo, A., Lodaya, V., Cole, C.B., Denroche, R.E., Spickett, C., Wagih, O., Pinheiro, P.O., Vora, P., Mohammadi-Shemirani, P., Chan, S., Nussbaum, Z., Zhang, X., Zhu, H., Ramamurthy, E., Kanuparthi, B., Iacocca, M., Ly, D., Kron, K., Verby, M., Cheung-Ong, K., Shalev, Z., Vaz, B., Bhargava, S., Yusuf, F., Samuel, S., Alibai, S., Baghestani, Z., He, X., Krastel, K., Oladapo, O., Mohan, A., Shanavas, A., Bugno, M., Bogojeski, J., Schmitges, F., Kim, C., Grant, S., Jayaraman, R., Masud, T., Deshwar, A., Gandhi, S., Frey, B.J., 2023. An rna foundation model enables discovery of disease mechanisms and candidate therapeutics. bioRxiv doi:10.1101/2023.09.20.558508.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al., 2022. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 .

Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020. Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer. pp. 104–120.

Cooray, T., Cheung, N.M., Lu, W., 2020. Attention-based context aware reasoning for situation recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Couairon, G., Grechka, A., Verbeek, J., Schwenk, H., Cord, M., 2022. Flexit: Towards flexible semantic image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18270–18279.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Ding, S., Lin, L., Wang, G., Chao, H., 2015. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition 48, 2993–3003.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Liu, Z., Zeng, M., et al., 2021. An empirical study of training end-to-end vision-and-language transformers. arXiv preprint arXiv:2111.02387 .

Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., et al., 2022. An empirical study of training end-to-end vision-and-language transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18166–18176.

Fain, M., Twomey, N., Ponikar, A., Fox, R., Bollegala, D., 2019. Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to sota. arXiv preprint arXiv:1911.12763 .

Fu, H., Wu, R., Liu, C., Sun, J., 2020. Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14570–14580.

Gao, H., Li, Y., Long, K., Yang, M., Shen, Y., 2024. A survey for foundation models in autonomous driving. arXiv preprint arXiv:2402.01105 .

Guerrero, R., Pham, H.X., Pavlovic, V., 2021. Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace learning, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3192–3201.

Huang, X., Liu, J., Zhang, Z., Xie, Y., 2023. Improving cross-modal recipe retrieval with component-aware prompted clip embedding, in: Proceedings of the 31st ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 529–537. URL: https://doi.org/10.1145/3581783.3612193, doi:10.1145/3581783.3612193.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR. pp. 4904–4916.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123, 32–73.

Lester, B., Al-Rfou, R., Constant, N., 2021. The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059.

Li, J., Li, D., Xiong, C., Hoi, S., 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086 .

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H., 2021a. Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems 34.

Li, J., Sun, J., Xu, X., Yu, W., Shen, F., 2021b. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, Association for Computing Machinery, New York, NY, USA. p. 173–182. URL: https://doi.org/10.1145/3460426.3463618, doi:10.1145/3460426.3463618.

Li, J., Xu, X., Yu, W., Shen, F., Cao, Z., Zuo, K., Shen, H.T., 2021c. Hybrid

Fusion with Intra- and Cross-Modality Attention for Image-Recipe Retrieval. Association for Computing Machinery, New York, NY, USA. p. 244–254. URL: https://doi.org/10.1145/3404835.3462965.

Li, L., Hu, C., Zhang, H., Maradapu Vera Venkata sai, A., 2024. Cross-modal image-recipe retrieval via multimodal fusion, in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3595916.3626389, doi:10.1145/3595916.3626389.

Li, L., Li, M., Zan, Z., Xie, Q., Liu, J., 2021d. Multi-Subspace Implicit Alignment for Cross-Modal Retrieval on Cooking Recipes and Food Images. Association for Computing Machinery, New York, NY, USA. p. 3211–3215. URL: https://doi.org/10.1145/3459637.3482149.

Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., Chang, S.F., 2022b. Clip-event: Connecting text and images with event structures, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16420–16429.

Li, M., Zareian, A., Lin, Y., Pan, X., Whitehead, S., Chen, B., Wu, B., Ji, H., Chang, S.F., Voss, C., Napierski, D., Freedman, M., 2020a. GAIA: A fine-grained multimedia knowledge extraction system, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online. pp. 77–86. URL: https://aclanthology.org/2020.acl-demos.11, doi:10.18653/v1/2020.acl-demos.11.

Li, M., Zareian, A., Zeng, Q., Whitehead, S., Lu, D., Ji, H., Chang, S.F., 2020b. Cross-media structured common space for multimedia event extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 2557–2568. URL: https://aclanthology.org/2020.acl-main.230, doi:10.18653/v1/2020.acl-main.230.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al., 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision, Springer. pp. 121–137.

Li, Z., Fan, Z., Tou, H., Wei, Z., 2022c. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. arXiv preprint arXiv:2201.12596 .

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Zhou, J., 2023. Remoteclip: A vision language foundation model for remote sensing. arXiv preprint arXiv:2306.11029 .

Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X., 2022. Prompt distribution learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5206–5215.

Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A., 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE transactions on pattern analysis and machine intelligence 43, 187–203.

Martinel, N., Piciarelli, C., Micheloni, C., Foresti, G.L., 2015. A structured committee for food recognition, in: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 484–492. doi:10.1109/ICCVW.2015.70.

Min, W., Liu, L., Wang, Z., Luo, Z., Wei, X., Wei, X., Jiang, S., 2020. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 393–401.

Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K., 2015. Im2calories: Towards an automated mobile vision food diary, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1233–1241. doi:10.1109/ICCV.2015.146.

Ofli, F., Aytar, Y., Weber, I., Al Hammouri, R., Torralba, A., 2017. Is saki# delicious? the food perception gap on instagram and its relation to health, in: Proceedings of the 26th International Conference on World Wide Web, pp. 509–518.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Ordonez, V., Kulkarni, G., Berg, T.L., 2011. Im2text: Describing images using 1 million captioned photographs, in: Proceedings of the 24th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 1143–1151.

Papadopoulos, D.P., Mora, E., Chepurko, N., Huang, K.W., Ofli, F., Torralba, A., 2022. Learning program representations for food images and cooking recipes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16559–16569.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer. pp. 180–189.

Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S., 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE international conference on computer vision, pp. 2641–2649.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR. pp. 8748–8763.

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

Salvador, A., Gundogdu, E., Bazzani, L., Donoser, M., 2021. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15475–15484.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A., 2017. Learning cross-modal embeddings for cooking recipes and food images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Sara, S., Cornia, M., Baraldi, L., Cucchiara, R., 2022. Retrieval-augmented transformer for image captioning, in: 19th International Conference on Content-based Multimedia Indexing.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A., 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 .

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D., 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval, in: Proceedings of the Fourth Workshop on Vision and Language, Association for Computational Linguistics, Lisbon, Portugal. pp. 70–80. URL: https://aclanthology.org/W15-2812, doi:10.18653/v1/W15-2812.

Sharma, P., Ding, N., Goodman, S., Soricut, R., 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: ACL.

Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K., 2022. How much can CLIP benefit vision-and-language tasks?, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=zf_Ll3HZWgy.

Shukor, M., Couairon, G., Cord, M., 2022a. Efficient vision-language pretraining with visual concepts and hierarchical alignment, in: 33rd British Machine Vision Conference (BMVC).

Shukor, M., Couairon, G., Grechka, A., Cord, M., 2022b. Transformer decoders with multimodal regularization for cross-modal food retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4567–4578.

Shukor, M., Dancette, C., Cord, M., 2023. ep-alm: Efficient perceptual augmentation of language models. arXiv preprint arXiv:2303.11403 .

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D., 2021. Flava: A foundational language and vision alignment model. arXiv preprint arXiv:2112.04482 .

Song, F., Zhu, B., Hao, Y., Wang, S., He, X., 2023. Car: Consolidation, augmentation and regulation for recipe retrieval. arXiv preprint arXiv:2312.04763 .

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations, in: International Conference on Learning Representations.

Suhail, M., Sigal, L., 2019. Mixture-kernel graph attention network for situation recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Wang, H., Lin, G., Hoi, S.C., Miao, C., 2021a. Learning structural rep-

resentations for recipe generation and food retrieval. arXiv preprint arXiv:2110.01209 .

Wang, H., Sahoo, D., Liu, C., Lim, E.p., Hoi, S.C., 2019. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11572–11581.

Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.p., Hoi, C.S., 2021b. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. IEEE Transactions on Multimedia .

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H., 2022a. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. arXiv preprint arXiv:2202.03052 .

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al., 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 .

Weinberger, K.Q., Blitzer, J., Saul, L., 2005. Distance metric learning for large margin nearest neighbor classification, in: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems, MIT Press. URL: https://proceedings.neurips.cc/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf.

Xie, Z., Liu, L., Li, L., Zhong, L., 2021a. Learning joint embedding with modality alignments for cross-modal retrieval of recipes and food images, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2221–2230.

Xie, Z., Liu, L., Wu, Y., Li, L., Zhong, L., 2021b. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. IEEE Transactions on Services Computing .

Xie, Z., Liu, L., Wu, Y., Zhong, L., Li, L., 2021c. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. ACM Transactions on Information Systems (TOIS) 40, 1–27.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 .

Zan, Z., Li, L., Liu, J., Zhou, D., 2020. Sentence-Based and Noise-Robust Cross-Modal Retrieval on Cooking Recipes and Food Images. Association for Computing Machinery, New York, NY, USA. p. 117–125. URL: https://doi.org/10.1145/3372278.3390681.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L., 2022. Lit: Zero-shot transfer with locked-image text tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18123–18133.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 .

Zhu, B., Ngo, C.W., 2020. Cookgan: Causality based text-to-image synthesis, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5518–5526. doi:10.1109/CVPR42600.2020.00556.

Zhu, B., Ngo, C.W., Chen, J., Hao, Y., 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- Existing general foundation models underperoform on computaional cooking tasks.
- Domain specific applications need more adapted pretraining approaches.
- Adapting existing general datasets of image-text pairs to be closer to food data.
- Vision Language Pretraining on adapted datasets helps cooking downstream tasks.
- Foundation models can be leveraged for Food models by injecting external knowledge.

Declaration of Interest Statement

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Matthieu Cord reports financial support was provided by French National Research Agency. Matthieu Cord reports a relationship with Valeo.ai that includes: employment. Mustafa Shukor reports a relationship with InterDigital, Inc that includes: employment.