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

Mustafa Shukor, Nicolas Thome, Matthieu Cord

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

Mustafa Shukor Nicolas Thome Matthieu Cord

Sorbonne University

{firstname.lastname}@sorbonne-universite.fr

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 is available here: https://github.com/mshukor/VLPCook.*

1. Introduction

Vision-Language Pretraining (VLP) [7, 17, 30, 67, 71] has become the general recipe to attain SoTA results on downstream unimodal and multimodal tasks, with the key success is learning a shared latent space where all modalities are aligned. This paradigm generally 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 [6, 75, 76]. 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 [1, 6, 22, 29, 53, 83].

However, these approaches are still evaluated on simple downstream tasks, to the detriment of more complex albeit important tasks. The current evaluation schema considers tasks such as VQA [2], Visual entailment [78], Image-Text Retrieval [52], Image Classification and other general benchmarks that 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, 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 [56], aiming at bridging the gap between VLP and Computational Cooking.

Computational Cooking or Food applications [19, 44, 47, 56] are one of the important applications that fit very well in this marginalized list, with no existing work to bridge the gap with VLP. In particular, Cross-Modal Food Retrieval [5, 55, 56, 64] which has gained a lot of attention in the recent years and 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 com-

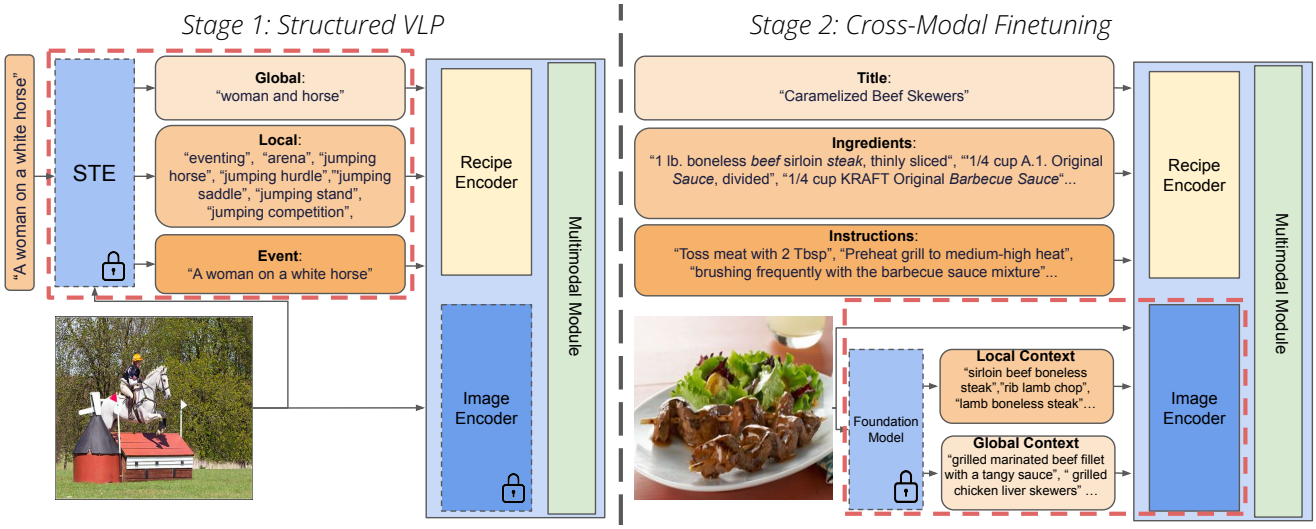

Figure 1. VLPCook framework with 2 sequential stages. Stage 1 (left) or VSLP (Sec. 3.1): the Structured Text Extraction (STE) module transforms the caption to a structured recipe-like input that is used to pretrain the model on a large corpus of structured text and images. Stage 2 (right) or Cross-Modal 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.

patible, structurally and semantically, to some extent, to fit in these models. In addition, and pushing on the environmentally responsible idea of reusing existing models, 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:

- 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.
- 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:

- 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.
- The first work showing the effectiveness of VLP in the cooking context, after experimenting with different kinds of existing food approaches.
- Despite what was reported [43] 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.
- 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.
- 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) [7, 67, 71] aims at learning vision-language representation by pretraining on datasets of images and texts ([1, 49, 53, 59, 61]). The model is then evaluated on several downstream tasks such as VQA [2], NLVR2 [70], image-text retrieval [52] and image captioning [40]. This line of research has shown promising success in the last few years, leading to state of art (SoTA) results [17, 29, 30] 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 sepa-

rate vision and language encoders, without significant cross modal interaction (*e.g.*, CLIP [53], ALIGN [23]). Despite their fast inference, they are data hungry and perform poorly on tasks that need deeper reasoning. To overcome these limitations, heavy fusion models use a cross modal interaction module [7, 25, 28, 34, 41, 67, 86] which is added on top of unimodal encoders [16, 30, 63, 82] leading to hybrid models. These hybrid approaches have succeeded to get SoTA results while training on reasonably sized datasets. On the learning side, the main training objectives can be categorised into contrastive (ITC [53], ITM [7]) and masked predictions (MLM [12], MIM [17, 63]). The models that work best are those that combine several objectives, however, at large scale, there are many attempts to unify pretraining tasks.

Leveraging Foundation Models Foundation models [1, 53, 65, 75, 76, 83] draw some similarity with VLP, however here the objective is to develop a general model that can be adapted to many unimodal and multimodal tasks. Here there is more emphasis on large scale, in terms of tranining data [53], and model size [83] and on unification of the architectural design and training objectives [75, 76]. In spite of being successful, due to the need for huge resources to traininig these models from scratch, researchers and practioners have leveraged them, without the burden of retraining; such as initialization and finetuning [62, 64], as frozen modules [11, 54, 68], enriching the input [57] and extracting visual concepts [63]. 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 [4], calorie estimation [46], image generation [88] and cross modal retrieval [56]. Since the inception of large scale food datasets such Recipe1M [56] followed by Recipe1M+ [43] the task of cross-modal retrieval have gained a lot of attention. In terms of performance and architectural designs, cross modal food retrieval work can be divided into transformer-based [21, 50, 55, 64] or transformer-free [5, 18, 56, 73, 74, 89] approaches, with a significant improvements of the former. Specifically, on the vision side, ViT [14] is used as an image encoder, and on the recipe side, standard [21] or hierarchical transformers [55, 64] are adopted. In terms of training objectives, almost all approaches use triplet loss [13, 58, 77] in addition to some regularization such as semantic triplet [5, 64], embedding classification [56], adversarial losses [73] and multimodal regularization with image-text matching objective [64]. In addition to food applications, learning from structured texts and images has been investigated in several domains and tasks, such as Medical applications [51], News applications [3], Multimedia Event extraction [36, 37] and Situation Recognition [10, 69]. In the context of VLP, few work have been recently proposed [35, 39], 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 solves 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 [48, 66, 87] and triplet loss [13, 77]. 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 [7]. 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; usually a one sentence describing a general event, or the scene in the image. Despite being easily scraped from the internet, and successful in many general downstream tasks, image captions are not directly aligned with some domains such

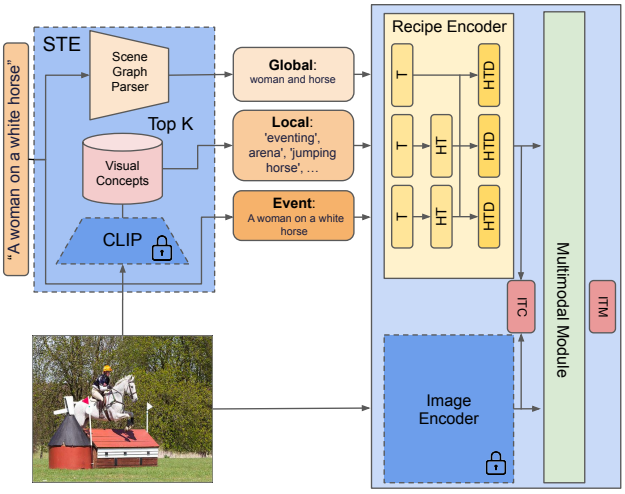

Figure 2. Illustration of our VSLP (Stage 1 of VLPCook). To enable VLP for food models, image-text pairs are transformed to image and structured-text pairs, that are compatible with hierarchical recipe encoders. The Structured Text Extraction (STE) module generates 3 entities; (a) global description ("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.

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 [55, 64] 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 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): we assume that the caption describes either the global scene or the main event in the image, and 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). As a way to filter out the caption and keep the main elements, we extract only the objects using Scene Graph Parsing (SGP) [60] 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. Relying on the caption alone is not optimal, as it contains only few seen objects, besides referring to global aspects of the scene. On the other hand, we do not want to be limited to seen objects and include unseen but relevant objects, which is the case for ingredients in food tasks (*e.g.* salt, sugar). This motivates us to leverage additional sources of information to extract all relevant, seen or unseen, objects. 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 [53], 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 [64]. 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 [5]. 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 [85] 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.

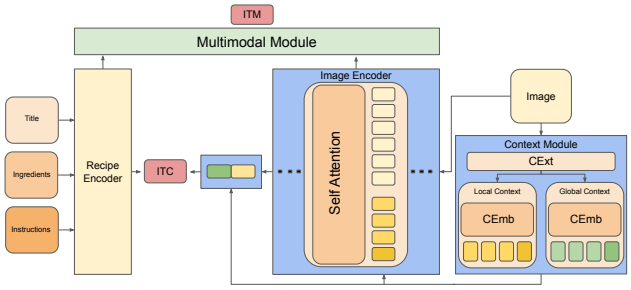

Figure 3. Illustration of our contextualized vision encoder (stage 2 of VLPCook). The ViT is contextualized by the context module, which extracts local and global context (CExt), then project them using a light-weight module (CEmb) to obtain the context tokens. Local context tokens are concatenated to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

3.2. Leveraging Foundation Models for Structured Downstream Tasks

We propose to leverage foundation models (CLIP [53]), 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 context inherits the features and biases in the pretrained CLIP, which excels in general cross-modal retrieval tasks. We adopt a vision transformer (ViT [14]) 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:

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

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 [79] or other approaches that concatenate object tags (OSCAR [38]) or visual concepts (ViCHA [63]) only at the input, without any distinction between local and global contexts. Our approach is also inspired by prompt tuning techniques [24, 27, 42] 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 [64], mainly we train the model using Adamine triplet loss [5] 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.

4. Experiments

In this section we detail the experimental results.

Datasets: We use several datasets; such as Recipe1M [56] (239 k, 51 k, 51 k pairs as training, validation and test set) where each example consists of a recipe (title, ingredients, instructions) and image pair. Recipe1M+ [43] 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 [40], Visual Genome [26] and SBU [49] to form a total of 2M pairs including around 1M different images.

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 [64]. 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 [12] encoder, followed by linear projection (more details in the appendix).

	10k					
	image-to-recipe			recipe-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Adamine [5]	14.8	34.6	46.1	14.9	35.3	45.2
R2GAN [89]	13.5	33.5	44.9	14.2	35.0	46.8
MCEN [20]	20.3	43.3	54.4	21.4	44.3	55.2
ACME [73]	22.9	46.8	57.9	24.4	47.9	59.0
SN [84]	22.1	45.9	56.9	23.4	47.3	57.9
IMHF [31]	23.4	48.2	58.4	24.9	48.3	59..4
Wang et. al [72]	23.4	48.8	60.1	24.6	50.0	61.0
SCAN [74]	23.7	49.3	60.6	25.3	50.6	61.6
HF-ICMA [32]	24.0	51.6	65.4	25.6	54.8	67.3
MSJE [80]	25.6	52.1	63.8	26.2	52.5	64.1
SEJE [81]	26.9	54.0	65.6	27.2	54.4	66.1
M-SIA [33]	29.2	55.0	66.2	30.3	55.6	66.5
DaC [18]	30.0	56.5	67.0	-	-	
X-MRS [21]	32.9	60.6	71.2	33.0	60.4	70.7
H-T (ViT) [55]	33.5	62.1	72.8	33.7	62.2	72.7
T-Food (ViT) [64]	40.0	67.0	75.9	41.0	67.3	75.9
T-Food (CLIP-ViT) [64]	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. VLPCook Results

Results on Recipe1M: *Comparison with SoTA:* Tab. 1 shows a comparison with existing approaches on the test set of Recipe1M. 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).

Qualitative Comparison with SoTA: we 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. [43]	17.0	38.0	48.0	17.0	42.0	54.0
VLPCook*	**45.2**	**75.9**	**84.0**	**47.3**	**77.6**	**85.3**

Table 2. 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.

Results on Recipe1M+: in Tab. 2, 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 [43]. However, there is a big gap between the scores on this dataset and those on Recipe1M, which makes it more challenging and more interesting to devise more complex approaches in the future.

Model	image-to-recipe			recipe-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	40.0	67.0	75.9	41.0	67.3	75.9
+ VSLP	41.1	67.5	76.1	42.4	68.1	76.5
+ VSLP & Context	41.1	68.0	76.9	42.8	69.2	77.8

Table 3. Ablation Study: Both VSLP and Context module bring significant improvement.

4.2. Ablation Study of VLPCook

Here we present the ablation study for some design choices, on the 1k setup of Recipe1M test set:

VLPCook (Sec. 3): In Tab. 3, 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.

Local and Global Context (Sec. 3.2): In Tab. 4, 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)

Figure 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.

(lines 2 and 3 Tab. 4) 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).

	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 4. 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.

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 [43], 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 [43]. Here 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. 5 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. 5 that pretraining on IST leads to better results.

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 5. VSLP on our IST dataset vs on Recipe1M+ (R1M+). Pretraining on R1M+ gives better results, however, for the same number of images, IST is a better choice.

4.3. Further Experiments

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 [53]: 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 [30]: 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. 6 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 [21]	64.0	88.3	92.6	63.9	87.6	92.6
H-T (ViT) [55]	64.2	89.1	93.4	64.5	89.3	93.8
T-Food [64]	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 6. Finetuning foundation models on Recipe1M.

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 alignment significantly helps solving multimodal downstream tasks. To echo this finding, we test the benefit of VLP for Food Recognition on Food101 [4] and the large ISIA Food500 [45]. We compare SoTA food models to our VLPCooK pre-trained with VSLP, following the linear probe setup on top of frozen ViTs. Table .7 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 7. Linear regression classification on the test sets of Food101 and ISIA Food500. Backbone (ViT) kept frozen.

4.4. Beyond Computational Cooking: Medical Domain

Although stage 1 of our approach has been tailored for computational cooking tasks (stage 2), its design is more generally concerned with the processing of structured documents, and can be seamlessly adapted to other domains. To support that, we consider structured data from very different domain, namely structured medical retrieval.

We experiment with Text-Image Retrieval for medical databases. We use the large scale ROCO dataset [51] 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 8, shows that our VSLP (VSLP) lead to additional $\sim$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 8. Comparison of different types of VLP on Image-Text Medical Retrieval on ROCO 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. An interesting follow up of this work, is to improve the textual structure extraction and going large scale in terms of pretraining data.

6. Acknowledgments

The authors would like to thank Rémy Sun for fruitful discussion. This work was partly supported by ANR grant VISA DEEP (ANR-20-CHIA-0022), and HPC resources of IDRIS under the allocation 2022-[AD011013415] and 2022-[A0121012449] made by GENCI.

Appendix

The Appendix is organized as follows; Sec. A elaborates on the implementation details, Sec. B presents the complete comparison of VLPCook with other SoTA approaches on Recipe1M and Recipe1M+ datasets. In section Sec. C, we conduct additional VSLP experiments. In Sec. D, we do a robustness analysis to missing recipe entities, where we show also the contribution of each of these entities for food retrieval. Finally, we show some qualitative examples on the extracted text (Sec. E) and the injected local and global context (Sec. F).

A. Implementation details

VLP of VLPCook: the model consists of a hierarchical transformer encoders and decoders on the recipe side, a ViT-B/16 [15] on the image side and a multimodal module [64]. For VLP, We start by pretraining this baseline with Adamine triplet (without semantic regularization losses) [5] and ITM losses ($\lambda = 1$), with learning rate (lr) 1e-5 and total batch size of 200, on 4 GPUs (50 per GPU) for 30 epochs. We pretrain on the 2M pairs of the IST dataset. Inspired by LiT [85] we freeze the image encoder during this stage.

Finetuning on Recipe1M: in the second finetuning stage, we follow the implementation details of recent work [64], mainly, batch size of 100, lr of 1e-5 (lr of 1e-6 for CLIP-ViT) and training for 120 epochs on the training set of Recipe1M. We optimize the model with the Adamine triplet (instance and semantic) with incremental margin (we start by a $\alpha_{inc} = 0.05$ and increase it by 0.005 each epoch until reaching 0.3) and ITM objective ($\lambda = 1$). The ViT is kept frozen for the first 20 epochs. Note that, we pretrain always with a ViT, even when we finetune with CLIP-ViT. We associate each image to 5 titles and 15 ingredients. These are extracted from the recipes of the training set of Recipe1M, using the CLIP-based retrieval approach. During training, we sample only 2 titles and 4 ingredients randomly in each batch. During Test we use all titles and ingredients. We concatenate the ingredients to the input of the ViT and the title to its output, before the linear projection to the latent space. The context is embedded by the first 2 layers of the BERT [12] encoder, then linearly projected to obtain the context tokens, we find it beneficial to use separate BERT encoders for each context.

Finetuning on Recipe1M+: for finetuning on Recipe1M+ [43], we adopt the same implementation details as for Recipe1M, however, due to the large number of images (*i.e.*, 13M) we extract the context from only 1 image for each recipe and use this context for all the other corresponding images. We finetune on 2 A100 GPUs, for 60 epochs, without the semantic triplet loss and keep the ViT frozen for the first 5 epochs.

Evaluation: We follow other work and report recall@{1, 5, 10} (R@k) and their sum (RSUM), in addition to the median rank (medR) on the 1k and 10 setups, averaged over 10 and 5 runs respectively.

Image-Text Medical Retrieval. We use the large scale ROCO dataset [51] 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 as we did in our STE module. The results with standard VLP are reported from [8, 9]. We follow other approaches and evaluate on 2k pairs of the test set of ROCO. To ensure reproductibility, we average the results obtained on 4 different 2k subsets. Here we do not use the context module.

B. Comparison with SoTA

We compare VLPCook with other SoTA for Cross-Modal Food Retrieval. Tab. 9 shows the results after finetuning on Recipe1M. We outperform other SoTA by a significant margin on the 1k (+2.1 R@1) and 10k (+1.9 R@1) setups. Pretraining on Recipe1M+ (R1M+) leads to additional improvements of +3 and +3.3 R@1 on the 1k and 10k setups respectively. We also show some qualitative results in Fig. 5 and 6.

The results of training on Recipe1M+ dataset are shown in Tab. 10. We show the first interesting results on this challenging dataset, after the work [43] that introduced this dataset. Despite the large improvements, these results reveal the difficulty of this dataset, that could be interesting for devising more sophisticated approaches in the future.

C. Additional VLP experiments

Vision and Structured-Language Pretraining Variants. In Tab. 11, we compare different design choices for VSLP. The baseline is our implementation of TFood. We show the effectiveness of VSLP, especially the IST dataset, by the superiority of B+VSLP (ours) compared to B+VLP (w/o structure), which is a baseline that takes the same caption as title, ingredients and instructions, without extracting any structure. We also compare with pretraining all modules

	1k								10k							
	image-to-recipe				recipe-to-image				image-to-recipe				recipe-to-image			
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
Salvador et al. [56]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
Adamine [5]	2.0	40.2	68.1	78.7	2.0	39.8	69.0	77.4	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
R2GAN [89]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
MCEN [20]	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
ACME [73]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
SN [84]	1.0	52.7	81.7	88.9	1.0	54.1	81.8	88.9	7.0	22.1	45.9	56.9	7.0	23.4	47.3	57.9
IMHF [31]	1.0	53.2	80.7	87.6	1.0	54.1	82.4	88.2	6.2	23.4	48.2	58.4	5.8	24.9	48.3	59..4
Wang et. al [72]	1.0	53.5	81.5	88.8	1.0	55.0	82.0	88.8	6.0	23.4	48.8	60.1	5.6	24.6	50.0	61.0
SCAN [74]	1.0	54.0	81.7	88.8	1.0	54.9	81.9	89.0	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
HF-ICMA [32]	1.0	55.1	86.7	92.4	1.0	56.8	87.5	93.0	5.0	24.0	51.6	65.4	4.2	25.6	54.8	67.3
MSJE [80]	1.0	56.5	84.7	90.9	1.0	56.2	84.9	91.1	5.0	25.6	52.1	63.8	5.0	26.2	52.5	64.1
SEJE [81]	1.0	58.1	85.8	92.2	1.0	58.5	86.2	92.3	4.2	26.9	54.0	65.6	4.0	27.2	54.4	66.1
M-SIA [33]	1.0	59.3	86.3	92.6	1.0	59.8	86.7	92.8	4.0	29.2	55.0	66.2	4.0	30.3	55.6	66.5
DaC [18]	1.0	60.2	84.0	89.7	-	-	-	-	4.0	30.0	56.5	67.0	-	-	-	
X-MRS [21]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
H-T (ViT) [55]	1.0	64.2	89.1	93.4	1.0	64.5	89.3	93.8	3.0	33.5	62.1	72.8	3.0	33.7	62.2	72.7
Papadopoulos et al. [50]	1.0	66.9	90.9	**95.1**	1.0	66.8	89.8	**94.6**	-	-	-	-	-	-	-	-
T-Food (ViT) [64]	1.0	68.2	87.9	91.3	1.0	68.3	87.8	91.5	2.0	40.0	67.0	75.9	2.0	41.0	67.3	75.9
T-Food (CLIP-ViT) [64]	1.0	72.3	90.7	93.4	1.0	72.6	90.6	93.4	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
VLPCook	1.0	73.6	90.5	93.3	1.0	74.7	90.7	93.2	2.0	45.3	72.4	80.8	2.0	46.4	73.1	80.9
VLPCook (R1M+)	1.0	**74.9**	**91.4**	93.7	1.0	**75.6**	**91.2**	93.6	2.0	**46.7**	**73.3**	**83.31**	2.0	**47.8**	**74.1**	**81.8**

Table 9. Comparison with other work on the Recipe1M dataset. medR (↓), Recall@k (↑) are 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.

	1k								10k							
	image-to-recipe				recipe-to-image				image-to-recipe				recipe-to-image			
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
Marin et al. [43]	8.6	17.0	38.0	48.0	6.8	17.0	42.0	54.0	-	-	-	-	-	-	-	-
VLPCook*	2.0	45.2	75.9	84.0	2.0	47.3	77.6	85.3	9.2	18.0	40.7	52.2	8.0	19.8	43.4	55.0

Table 10. Comparison with other work on the Recipe1M+ dataset. medR (↓), Recall@k (↑) are 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. All models are trained on the training set of Recipe1M+. * means without pretraining.

(B+VSLP (+Unfreeze Vis. Enc.)) and show that this degrades the performance. Finally, we use an object detector (VinVL [86]) to extract the objects or local entities in the image, instead of our CLIP-based approach and show that both are competitive in the pretraining stage.

	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 11. Ablation study on VSLP. Different variants of VSLP.

VLP of Existing Food Models. We now validate that VLP consistently improves a wide variety of existing food models. We experiments with 2 kinds of approaches; with standard transformer (*e.g*, BERT) such X-MRS [21] and VLPCook-B (BERT) (our Baseline where we replace the recipe encoder by a BERT) and with hierarchical transformers such as TFood. We do not change the training procedure for these methods, the only difference is in the pretraining stage, or initialization. We train on the 2M pairs. The BERT-based models are trained with image captions (training on IST can be found in the appendix) and those with hierarchical transformers with our transformed datasets (structured text). Results are reported in Tab. 12, that shows a consistent improvement for all SoTA with VLP. This validates the benefit of using VLP for cross-modal food retrieval and shows the effectiveness of our approach to transform captions to structured text.

Pretraining on Recipe1M+ In this section, we analyse the influence of the number of images and recipes for VSLP. We pretrain on different subset sizes of Recipe1M+ dataset. From Tab. 13, we can notice that there is a significant improvement when adding more images. Interestingly, for comparable number of images, pretraining on IST gives

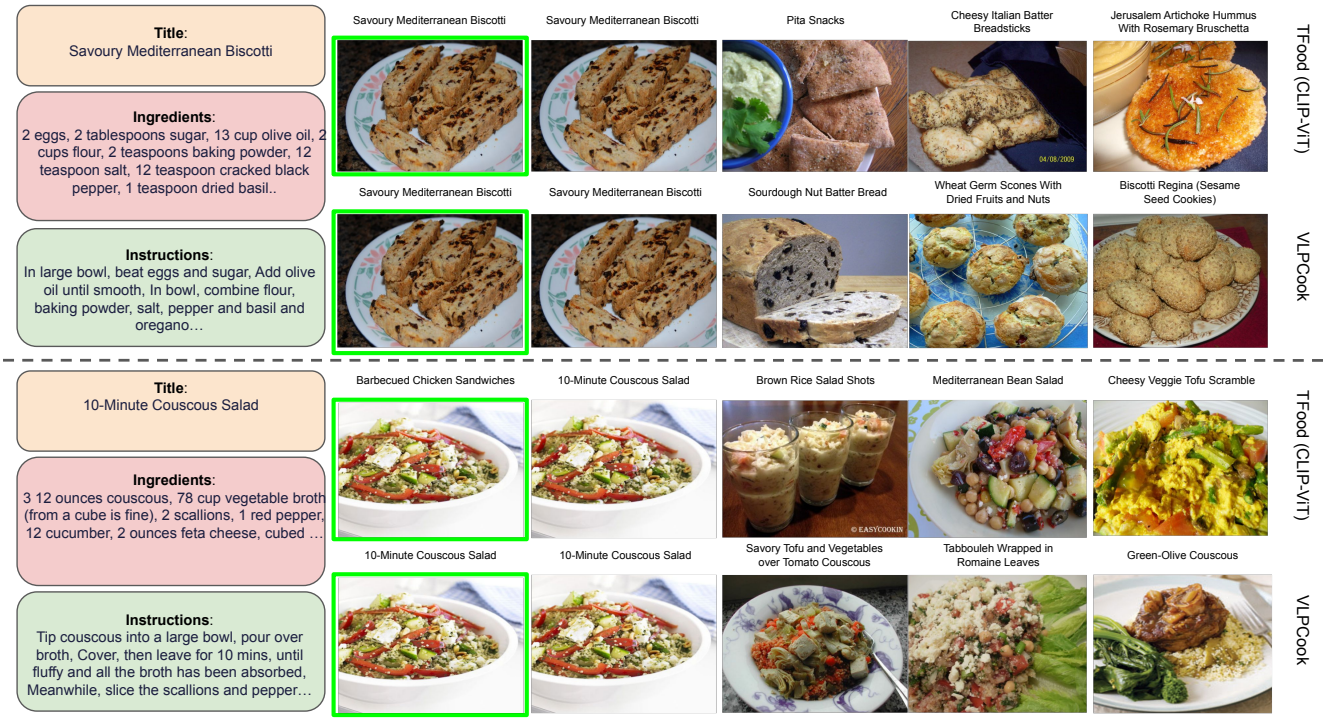

Figure 5. Recipe to image qualitative results of VLPCook on the Recipe1M test set. The image in green is the ground truth, followed by the top 4 retrieved images in order. For VLPCook, we can notice that all images semantically resemble the ground truth in addition to successfully retrieving the correct image.

Model	VLP	image-to-recipe			recipe-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
XMRS	✗	60.9	85.6	90.8	61.2	85.9	91.0
	✓	61.8	86.3	91.6	62.7	86.7	91.7
VLPCook-B (BERT)	✗	61.4	84.1	88.8	61.3	84.3	89.0
	✓	63.4	85.3	89.5	63.2	85.3	89.7
TFood	✗	68.2	87.9	91.3	68.3	87.8	91.5
	✓	69.5	88.0	91.4	69.7	88.1	91.5

Table 12. Results of VLP with existing food approaches. We see consistent improvement with VLP.

Pretraining Dataset	# Images	image-to-recipe			recipe-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
IST	1M	73.6	90.5	93.3	74.7	90.7	93.2
R1M+	1.3M	73.4	90.7	93.2	73.8	90.8	93.1
R1M+	6.5M	73.9	91.0	93.6	74.8	91.2	93.7
R1M+	13M	74.9	91.4	93.7	75.6	91.2	93.6

Table 13. VLPCook pretrained on IST and subsets of Recipe1M+ with different number of images; 1.3M (10%), 6.5M (50%) and 13M (100%).

Pretraining Dataset	% of Images	# Recipes	image-to-recipe			recipe-to-image		
			R@1	R@5	R@10	R@1	R@5	R@10
R1M+	10%	0.9M	73.4	90.7	93.2	73.8	90.8	93.1
R1M+	10%	0.45M	72.7	90.4	93.5	73.5	90.8	93.6
R1M+	100%	0.9M	74.9	91.4	93.7	75.6	91.2	93.6
R1M+	100%	0.45M	73.9	90.8	93.4	74.6	91.0	93.5

Table 14. VLPCook pretrained on IST and subsets of Recipe1M+ with different number of recipes; 0.45M (50%), and 0.9M (100%).

better performance.

When reducing the number of recipes in Tab. 14, we can notice also a significant degradation.

Interestingly, reducing the number of recipes or images to half, leads to comparable results (73.9 R@1 for 6.5M images in Tab. 13 or 0.45M recipes in Tab. 14).

D. Robustness to missing recipe entities

Here we analyse how much our model is robust against missing recipe entities. In addition, this will help to understand the importance of each element, and how much they contribute to find the right visual representation. This may also have some important applications in several scenarios (*e.g.* in case we have a specific ingredients, and we are wondering what can we make from them). The results are shown in Tab. 15. We can notice that the most important elements are the ingredients, then the instructions and finally

the title. Compared to TFood (CLIP-ViT) [64], in general we are more robust, except for missing ingredients. This indicates that our model rely heavily on the ingredients to find the image which might be caused by the local context (ingredients injected in the vision encoder) that might steer the model to focus more on the ingredients.

Missing entity	Model	image-to-recipe			recipe-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
Ttl	TFood (CLIP-ViT)	65.6	87.8	91.8	64.2	86.9	91.1
	VLPCook	68.6	88.1	92.0	68.4	87.6	91.3
Ing	TFood (CLIP-ViT)	40.6	69.6	78.6	30.4	57.5	67.3
	VLPCook	36.5	65.7	75.3	24.9	52.4	63.9
Ins	TFood (CLIP-ViT)	62.1	84.9	90.1	57.5	82.3	88.2
	VLPCook	64.1	85.7	90.0	62.0	83.8	88.6
Ttl+Ins	TFood (CLIP-ViT)	45.5	72.3	80.6	34.5	61.8	72.3
	VLPCook	51.0	76.9	83.6	42.9	69.6	78.1

Table 15. Robustness to missing recipe entities. The ingredients contribute more to finding the corresponding example, then the instructions, and finally the title.

E. Structured Text Extraction (STE)

We illustrate in Fig. 7 some qualitative examples of the structured text, obtained after transforming image captions using the STE module. We can see that the local elements are related mostly to the center of the image, describe the main or central object, and redundant. While such extracted information proved to be useful for food retrieval, devising other approaches that extracts information about all seen objects, with richer details, can help for tasks requiring more complex reasoning.

F. Local and Global Textual Concepts

Fig. 8 shows the extracted context associated with each image. We successfully extract relevant contexts describing the recipe. However, we have also the redundancy in the local context, which might be due to the biases in the CLIP to the central objects in the image.

Figure 6. Recipe to image qualitative results of VLPCook on the Recipe1M test set. The image in green is the ground truth, followed by the top 4 retrieved images in order.

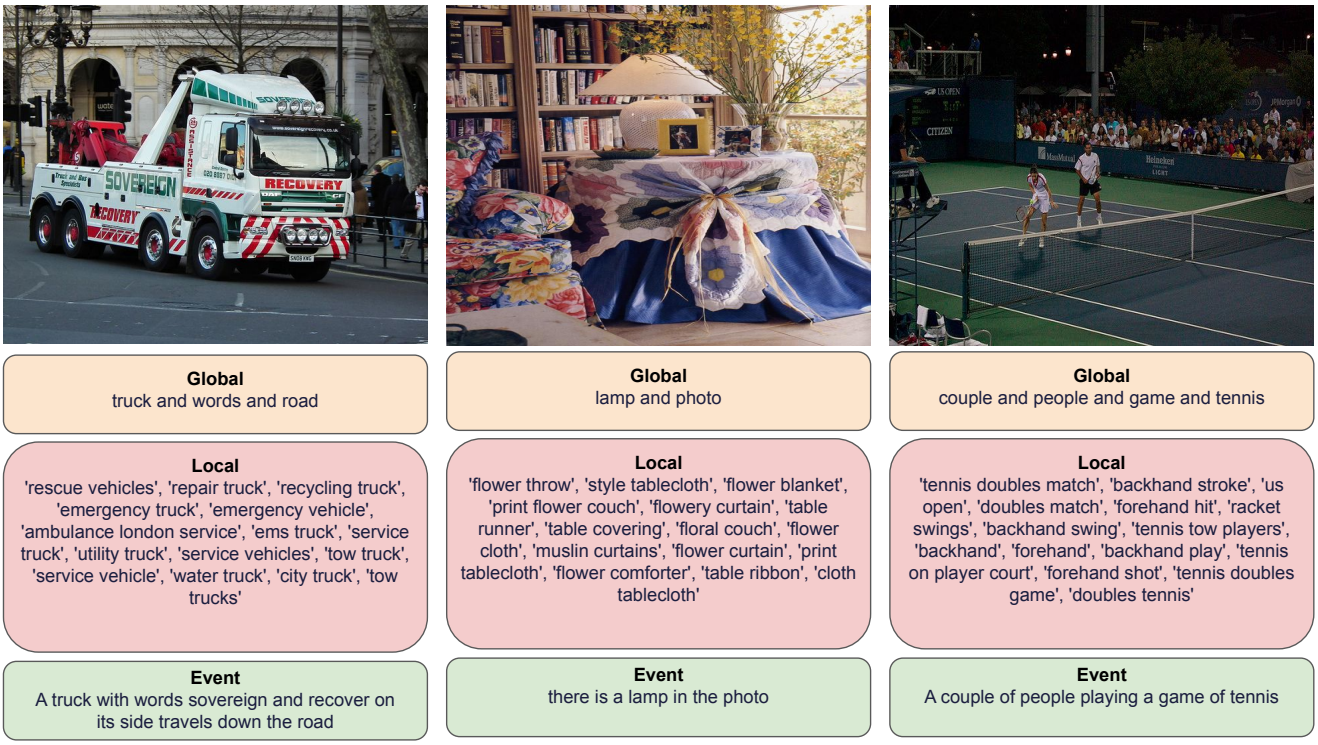

Figure 7. Illustration of the structured text, extracted by the STE module. For each image, we extract a global information using SGP , local information using CLIP-based retrieval and the event which is simply the caption.

Local Context	Local Context	Local Context
'sirloin beef boneless steak', 'loin strip boneless steak', 'steak sirloin boneless', 'sirloin boneless steak', 'pork sirloin boneless chop', 'rib boneless steak', 'lamb boneless steak', 'veal boneless steak', 'chicken skinless boneless tenderloin', 'eye rib boneless steak', 'sirloin boneless pound steak', 'sirloin beef steak', 'strip boneless steak', 'beef boneless tenderloin', 'rib lamb chop'	'sugar cups confection', 'pastry cups cream', 'pastry cups flour', 'milk cups cream', 'cups confection', 'food angel cups cak', 'chocolate semisweet cups morsel', 'pudding cups mix', 'cheese cream cups product', 'dessert cups sauc', 'cake cups flour', 'sugar cup confection', 'chocolate cups waf', 'chocolate cups piec', 'cake cups mix'	'tandoori chicken', 'wing tablespoons sauc', 'bbq chicken', 'bbq wing', 'tandoori tsp masala', 'wing tablespoon sauc', 'chicken skinless w', 'tandoori tbsp spic', 'spicy chicken', 'barbecue smoky sauc', 'chicken wings w', 'chicken fry', 'chicken tablespoon season', 'barbecue spicy tablespoons sauc', 'rotisserie boneless chicken'
Global Context	**Global Context**	**Global Context**
'grilled marinated beef fillet with a tangy sauce', 'grilled marinated pork fillet', 'chicken-liver salad with hot bacon dressing and croutons', 'grilled chicken liver skewers', 'warm pork fillet salad with honey dressing'	'pudding cupcake cones', 'pizzelle dessert cups', 'pastry cups (can substitute frozen puff pastry)', 'fluted kisses cups with peanut butter filling', 'mini chocolate meringue pie tarts in baked wonton shells'	'grilled tandoori chicken wings with coriander yogurt', 'tandoori touchdown wings with mint-mango chutney recipe harvardcommon', 'tandoori-style grilled chicken wings', 'tamarind-chipotle chicken wings', 'indian style chicken wings'

Figure 8. Illustration of the local and global concepts. Both concepts are extracted using CLIP-based retrieval. The local concepts consists of ingredients, and the global ones as recipe titles.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2, 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2

[3] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 3

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 3, 8

[5] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 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*, pages 35–44, 2018. 1, 3, 5, 6, 9, 10

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2, 3

[8] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 679–689, Cham, 2022. Springer Nature Switzerland. 9

[9] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022. 9

[10] Thilini Cooray, Ngai-Man Cheung, and Wei Lu. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[11] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18270–18279, 2022. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 6, 9

[13] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 9

[16] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021. 3

[17] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 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*, pages 18166–18176, 2022. 1, 2, 3

[18] Mikhail Fain, Niall Twomey, Andrey Ponikar, Ryan Fox, and Danushka Bollegala. Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to sota. *arXiv preprint arXiv:1911.12763*, 2019. 3, 6, 10

[19] Jill Freyne and Shlomo Berkovsky. Intelligent food planning: Personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 321–324, New York, NY, USA, 2010. Association for Computing Machinery. 1

[20] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. 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*, pages 14570–14580, 2020. 6, 10

[21] Ricardo Guerrero, Hai X Pham, and Vladimir Pavlovic. 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*, pages 3192–3201, 2021. 3, 6, 8, 10

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation

learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[24] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 5

[28] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 3

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2

[30] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3, 8

[31] Jiao Li, Jialiang Sun, Xing Xu, Wei Yu, and Fumin Shen. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 173–182, New York, NY, USA, 2021. Association for Computing Machinery. 6, 10

[32] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. *Hybrid Fusion with Intra- and Cross-Modality Attention for Image-Recipe Retrieval*, page 244–254. Association for Computing Machinery, New York, NY, USA, 2021. 6, 10

[33] Lin Li, Ming Li, Zichen Zan, Qing Xie, and Jianquan Liu. *Multi-Subspace Implicit Alignment for Cross-Modal Retrieval on Cooking Recipes and Food Images*, page 3211–3215. Association for Computing Machinery, New York, NY, USA, 2021. 6, 10

[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[35] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 3

[36] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online, July 2020. Association for Computational Linguistics. 3

[37] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online, July 2020. Association for Computational Linguistics. 3

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 5

[39] Zejun Li, Zhihao Fan, Huaixiao Tou, and Zhongyu Wei. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. *arXiv preprint arXiv:2201.12596*, 2022. 3

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[42] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 5

[43] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019. 2, 3, 6, 7, 9, 10

[44] Niki Martinel, Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. A structured committee for food recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 484–492, 2015. 1

[45] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 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*, pages 393–401, 2020. 8

[46] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2calories: Towards an automated mobile vision food diary. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1233–1241, 2015. 3

[47] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi Al Hammouri, and Antonio Torralba. 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*, pages 509–518, 2017. 1

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[49] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA, 2011. Curran Associates Inc. 2, 6

[50] Dim P Papadopoulos, Enrique Mora, Nadiia Chepurko, Kuan Wei Huang, Ferda Ofli, and Antonio Torralba. Learning program representations for food images and cooking recipes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16559–16569, 2022. 3, 10

[51] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 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*, pages 180–189. Springer, 2018. 3, 8, 9

[52] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 8

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[55] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 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*, pages 15475–15484, 2021. 1, 3, 4, 6, 8, 10

[56] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3, 6, 10

[57] Sarto Sara, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *19th International Conference on Content-based Multimedia Indexing*, 2022. 3

[58] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[59] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[60] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. 4

[61] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[62] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. 3

[63] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. In *33rd British Machine Vision Conference (BMVC)*, 2022. 3, 5

[64] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4567–4578, 2022. 1, 3, 4, 5, 6, 8, 9, 10, 12

[65] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021. 3

[66] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3

[67] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-

linguistic representations. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[68] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 3

[69] Mohammed Suhail and Leonid Sigal. Mixture-kernel graph attention network for situation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[70] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[71] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1, 2

[72] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. Learning structural representations for recipe generation and food retrieval. *arXiv preprint arXiv:2110.01209*, 2021. 6, 10

[73] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. 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*, pages 11572–11581, 2019. 3, 6, 10

[74] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and CH Steven Hoi. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia*, 2021. 3, 6, 10

[75] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 1, 3

[76] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 3

[77] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. 3

[78] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1

[79] Zhongwei Xie, Ling Liu, Lin Li, and Luo Zhong. 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*, pages 2221–2230, 2021. 5

[80] Zhongwei Xie, Ling Liu, Yanzhao Wu, Lin Li, and Luo Zhong. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. *IEEE Transactions on Services Computing*, 2021. 6, 10

[81] Zhongwei Xie, Ling Liu, Yanzhao Wu, Luo Zhong, and Lin Li. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–27, 2021. 6, 10

[82] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3

[83] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 3

[84] Zichen Zan, Lin Li, Jianquan Liu, and Dong Zhou. *Sentence-Based and Noise-Robust Cross-Modal Retrieval on Cooking Recipes and Food Images*, page 117–125. Association for Computing Machinery, New York, NY, USA, 2020. 6, 10

[85] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 5, 9

[86] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3, 10

[87] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 3

[88] Bin Zhu and Chong-Wah Ngo. Cookgan: Causality based text-to-image synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5518–5526, 2020. 3

[89] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 6, 10