Multimodal Contrastive Decoding for Image Captioning

Victor Wang
The University of Texas at Austin
victorwang37@utexas.edu

Will Beason
The University of Texas at Austin
beason@utexas.edu

Abstract

While Large Vision-Language Models have been the subject of much recent work, there is an open question of how to handle hallucinations resulting from linguistic priors. While these models usually correctly identify large features of images, they often generate irrelevant text that is otherwise linguistically probable. Contrastive Decoding is a recent technique that shows much promise in reducing over-reliance on such priors, reducing hallucinations and improving performance on language tasks. In this paper, we propose an application of Contrastive Decoding to multilingual image captioning. Specifically, we contrast an expert, which receives both an image and its caption, with one or two amateurs, which receive only either the image or the caption. We find that using only one amateur, the text-only one, works best. Our results show questionable improvements in generating English captions when provided a Chinese caption.

Code: https://github.com/dubai03nsr/multimodal-contrastive-decoding.

1 Introduction

Large Vision Language Models (LVLMs) have proven reliable in a variety of multimodal tasks (Radford et al., 2021; Lu et al., 2019; Zellers et al., 2021; Chen et al., 2023; Chowdhery et al., 2022). Multimodal machine translation (MMT), or multilingual image captioning, is a task that has served as a testbed for models to jointly reason over visual and textual content (Elliott et al., 2016). The inputs to the task are an image and a caption for the image, and the goal is to provide a caption for the image in another language. Unlike vanilla visual question answering (VQA) (Agrawal et al., 2016), where the content unilaterally lies in the image while the text serves merely as a prompt, MMT involves drawing on the content of both modalities. Furthermore, a practical motivation of MMT is the multimodal nature of human interaction, such as sending an email with a photo attachment. Indeed, such multimodal tasks have seen extension into other modalities, such as audio and video (Harwath et al., 2018; Wang et al., 2019).

In multimodal tasks like MMT, models can derive richer representations by jointly conditioning on the modalities than independently (Bagher Zadeh et al., 2018; Tsai et al., 2019). Nonetheless, undesirable biases from single-modal priors may remain present despite joint conditioning (Ramakrishnan et al., 2018). Therefore, there is potential for improvement by filtering out single-modal biases from the multimodal representation.

Contrastive Decoding (CD). CD is a technique for improving the performance of models by contrasting an "expert model" with a worse-performing "amateur model" that exhibits some undesirable behavior also present in the expert model (Li et al., 2023). In this paper, we propose contrasting a multimodal model with multiple single-modal models. That is, our expert model will be a multimodal model that combines text and image, and our amateur models will be either image-only or text-only. Since single-modal amateurs condition on incomplete information, we hypothesize that contrasting them against the expert will mitigate single-modal biases. Investigating the qualitative effects of such biases is outside the scope of this paper, but we do believe it is important future work – particularly measuring the prevalence of hallucinations relative to the modality.

2 Related Work

CD in Language. The idea of contrasting with respect to context in text generation has been explored (He et al., 2019; Li et al., 2016; Waldendorf et al., 2024; Shi et al., 2023). In the work of Li et al. (2023), the expert and amateur are frozen models of different sizes from the same family, and the amateur conditions on only the last context token

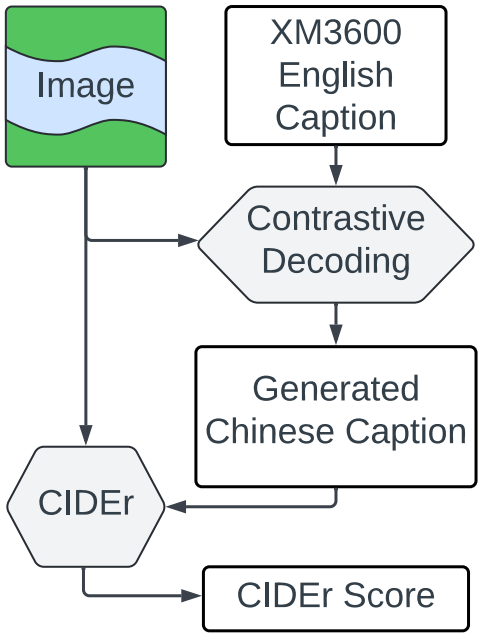

Figure 1: A captioning task from English to Chinese. CIDEr compares the generated Chinese text with the original image to determine the quality of the generated caption. We also generate tasks by providing XM3600 Chinese captions and requesting the LVLM generate a caption in English.

rather than the full context. Sennrich et al. (2024) apply CD to machine translation by giving the amateur model the wrong source text or the wrong target language.

Multimodal CD. Ramakrishnan et al. (2018) mitigate language bias by introducing a question-only adversary to learn image-dependent encodings and predictions. We note that this method technically differs from CD in that it proposes a training procedure rather than a decoding one. Leng et al. (2023) mitigate object hallucinations by giving the amateur a noisified image. A difference with our work is that we consider sentence-length generation and explore employing multiple amateurs. Subsequent to the beginning of our project, Zhu et al. (2024) find that using an image-biased model as the expert and the vanilla multimodal model as the amateur improves caption quality, corroborating our findings.

3 Method

3.1 Multimodal CD

In our approach to multimodal machine translation (MMT), the expert model receives the image and caption, the text amateur receives only the caption, and the image amateur receives only the image, as depicted in Figure 3.

Consider a given decoding step. All models condition on the prior generated tokens. Denoting the models' respective probability distributions (over tokens to decode) as $p_{\text{exp}}, p_{\text{txt}}, p_{\text{img}}$, we define the CD score distribution as

$$s_{\text{CD}} = \log p_{\text{exp}} - \lambda_{\text{txt}} \log p_{\text{txt}} - \lambda_{\text{img}} \log p_{\text{img}} \quad (1)$$

for hyperparameters $\lambda_{\text{txt}}, \lambda_{\text{img}}$. Following Li et al. (2023), we add a constraint to ensure we only decode tokens that the expert considers plausible. Denoting $\mathcal{V}$ as the set of vocabulary tokens, define

$$\mathcal{V}_\alpha = \{x \in \mathcal{V} \mid p_{\text{exp}}(x) \geq \alpha \max_{w \in \mathcal{V}} p_{\text{exp}}(w)\}$$

for a hyperparameter $\alpha \in [0, 1]$; we use $\alpha = 0.1$. We then generate

$$\operatorname*{argmax}_{x \in \mathcal{V}_\alpha} s_{\text{CD}}(x).$$

3.2 MiniCPM-V Model

The multilingual vision-language model we use is MiniCPM-V, a 3B parameter English/Chinese multilingual model capable of multimodal interactions in text and images (Hu et al., 2024). Per standard CD methodology, we freeze the model weights. Prompts are in Table 2.

4 Experiments

4.1 XM3600 Dataset

We use XM3600 (Thapliyal et al., 2022), a multilingual image captioning dataset consisting of 3,600 images from the Open Images dataset with 1-2 captions per image in each of 36 languages. For each supported language, there are about 100 images taken from a region in which the language is spoken, allowing us to analyze model bias by evaluating on geographic subsets of the dataset.

The annotation guidelines for XM3600 specify that each caption be one sentence long and depict the visual contents of the image. Thus, the image likely includes a strictly richer set of information compared to the caption, undermining the premise that the image and text offer complementary information. As compensation, we downsize the images by 5x5. Images in XM3600 have inconsistent aspect ratios, but all contain about the same area of 300,000 pixels, so downsizing yields images of about 12,000 pixels each.

The evaluation of our model on the dataset is depicted in Figure 1. We only consider the English and Chinese captions for this work.

4.2 Evaluation Metrics

The task at hand can be formulated either as text-assisted image captioning or as image-assisted machine translation. For each unassisted task, the conventional metrics differ. We believe that for the XM3600 dataset we have chosen, image captioning is the more fitting task formulation, because the annotation guidelines specify to write a caption that describes the visual contents of the image, and the captions of different languages are not written together or meant to be translations of each other. Nevertheless, we include a conventional metric from each task: CIDEr (Vedantam et al., 2015), a lexical metric, and COMET (Rei et al., 2020), a model-based metric. Model-based metrics have the advantage that they capture semantics in a more nuanced way, but lexical metrics have the advantage that they scale to many languages without dependence on language-specific model training and performance. Indeed, CIDEr is the metric of choice in the work introducing XM3600 (Thapliyal et al., 2022).

4.3 Tuning $\lambda_{\text{txt}}, \lambda_{\text{img}}$

Figure 2 shows the results of tuning $\lambda_{\text{txt}}, \lambda_{\text{img}}$ (Equation 1) on a subset of XM3600 consisting of 100 random images. A value of 0 on the x-axis corresponds to the vanilla multimodal expert model. When scaling $\lambda_{\text{txt}}, \lambda_{\text{img}}$ in step (left), performance begins similar to the expert but soon drops. We then try one amateur at a time (center, right) and find that the text amateur alone with $\lambda_{\text{txt}} = 0.1$ works best. This result is consistent with the finding from Zhu et al. (2024), that it is preferable to contrast more against language content. Although the trends for $\text{en}\rightarrow\text{zh}$ and $\text{zh}\rightarrow\text{en}$ swap when swapping the CIDEr and COMET metrics, $(\lambda_{\text{txt}}, \lambda_{\text{img}}) = (0.1, 0)$ appears the best overall, and is what we use for the remaining experiments.

4.4 Main Results

Table 1 shows the results on the full XM3600 (3,600 images) as well as the subsets corresponding to regions where the source or target language is spoken (100 images each). Based on the CIDEr metric, our model improves performance on $\text{zh}\rightarrow\text{en}$ for the whole dataset or the English subset. However, the COMET metric reveals far less distinction. Notably, both metrics report a significant drop with $\lambda_{\text{txt}} = 0.1$ for the $\text{en}\rightarrow\text{zh}$ direction on the Chinese subset.

5 Conclusion

Consistent with related work in machine translation, using contrastive decoding to improve performance vis-a-vis an overreliance on linguistic priors does improve performance in some circumstances. This improvement appears specific to the specific pair of languages and the direction of translation, as shown by the different behavior between captioning in Chinese when given English versus the inverse. Further, where there is an improvement, we find the improvements to be most significant when giving a small weight to the contrasted-against amateurs, and that reducing reliance on the text-only amateur yields larger performance gains than on the image-only amateur.

6 Future Work

A major limitation of our work is the lack of human evaluation. As such, we are unable to speculate qualitatively on why performance improved: whether this was due to a reduction in linguistic prior-induced hallucinations or some other factor. Future work should include human evaluation of the captions to look for patterns in the improvements, as this could lead to more targeted avenues of research.

There are several simple alternatives to our method that we did not try, but could yield good results as well. For one, we could use a text-only amateur that was provided double-translated text, the idea being that the double-translated text would contain more influence from linguistic priors. Similar to Leng et al. (2023), we could also try an amateur which receives an image at a lower resolution than the expert, still using our method of iterative decoding rather than a set of short answers.

Another open question is how to best combine multiple amateurs. Li et al. (2023) only considers a single amateur, whose log probabilities are weighted equally to those of the expert. Our work furthers this by attempting to combine multiple amateurs, but we only explored a small number of ways to combine them. Another option is to add a small number of trainable parameters to produce an input-dependent weighting. Different amateurs may express different types of biases that are de-

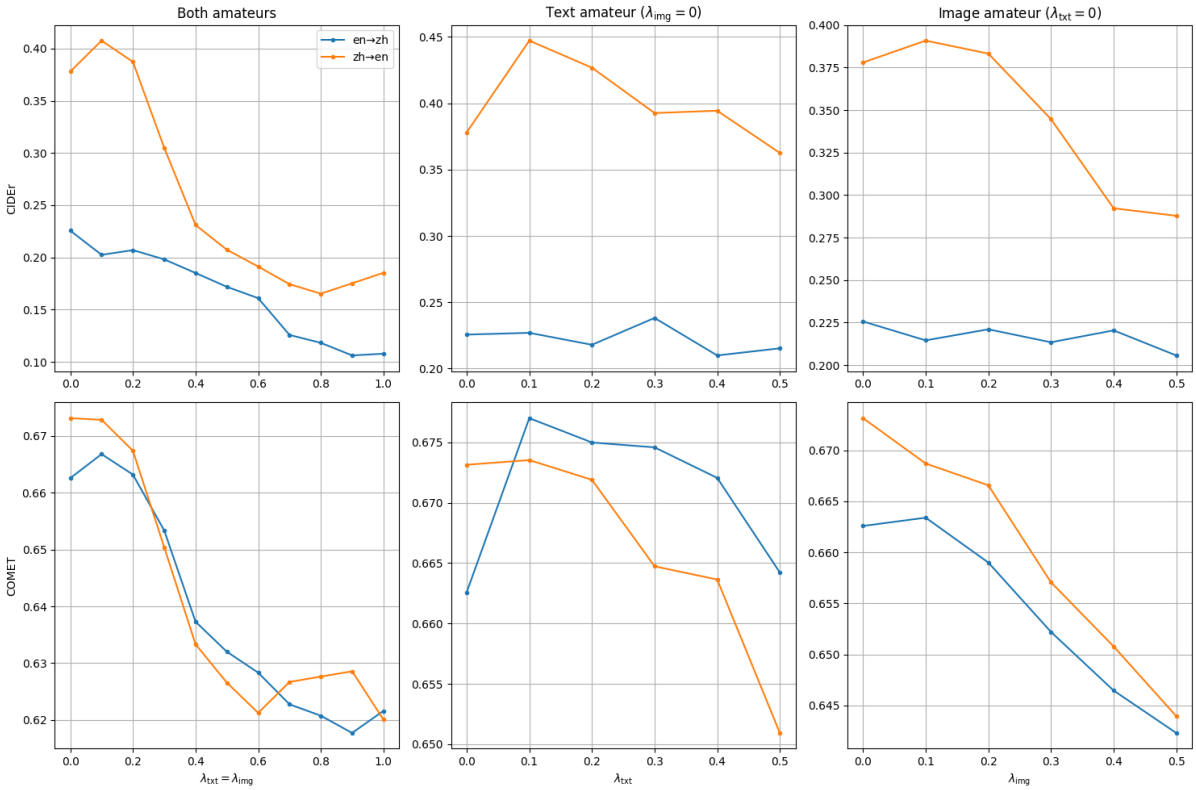

Figure 2: Tuning $\lambda_{\text{txt}}, \lambda_{\text{img}}$ on 100 images.

		region: all		region: en		region: zh	
		en→zh	zh→en	en→zh	zh→en	en→zh	zh→en
CIDEr	Expert	**0.216**	0.333	**0.272**	0.381	**0.217**	**0.410**
	$\lambda_{\text{txt}} = 0.1$	0.202	**0.351**	0.239	**0.419**	0.197	0.392
COMET	Expert	0.656	**0.669**	**0.670**	**0.667**	**0.626**	**0.656**
	$\lambda_{\text{txt}} = 0.1$	**0.660**	0.666	0.669	0.663	0.625	0.638

Table 1: Evaluation on XM3600 and geographic subsets.

sirable to reduce in models, and so a good way to merge them would make CD approaches even more flexible.

Lastly, we suggest this work should be repeated in other language pairs. Our inconsistent results in captioning in Chinese and English suggest that the behavior of CD approaches may be specific to language and the direction of translation. Thus, CD work should strive to be multilingual as it may not yield performance gains in every language.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. Pali: A jointly-scaled multilingual language-image model.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding.

Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations.

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

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590. IEEE.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. IBD: Alleviating hallucinations in large vision-language models via image-biased decoding.

A Detailed Contrastive Decoding Diagram

What follows is a detailed diagram of the contrastive decoding process we use.

B Prompts

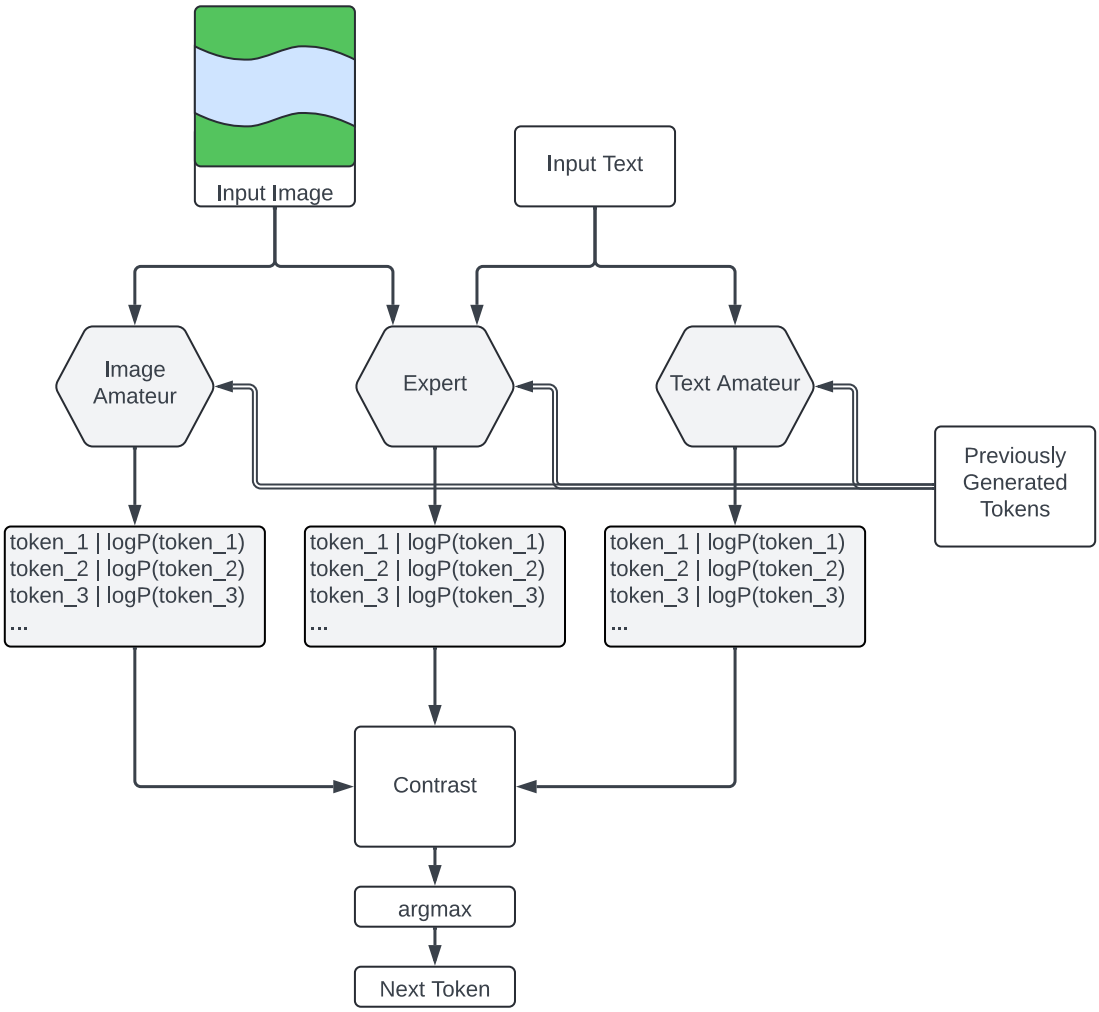

Figure 3: Detailed diagram of the contrastive decoding process, including the expert and both amateur models. For each generated token, every model receives the previously-generated tokens as input.

	Model	Prompt
en→zh	exp	这是图像的英文说明：`src_text` 用1句话中文描述这幅图像。
	txt	翻译成中文：`src_text`
	img	用1句话描述这幅图像。
zh→en	exp	Here is the Chinese caption of the image: `src_text` Describe the image in 1 sentence in English.
	txt	Translate this to English: `src_text`
	img	Describe the image in 1 sentence.

Table 2: Expert and amateur prompts.