FoodGen: Generating and Captioning Food Images with Vision-Language Models

Karina Abubakirova
Mohamed Bin Zayed University of Artificial Intelligence
karina.abubakirova@mbzuai.ac.ae

Khawla Ali Hasan Ali Almarzooqi
Mohamed Bin Zayed University of Artificial Intelligence
khawla.almarzooqi@mbzuai.ac.ae

Nazira Dunbayeva
Mohamed Bin Zayed University of Artificial Intelligence
nazira.dunbayeva@mbzuai.ac.ae

Abstract

Food image captioning poses unique challenges due to the fine-grained nature of culinary categories and the scarcity of high-quality textual annotations. We introduce **FoodGen**, a fully self-supervised and domain-adaptive framework that generates semantically rich and visually grounded captions for food images—without relying on manual labeling. Our pipeline combines instruction-tuned captioning (InstructBLIP), semantic filtering (CLIPScore and a novel SLA Rank), and optional linguistic refinement (LLaMA2) to curate high-quality pseudo-captions. We fine-tune InstructBLIP on multiple supervision regimes, including filtered pseudo-labels, LLM-refined outputs, and a human-annotated benchmark (Food500Cap). Comprehensive experiments across Food-101 (unlabeled) and Food500Cap (labeled) demonstrate significant gains in both reference-based (BLEU, METEOR, CIDEr) and reference-free (CLIPScore, SLA Rank) metrics. Our proposed SLA Rank further enables fine-grained alignment assessment between captions and food labels, complementing existing semantic retrieval scores. FoodGen achieves substantial improvements over zero-shot baselines, underscoring the importance of domain-specific adaptation in vision-language tasks. Our framework offers a scalable solution for grounded caption generation in fine-grained domains. Code and data are publicly available at: https://github.com/nazirad19/ai702-finalproject.

1 Introduction

Image captioning, the task of generating natural language descriptions for images—has seen remarkable progress with the advent of large-scale vision-language models (VLMs) such as BLIP-2 Li et al. [2023], InstructBLIP Dai et al. [2023], and GPT-4V. While these models perform well on general web-scale datasets, they often underperform in domain-specific settings that demand fine-grained visual understanding. One such domain is food imagery, where effective captions must capture subtle distinctions in ingredients, texture, presentation, and style—factors that are frequently overlooked by general-purpose systems.

Despite growing interest in food-related applications—from dietary tracking and recipe generation to restaurant recommendation—progress is hindered by the scarcity of richly annotated datasets.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Widely used datasets like Food-101 Bossard et al. [2014] provide only coarse class labels, while more detailed resources such as Food500Cap Ma et al. [2023] remain costly and time-consuming to scale due to the need for expert annotations. Moreover, standard evaluation metrics like CLIPScore and SLA often misalign with human preferences, rewarding literal or short captions over more fluent, informative descriptions.

To address these challenges, we introduce **FoodGen**, a fully self-supervised and domain-adaptive pipeline for generating high-quality captions for food images without relying on manual labels. FoodGen integrates instruction-tuned vision-language models, semantic filtering, and large language model (LLM) refinement to produce captions that are both semantically grounded and linguistically expressive.

The FoodGen pipeline operates in four stages: (1) initial caption generation using InstructBLIP, (2) semantic filtering via CLIPScore Hessel et al. [2021] and SLA Rank Ma et al. [2023], (3) optional refinement using LLaMA2 Touvron et al. [2023] to enhance fluency and domain specificity, and (4) fine-tuning of InstructBLIP on the curated pseudo-caption corpus. This enables FoodGen to adapt to fine-grained visual distinctions without any human-in-the-loop supervision. We evaluate FoodGen on both Food-101 (pseudo-labeled) and Food500Cap (human-annotated) datasets using a broad set of evaluation criteria, including both retrieval-based metrics (CLIPScore, SLA Rank, BLIP2-ITM) and reference-based metrics (BLEU Papineni et al. [2002], METEOR Banerjee and Lavie [2005], ROUGE-L Lin [2004], and CIDEr Vedantam et al. [2015]). Our experiments reveal key trade-offs between semantic alignment and linguistic fluency, and highlight limitations of existing metrics in capturing human-desired caption quality.

Contributions

Our contributions are as follows:

- We propose **FoodGen**, a modular and scalable pipeline for food image captioning that combines instruction-tuned captioning, semantic filtering, and LLM-based refinement—entirely without human-labeled data.
- We introduce SLA Rank as a complementary semantic evaluation metric and identify the misalignment between embedding-based metrics and human judgment in the food domain.
- We show that fine-tuning InstructBLIP on filtered and refined captions significantly improves caption quality across both alignment and reference-based metrics, outperforming state-of-the-art baselines such as BLIP, GIT, and OFA.
- We provide a high-quality, pseudo-labeled version of Food-101 with over 20,000 refined captions, enabling scalable experimentation in low-annotation regimes.
- Our work establishes a robust benchmark for fine-grained, domain-specific captioning and lays the groundwork for adapting vision-language models to other structured visual domains.

2 Related Work

2.1 Image Captioning

Image captioning has evolved from early CNN-RNN models Vinyals et al. [2015], Xu et al. [2015], Karpathy and Fei-Fei [2015] to Transformer-based architectures Cornia et al. [2020], Huang et al. [2021] and large-scale vision-language pretraining Li et al. [2020], Chen et al. [2020]. Pretrained VLMs such as BLIP Li et al. [2022], GIT Wang et al. [2022], Flamingo Alayrac et al. [2022], and BLIP-2 Li et al. [2023] unify vision-language understanding via contrastive, matching, and generative objectives, achieving strong results on general datasets like COCO Lin et al. [2014].

Instruct-tuned models (e.g., InstructBLIP Dai et al. [2023]) improve generalization through prompt-based supervision, but often underperform in visually complex, domain-specific settings, such as food or medical imagery, where fine-grained features and vocabulary are essential.

2.2 Domain-Specific Captioning

Domain-specific captioning has been explored in areas such as fashion Ak et al. [2019], medical imaging Boecking et al. [2022], and food Marin et al. [2019]. However, most food-related datasets (e.g., Food-101 Bossard et al. [2014], Recipe1M Salvador et al. [2017]) focus on classification or retrieval and lack descriptive captions. Food500Cap Ma et al. [2023] addresses this by introducing 24K human-annotated captions across 500 food categories with regional tags, enabling fine-grained evaluation.

Nonetheless, general-purpose VLMs like CLIP Radford et al. [2021] and BLIP struggle with cultural nuances and ingredient-specific semantics in food images Ma et al. [2023]. FoodLMM Yin et al. [2023] and LLaVA-Chef Mohbat and Zaki [2024] tackle this via instruction tuning and domain-curated metadata, though both rely heavily on labeled supervision.

2.3 Self-Supervised Captioning with CLIP-based Filtering

To mitigate the cost of manual annotation, recent work has adopted self-supervised pipelines that generate pseudo-captions using pretrained generative models and filter them based on alignment metrics. While CLIP Radford et al. [2021] is not generative, its image-text embeddings enable scalable retrieval and scoring. CLIPScore Hessel et al. [2021], based on cosine similarity in the CLIP embedding space, offers a reference-free metric that correlates well with human judgment.

We follow this direction by generating captions with InstructBLIP, refining them using LLaMA2 Touvron et al. [2023], and selecting high-quality outputs based on CLIPScore. This yields a scalable, domain-aligned caption corpus without requiring manual labels.

2.4 Language Model Refinement and Instruction Tuning

Instruction tuning improves model responsiveness to task-specific prompts. InstructBLIP Dai et al. [2023] demonstrates strong zero-shot generalization, but may hallucinate or under-describe fine-grained elements in domain settings. To address this, we incorporate a language refinement stage using LLaMA2 Touvron et al. [2023], which enhances fluency and factual correctness. This hybrid approach combines grounded vision-language modeling with language model fluency for robust caption quality.

2.5 Caption Evaluation

Traditional metrics such as BLEU Papineni et al. [2002], METEOR Banerjee and Lavie [2005], ROUGE Lin [2004], and CIDEr Vedantam et al. [2015] require references and are suboptimal in open-domain or self-supervised settings. Embedding-based metrics like CLIPScore Hessel et al. [2021] and BLIP2-ITM provide semantic alignment sensitivity without ground truth. SLA Rank Ma et al. [2023] further proposes task-specific coverage metrics for food attributes.

In this work, we rely on CLIPScore for both caption selection and evaluation, ensuring semantic fidelity in the absence of labeled references.

3 Methodology

We propose **FoodGen**, a self-supervised, domain-adaptive framework for food image captioning that operates without human-provided labels. Leveraging recent advances in vision-language modeling, FoodGen generates, filters, and optionally refines pseudo-captions to fine-tune a captioning model for the food domain. Figure 1 outlines the five-stage pipeline: (1) caption generation, (2) semantic filtering, (3) optional refinement, (4) domain-specific fine-tuning, and (5) evaluation.

3.1 Data Preparation

We use the Food-101 dataset Bossard et al. [2014] consisting of 101,000 images across 101 food categories. Since it lacks textual annotations, we apply stratified sampling to select 20,000 images, ensuring balanced class representation while maintaining computational feasibility. All images are

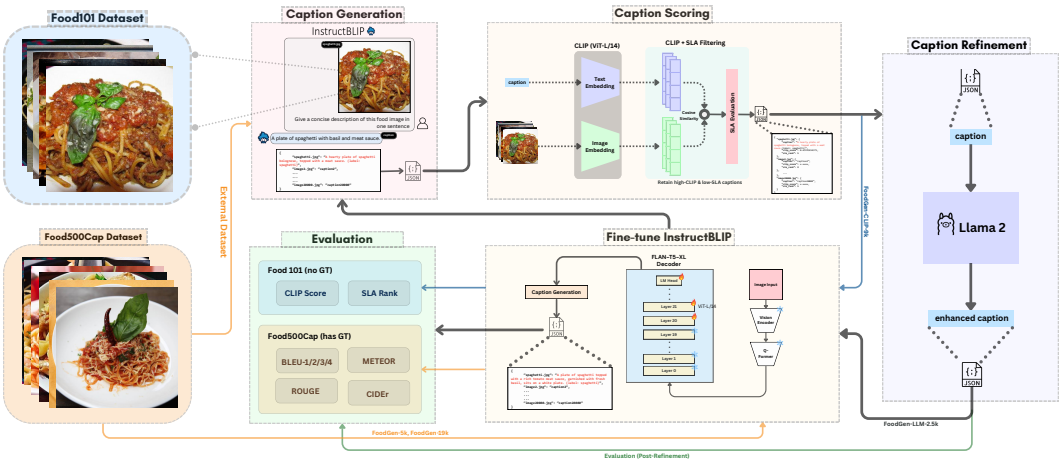

Figure 1: Overview of the FoodGen pipeline. Unlabeled food images are captioned by InstructBLIP, filtered using CLIPScore and SLA Rank, optionally refined by LLaMA2, and used to fine-tune the captioning model. Evaluation uses both reference-free and reference-based metrics

resized to 224×224 pixels to match transformer-based encoders. Importantly, class labels are withheld to preserve the self-supervised setting.

For supervised fine-tuning and reference-based evaluation, we employ the Food500Cap dataset, which contains high-quality, expert-written captions. This dataset is used exclusively in explicitly supervised settings.

3.2 Caption Generation with InstructBLIP

We initiate pseudo-captioning using InstructBLIP-FlanT5-XL Dai et al. [2023], a state-of-the-art vision-language model trained for instruction-following. Captions are generated using the prompt:

> *"Give a short, concise description of this food image in one sentence. Do not add extra details or guess unobservable context."*

Decoding is performed using beam search (beam size = 3, max length = 75 tokens). Minimal post-processing is applied to correct grammar and eliminate redundancy. Due to varying caption quality, subsequent filtering is essential to ensure semantic reliability.

3.3 Semantic Filtering with CLIP and SLA Rank

To assess and rank caption relevance, we compute **CLIPScore** Hessel et al. [2021] using CLIP ViT-L/14 Radford et al. [2021], defined as:

$$\text{CLIPScore}(I, T) = 2.5 \cdot \frac{\langle f_I(I), f_T(T) \rangle}{|f_I(I)||f_T(T)|} \tag{1}$$

We retain the top-K captions per image. To further enhance semantic coherence and reduce noisy examples, we introduce **Semantic-Linguistic Alignment (SLA) Rank**, which measures cosine similarity among candidate captions for a single image. Captions with unusually high SLA Rank (indicative of over-generic phrasing) are filtered out. This dual-filtering mechanism encourages both relevance and diversity.

3.4 Caption Refinement via LLaMA2

To improve linguistic fluency and domain specificity, we optionally refine filtered captions using LLAMA2-7B Touvron et al. [2023]. The model is prompted with:

> *"Rewrite the following food description in one sentence, making it more vivid and natural. Mention the food item clearly without repetition or generic phrasing."*

Responses are sampled using top-$p = 0.95$ and temperature $T = 0.7$. This refinement stage yields a smaller, high-quality corpus of approximately 3,000 captions that exhibit improved naturalness without sacrificing grounding.

3.5 Fine-Tuning InstructBLIP

We adapt InstructBLIP to the food domain using three types of supervision:

- **Filtered pseudo-captions** (CLIP + SLA).
- **LLaMA2-refined captions**.
- **Ground-truth captions** from Food500Cap.

Fine-tuning is conducted by freezing the vision encoder, unfreezing the final two transformer blocks and the language modeling head, and integrating LoRA adapters for parameter-efficient learning. Models are trained for 5 epochs using Adam (lr $= 1 \times 10^{-5}$, batch size = 1), with early stopping based on validation loss.

3.6 Inference and Evaluation

We evaluate model performance under two settings:

- **Food-101 (Unsupervised)**: Evaluated using reference-free metrics only.
- **Food500Cap (Supervised)**: Enables reference-based evaluation using human-written ground truth.

This multi-perspective evaluation provides comprehensive insight into both semantic fidelity and linguistic quality under different supervision regimes.

4 Experiments

4.1 Datasets

Food101 Bossard et al. [2014] includes 101,000 images across 101 categories without captions. We use it for reference-free evaluation (CLIPScore, SLA Rank).

Food500Cap Lab [2024] contains 25,000 food images with human-authored captions across 500 classes. We use 19,000 images for training and 6,000 for testing. Evaluation includes BLEU, METEOR, ROUGE-L, and CIDEr.

All images are resized to 224×224, and all captions are lowercased and cleaned.

4.2 Experimental Setup

Model. We fine-tune `instructblip-flan-t5-xl`, freezing the vision encoder and unfreezing the last 6 decoder layers and language modeling head. Training spans 3 epochs using AdamW ($\texttt{lr} = 3 \times 10^{-5}$), label smoothing (0.1), and delayed weight decay. Batch size ranges from 2 to 8.

Prompting. Prompts follow the instruction: *"Give a short, concise description of this food image in one sentence. Do not add extra details or guess unobservable context."* Inference uses beam search with $\texttt{num\_beams} = 5$, $\texttt{top\_p} = 0.7$, and $\texttt{max\_length} = 75$.

4.3 Evaluation Protocol

Reference-Free Metrics:

- **CLIPScore** Hessel et al. [2021]: Measures semantic alignment between image and caption via cosine similarity of CLIP embeddings.
- **SLA Rank**: An intra-set retrieval metric that ranks captions based on their cosine similarity to the corresponding image, assessing discriminative grounding.
- **BLIP2 ITM** Li et al. [2023]: Image-text matching probability score (optional).

Reference-Based Metrics (Food500Cap only):

- **BLEU-1/2/3/4** Papineni et al. [2002]: n-gram precision-based overlap.
- **ROUGE-L** Lin [2004]: Longest common subsequence recall.
- **METEOR** Banerjee and Lavie [2005]: Harmonic mean of precision and recall with synonym matching.
- **CIDEr** Vedantam et al. [2015]: Consensus-based captioning score.

4.4 Ablation Studies and Model Variants

Model Variants. We develop four FoodGen variants with differing supervision quality, dataset scale, and fine-tuning depth (Table 1).

Table 1: Training configurations for FoodGen variants

Variant	Dataset	Size	Decoder Blocks	Unfrozen Layers	Focus
FoodGen-LLM-2.5k	LLaMA-Refined (Food101)	2.5K	2	Last 2 + LM head	Rich, verbose captions
FoodGen-CLIP-9k	CLIP-filtered (Food101)	9K	2	Last 2 + LM head	High visual alignment
FoodGen-5k	Food500Cap Subset	5K	6	Last 6 + LM head	Human-authored fluency
FoodGen-19k	Full Food500Cap	19K	6	Last 6 + LM head	Scaled-up supervision

4.5 Quantitative Results

4.5.1 Reference-Free Evaluation

We evaluate all models on Food101 using CLIPScore and SLA retrieval metrics. As shown in Table 2, FoodGen-19k achieves the best SLA alignment, while FoodGen-5k delivers strong fluency with marginally lower grounding performance.

Table 2: Reference-Free Evaluation using Food101

Model	CLIPScore	SLA Rank	SLA@1
InstructBLIP (baseline)	0.281	59.96	11.0%
FoodGen-5k	0.266	82.1	13.2%
FoodGen-19k	0.268	**79.4**	**16.5%**

4.5.2 Reference-Based Evaluation

Table 3 summarizes performance on the Food500Cap test split. FoodGen-5k substantially improves over the baseline in CIDEr (from 0.15 to 0.39), BLEU, ROUGE, and METEOR, demonstrating strong gains from curated human supervision.

Experiment	AVG CLIP	SLA@1	Avg SLA	BLEU-1	BLEU-4	ROUGE-L	METEOR	CIDEr
InstructBLIP (our baseline)	0.276	1276	23.61	0.224	0.025	0.207	0.123	0.148
FoodGen-5k	0.266	1315	25.92	**0.278**	**0.046**	**0.269**	0.137	**0.393**
FoodGen-19k	0.268	1003	21.64	0.247	0.043	0.258	**0.164**	0.305
Ground Truth	0.269	1282	24.90	-	-	-	-	-

Table 3: Evaluation on the Food500Cap dataset. BLEU, ROUGE, METEOR, and CIDEr reflect reference-based quality. SLA/CLIP scores assess retrieval alignment.

4.5.3 Cross-Experiment Comparison

The comparison in Table 4 underscores key trade-offs. While FoodGen-5k leads in fluency and reference metrics, FoodGen-CLIP-9k excels in visual grounding. FoodGen-19k balances both fluency and alignment, benefiting from scaled human supervision.

Table 4: Cross-Experiment Metric Comparison

Metric	FoodGen-LLM-2.5k	FoodGen-CLIP-9k	FoodGen-5k	FoodGen-19k
Avg SLA Rank	81.38	**74.96**	82.09	79.39
SLA@1	1,608	1,722	**1,762**	1,654
Avg CLIP Score	0.268	**0.2705**	-	0.2678
CIDEr	–	–	**0.3935**	0.305

Therefore,

- **FoodGen-LLM-2.5k**: Verbose captions, weaker alignment.
- **FoodGen-CLIP-9k**: Best alignment, but lacks fluency.
- **FoodGen-5k**: Strongest overall fluency and caption quality.
- **FoodGen-19k**: Scales well with human data, but risks generic outputs.
- Using 6 decoder layers improved language quality across all experiments.

4.6 Comparison to Prior Work

Table 5 benchmarks FoodGen-5k against prior models on Food500Cap. Our model surpasses existing methods across all reference-based metrics, achieving a +25.7 CIDEr gain over BLIP_{Dec}.

Table 5: Comparison with prior work on Food500Cap Ma et al. [2023]. FoodGen-5k significantly outperforms baselines across all metrics

Model	BLEU@4	METEOR	ROUGE-L	CIDEr
BLIP_{Dec}	2.61	8.71	20.33	13.62
GIT	2.00	8.81	16.78	9.92
OFA	2.64	9.14	17.89	14.01
InstructBLIP (our baseline)	2.50	12.30	20.70	14.80
FoodGen-5k	**4.60**	**13.70**	**26.90**	**39.30**

4.6.1 Retrieval Evaluation on Food101

We additionally compare BLIP, BLIP-2, and FoodGen using CLIP-based SLA retrieval on Food101 (Table 6). FoodGen-5k surpasses both baselines, achieving +357 improvement in SLA@1 over BLIP-2.

4.7 Qualitative Results

Figure 2 presents representative samples across training stages, highlighting progressive gains in caption fluency, specificity, and visual-semantic alignment.

Table 6: Retrieval metrics on Food101 (higher SLA@1 = better).

Model	CLIP	SLA Rank ↓	SLA@1	SLA@5	SLA@10	SLA@20
BLIP	0.267	130.0	641	1907	3100	4887
BLIP-2	0.276	58.9	1870	4676	6555	8833
FoodGen-5k	**0.305**	**32.8**	**2268**	**6031**	**8412**	**10891**

Figure 2: Qualitative comparison of generated captions across three stages. Fine-tuned captions show notable improvements in fluency and label alignment

Performance Analysis. Baseline captions, though fluent, often lack semantic granularity—frequently resorting to generic terms (e.g., "oysters") without specifying preparation or ingredients. This gap is reflected in modest CLIPScore and retrieval accuracy (e.g., $SLA@1 \simeq 11\%$).

In contrast, **FoodGen-5k** yields more precise and contextually rich descriptions, capturing dish-specific nuances with higher consistency. Notably, it balances linguistic fluency with stronger visual grounding. However, models trained on pseudo-labeled data (e.g., **FoodGen-CLIP-9k**) tend to prioritize alignment over naturalness, often generating mechanically accurate yet less human-like captions.

These results underscore the importance of supervision quality in domain-specific captioning and suggest promising directions for harmonizing natural language generation with vision-grounded semantics.

5 Limitations

Despite the effectiveness of the FoodGen framework, it does have some limitations. The model is primarily designed for food image captioning and may not generalize well to other domains without significant adaptation. Additionally, its reliance on domain-specific datasets like Food-101 and Food500Cap limits its ability to handle diverse or rare food categories. The system also depends on large pretrained models, which require substantial computational resources, potentially limiting real-time deployment in resource-constrained environments. Future work will aim to address these issues by incorporating more diverse datasets, improving model generalization, and enhancing evaluation methods.

6 Conclusion

We presented **FoodGen**, a self-supervised and domain-adaptive framework for food image captioning that bridges the gap between generic vision-language models and fine-grained, domain-specific understanding. By combining instruction-tuned captioning, semantic filtering through CLIPScore and SLA Rank, and optional LLM-based refinement, FoodGen generates high-quality captions without the need for manual annotations. Our results demonstrate that domain-specific fine-tuning significantly improves both the fluency and semantic alignment of generated captions, as measured by standard reference-based metrics (BLEU, METEOR, CIDEr) and embedding-based alignment scores (CLIPScore, SLA Rank). Importantly, we show that supervision quality—not just quantity—is critical for effective adaptation, and that lightweight refinement (e.g., via LLaMA2) can further enhance linguistic quality without sacrificing alignment. Qualitative and quantitative evaluations across Food-101 and Food500Cap confirm the effectiveness of our approach, especially in capturing subtle visual distinctions critical to food understanding. We also propose SLA Rank as a practical and interpretable measure of semantic-label alignment, complementing existing metrics. Future directions include integrating interactive feedback mechanisms, extending to few-shot learning for unseen categories, and incorporating multimodal reasoning for richer, more grounded captions. This work lays the foundation for more accurate and context-aware captioning in food and other structured visual domains.

References

Kamal Ak, Mekhal Gul, Zhenhua Ma, and Yi Yang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *ICCV Workshops*, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, Antoine Miech, Alexander Ivanov, Ian Barr, Joanna Bitton, Po-Sen Huang, Arthur Mensch, Kyle Millican, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, 2005.

Benedikt Boecking, Naoya Usuyama, Siddharth Bannur, Victor Valenko, and Jonathan Raiman. Making the most of text semantics to improve biomedical vision–language processing. *Nature Communications*, 13(1): 1–10, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

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 *ECCV*, 2020.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

Zhiliang Dai, Zhe Zhu, Yuwei Zhang, Zhenhai Wang, Junnan Li, and Steven C.H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021.

Hao Huang, Xinlei Wang, Yanzhang Zhu, Mohit Bansal, Dongju Lee, Amanpreet Singh, Xiaowei Chen, Rowan Zellers, Mark Yatskar, and Devi Parikh. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

AdvancedCV Lab. Food500cap: Curated dataset with ground-truth captions, 2024. https://github.com/advancedcvlab/food500cap.

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

Junnan Li, Dongxu Zhang, and Steven CH Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

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 *ECCV*, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004.

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 *ECCV*, 2014.

Zhen Ma, Mingyue Pan, Wen Wu, Kaizhu Cheng, Jianfei Zhang, Shuaiqiang Huang, and Jianlong Chen. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *ACM MM*, 2023.

Javier Marin, Christian Abbet, Michael Gygli, and Luc Van Gool. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. In *TPAMI*, 2019.

Fady Mohbat and Mohammed J. Zaki. Llava-chef: A multi-modal generative model for food recipes. In *ACM WebConf*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

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. *International Conference on Machine Learning (ICML)*, 2021.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*, 2017.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Amanpreet Batra, Tim Bambroo, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

Zhe Wang, Jiyang Yang, Xiaohui Yang, Yixiao Bai, Yujun Tong, and Jiebo Luo. Git: A generative image-to-text transformer for vision and language. In *CVPR*, 2022.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Yao Yin, Hongyu Qi, Bin Zhu, Jun Chen, Y-G. Jiang, and Chong-Wah Ngo. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023.