VACoDe: Visual Augmented Contrastive Decoding

Sihyeon Kim[1 *] **Boryeong Cho**[1 *] **Sangmin Bae**[1] **Sumyeong Ahn**[2 †] **Se-Young Yun**[1 †]

Abstract

Despite the astonishing performance of recent Large Vision-Language Models (LVLMs), these models often generate inaccurate responses. To address this issue, previous studies have focused on mitigating hallucinations by employing contrastive decoding (CD) with augmented images, which amplifies the contrast with the original image. However, these methods have limitations, including reliance on a single augmentation, which is restrictive for certain tasks, as well as the high cost of using external knowledge. In this study, we address these limitations by exploring how to utilize multiple image augmentations. Through extensive experiments, we observed that different augmentations produce varying levels of contrast depending on the task. Based on this observation, we introduce a novel method called VACoDe, Visual Augmented Contrastive Decoding. This method adaptively selects the augmentation with a big contrast for each task using the proposed softmax distance metric. Our empirical tests show that VACoDe outperforms previous methods and improves output quality in various vision-language tasks. Additionally, VACoDe can be universally applied across different model types and sizes without additional training or the use of external models and data.

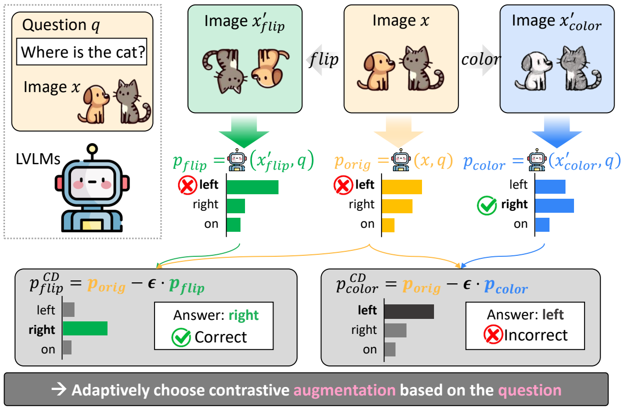

Question q
Where is the cat?
Image x
LVLMs
Image x'_{flip}
flip
Image x
color
Image x'_{color}
$p_{flip} = (x'_{flip}, q)$
left
right
on
$p_{orig} = (x, q)$
left
right
on
$p_{color} = (x'_{color}, q)$
left
right
on
$p^{CD}_{flip} = p_{orig} - \epsilon \cdot p_{flip}$
left
right
on
Answer: right
Correct
$p^{CD}_{color} = p_{orig} - \epsilon \cdot p_{color}$
left
right
on
Answer: left
Incorrect
→ Adaptively choose contrastive augmentation based on the question

Figure 1. Overview of the problem we focus on. When using LVLMs, visual augmentations have different effects on their outputs. For example, if the question is "where is the cat?" and the correct answer is *right*, applying a flip augmentation can alter the input image, resulting in a contrastive answer, *left*. This contrastive information is beneficial for increasing the answer's probability in CD. Conversely, using color augmentation for this question is unsuitable, as it does not generate a contrastive output distribution. Therefore, the main challenge is how to adaptively select the most effective augmentation to improve CD performance in LVLMs.

1. Introduction

Pre-trained Large Vision Language Models (LVLMs) (Liu et al., 2024a; Ye et al., 2023; Zhu et al., 2023; Dai et al., 2024; Li et al., 2022; 2023a; Radford et al., 2021) have gained prominence due to their capability to understand multiple data formats, especially vision and language, simultaneously. These models have demonstrated exceptional performance in various tasks like zero-shot image classification (Yao et al., 2021), image-text retrieval (Li et al., 2022), visual question answering (Liu et al., 2024a), and image captioning (Li et al., 2023a). Most recent large-scale VLMs, such as LLaVA (Liu et al., 2024a), MiniGPT-4 (Zhu et al., 2023), and InstructBLIP (Dai et al., 2024), utilize autoregressive transformers to expand their functionality, enabling them to generate more complex outputs.

However, language decoders sometimes produce incorrect outputs, a phenomenon often called hallucination. Among various methodologies (Wei et al., 2022; Rose et al., 2023; Shao et al., 2024), one promising approach is contrastive decoding (CD) (Li et al., 2023b), which generates final answers by examining multiple candidate responses and leveraging their contrastiveness. In detail, they operate in two stages: (1) generating output distributions given both the original and contrastive prompts, and (2) subtracting the two distributions to reduce the likelihood of hallucinated tokens. The effectiveness of this approach depends on how well the contrasting prompts are constructed.

[*]Equal contribution. [†]Corresponding author. [1]KAIST AI [2]CSE, Michigan State University. Correspondence to: Sumyeong Ahn <sumyeong@msu.edu>, Se-Young Yun <yunseyoung@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

There have been a few works on generating contrastive images (Leng et al., 2023; Wan et al., 2024; Chen et al., 2024), which aim to increase sample variance by manipulating features in images through the addition of noise or cropping. However, applying single augmentations to all samples cannot always guarantee contrastive images, as the salient features in a visual prompt may vary depending on the text question in vision-language tasks.

We illustrate an example in Figure 1, where the model receives position-related question "*where is the cat?*," and generates the incorrect output *left*, given the output distribution p_{orig}. If a position-related augmentation such as flipping is applied to a given image, the output distribution p_{flip} would likely be heavily skewed towards *left*. However, when applying augmentation that is less relevant to the target features (*i.e.,* position), such as color augmentation, the augmented output distribution p_{color} may be similar to p_{orig}. Consequently, the contrastive decoded logit $p^{\text{CD}}_{\text{color}}$ may have the wrong answer while $p^{\text{CD}}_{\text{flip}}$ corrects the answer. In light of this, to generate appropriate and sufficient contrastiveness to ensure the model provides the correct answer, selecting the proper augmentation operation is significantly required.

Contributions. In this paper, we address the challenge of enhancing contrastive decoding performance by formulating the selection of proper augmentation. Our contributions are summarized as follows:

- We explore the effect of visual augmentation on various vision-language tasks. Our findings indicate that applying different augmentation operations alters the output distribution of VLMs, subsequently affecting the response. From the CD perspective, the choice of proper augmentation is critical: selecting contrastive augmentations that introduce beneficial contrast can enhance performance, while unsuitable augmentations can lead to a decline in performance.

- Based on the findings, we introduce an algorithm called VACoDe that selects the most contrastive augmentation to empower CD capability without additional training or using external models. The algorithm consists of three main steps: (1) provide various types of augmented images to VLMs and generate multiple outputs. (2) Assess the difference between the original output distribution and the augmented output distributions. (3) Identify the most contrastive augmented image, characterized by the largest gaps, and produce the final output by CD.

- Extensive empirical tests verify that the proposed decoding method is superior to previous decoding techniques in VLMs. Furthermore, we observe evidence of why those augmentations work effectively in the contrastive decoding mechanism.

2. Preliminaries

Here, we provide a concise summary of background information to aid in understanding this research. We further provide the related literatures in Appendix G.

Visual data augmentation (VA).

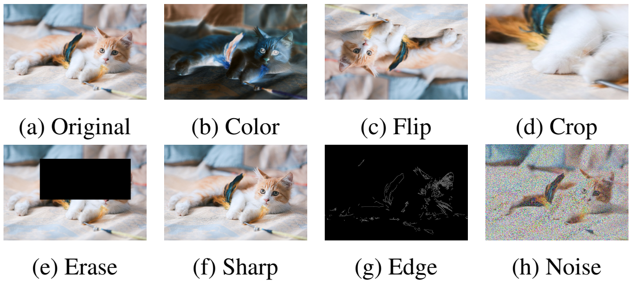


Figure 2. Visual augmentations utilized in this paper.

VA consists of long-established techniques that modify visual data to produce desired images for computer vision research, such as enhancing sharpness, adjusting color jitters, and more. These augmentation techniques are employed to increase the diversity of sample data, thereby mitigating overfitting issues in environments with limited samples. We focus on the framework:

$$v' = \mathcal{O}_o(v),$$

where o represents an augmentation operation within the set $\mathcal{A}$. The descriptions of the augmentations that we used are in Section A.1, and the examples are illustrated in Figure 2.

Contrastive decoding (CD). In NLP domain, it usually operates by generating two outputs using two different models: an expert model that produces the original outputs and an amateur model that generates contrastive outputs, then performs decoding based on the contradictions between them. It has also been explored in the VLM by using manipulated images to create contrastiveness. This method involves using the image to remove unrelated information, such as hallucinations, by subtracting the image with amplified contrastive information from the original one. The process operates as follows:

$$\begin{aligned} p_{\text{CD}}(y|v,\mathcal{O},q) = \texttt{SOFTMAX}\Big((1+\alpha)f(y|v,q) \\ -\alpha f(y|\mathcal{O}(v),q)\Big), \end{aligned} \tag{1}$$

where $f(\cdot)$ is the model output logit obtained from VLM and $\mathcal{O}(v)$ as augmented image. To amplify the hallucination inherent in the VLM, VCD (Leng et al., 2023) added noise to the image and CRG (Wan et al., 2024) used object-wise erasing with the provided bounding box labels.

3. VACoDe: Visual-Augmented Contrastive Decoding

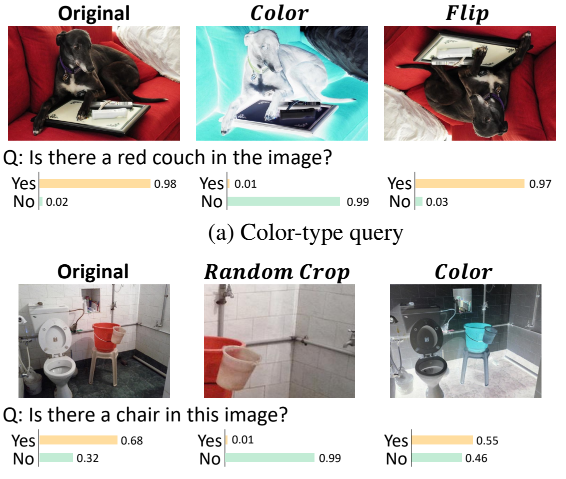
This section explores the impact of VAs on LVLM, focusing specifically on contrastive decoding. In essence, we demonstrate that certain VAs cause either contrast or persistence, implying that the output distribution either varies or stays consistent with the augmented image for the given query. Furthermore, we detail our discovery that contrastive augmentation can be identified using the proposed score, which relies on the softmax distance. Building on these insights, we present a novel algorithm named VACoDe, which leverages both the original and augmented images for CD.

3.1. Appropriate Augmentations Enhance the Contrast

Initially, we investigate the effect of VA on LVLM decoding. We hypothesize that a specific contrastive set of VAs exists for each query, causing it to lose critical features necessary to answer the given query correctly. To verify this, we manually select query type-contrastive augmentation pairs as described in Table 1 on the MME dataset (Fu et al., 2024), which provides questions of color, existence, and position categories. Note that each augmentation is contrastive to corresponding query types while persistent to other query types here. After that, we check its effectiveness by evaluating the output probability. The setting details for this investigation can be found in Section A.3.

Table 1. Manually selected query type-contrastive augmentation pairs.

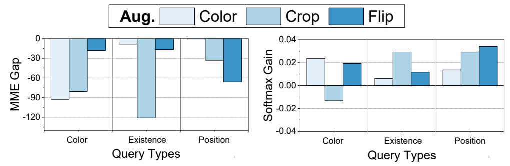
Query type	Contrast. aug.
Color	Color
Existence	Random Cropping
Position	Flip



(a) Color-type query

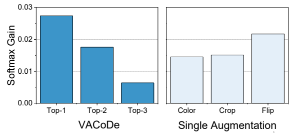
(b) Existence-type query

Figure 3. A detailed analysis of augmentation-question pairs reveals that (a) in color-type query, color augmentation produces a contrastive distribution, whereas flipping does not. Similarly, (b) shows that the existence query is influenced by random cropping.



(a) MME score drop. (b) The softmax gain.

Figure 4. On each question type in the MME dataset, (a) MME score drop of augmented images and (b) the softmax output gain after CD are measured on different augmentations.



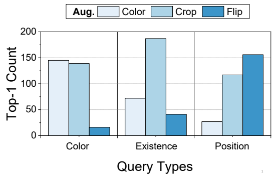


Figure 5. Softmax output of ground truth increases after CD. The top1 augmentation with the largest distance gets the best increment, which is higher than single augmentation result.

Figure 6. The number of selected augmentations with the largest distance D in each category shows that, for each category, contrastive augmentation was chosen most frequently.

Contrastive augmentations decrease performance. Using the manually selected query-augmentation pairs, we analyze their outputs both qualitatively and quantitatively. For qualitative analysis, as described in Figure 3, we show examples of contrastive and persistent augmentations for given queries using LLaVA-1.5 7B. As shown in Figure 3a, when the model is asked a color-related question with the original image, it provides the correct answer. However, when given the color-augmented image, which is a contrastive augmentation on the color-type query, the model generates an incorrect answer. Conversely, if a flip augmentation, *i.e.,* persistent augmentation, is applied, the model correctly predicts the answer. This is because flipping does not alter the critical color features that are needed to respond accurately. Similarly, for the existence-type query, as shown in Figure 3b, random cropping corresponds to a contrastive augmentation for the existence-type question since it removes the objects needed to answer correctly. For instance, the portion containing the "chair" may be cropped out, leading to an incorrect response. However, coloring the image does not affect the presence of the "chair," allowing the model to provide the correct answer. These results are also reflected in the quantitative analysis, as shown in Figure 4. It is conducted with the same setting using LLaVA-1.5 13B. Figure 4a shows the MME score difference when augmentation is applied. For instance, when the question type is "existence," applying random cropping to the input image lowers the MME score compared to using the original image. It means that contrastive augmentation can lead to a performance decrease.

Additionally, we aim to verify whether this contrastive augmentation can be advantageous in a CD setting. For this, we measure the softmax gain, called Gain score, as follows:

$$\begin{aligned}\texttt{Gain}(v, q, y_{\text{GT}}, \mathcal{O}) = p_{\text{CD}}(y_{\text{GT}}|v, \mathcal{O}, q) \\ -\texttt{SOFTMAX}(f(y_{\text{GT}}|v, q)).\end{aligned} \tag{2}$$

This score measures the increase of the softmax on the ground truth value from the original decoding to the CD output. As illustrated in Figure 4b, utilizing CD methods with contrastive augmentation results in a significant increase in the Gain score. For instance, in the case of existence-type questions, we observe the highest gain when using a random crop. Since we rely on impractical information, such as manually selected contrastive augmentation, the remaining challenge is to identify the contrastive augmentation for each query without human intervention. Moving forward, we focus on tackling this challenge.

3.2. Maximizing Contrast: Selecting Augmentation with the Biggest Distance

To address the challenge of selecting contrastive augmentation, we first set our intuition that the augmentation resulting in the most different output can serve as a contrastive augmentation. To measure the difference, we use one of the useful metrics, the L_2 norm, called distance D, defined as follows:

$$\begin{aligned}D(p(v), p(\mathcal{O}(v))) = \left\| \Big(p(v) - p(\mathcal{O}(v))\Big) \right\|_2 \\ \text{, where } p(v) = \texttt{SOFTMAX}(f(y|v, x)).\end{aligned} \tag{3}$$

Note that it can be changed to other types of distance metrics, such as the L_n norm, KL divergence, and so on. Analysis of this metric is also included in Appendix C.

Choosing augmentation with distance D. To verify our hypothesis – *a bigger distance D can have the most contrastiveness* – we measure distance D following Eq. (3) and the Gain score defined as Eq. (2), under the aforementioned experimental conditions. To examine the correlation between the distance D and the Gain score, we sort the augmentations based on the distance D, and analyze the average Gain score on each ranking. As shown in Figure 5, we confirm that the augmentation with the greatest D results in the biggest average increase in the Gain score. This implies that selecting the augmentation with the highest D yields the best performance improvement. Additionally, the top-ranked group shows a higher increase than other single augmentations. Moreover, Figure 6 shows how frequently each augmentation is selected as having the highest D score on each question type. The most frequent augmenations correspond to contrastive augmentation, which aligns with intuition in Section 3.1. This suggests using the augmentation with the largest D to select the contrastive augmentation o for each query q.

3.3. VACoDe: Visual-Augmented Contrastive Decoding

Based on the above observations, we propose VACoDe to automatically select an appropriate augmentation for each query by utilizing the distance D. The entire procedure is summarized in Algorithm 1.

In the initial decoding phase with the given question, we adapatively select contrastive augmentation by calculating the distance metric D and choosing the augmentation with the maximum distance. This chosen augmentation $\hat{o}$ is then used for the remainder of the sequence decoding process. Once the contrastive augmentation is determined, LVLM calculates word probability $p_{\texttt{VACoDe}}$ using Eq. (1). Subsequently, among the whole vocabulary V, the candidate word set $V_{\text{cand}} \in V$ is defined to select more reliable words following the original CD algorithm (Li et al., 2023b). This process is repeated iteratively to generate the output y.

For VACoDe, we use two scenarios to define candidate augmentations: *all* and *selection*. *All* uses all the augmentations as augmentation candidate set $\mathcal{A}$. However, some augmentations may be ineffective or replaceable. In this case, excluding these augmentations may work as a way to eliminate noisy augmentations. So we introduce the *selection* strategy that leverages validation to choose a subset of augmentations $\mathcal{A}' \in \mathcal{A}$ to use more effective augmentations only. Detailed settings and methods are explained in Appendix F.

4. Experiments

In this section, we aim to validate the superiority of our method through both qualitative and quantitative analyses. The experimental and implementation details are in Appendix A, and the ablation on different decoding strategies is in Appendix E.

4.1. Experimental Settings

Datasets and evaluation metrics. We conduct experiments using three datasets: MME (Fu et al., 2024), MMBench (Liu et al., 2024b), and VQAv2 (Goyal et al., 2017). Each dataset consists of image-question pairs to evaluate how well LVLM generates robust and correct answers to various questions. The details of the datasets can be found in Section A.2.

Models. We evaluate the performance of VACoDe on three pretrained baseline LVLM foundation models: LLaVA-1.5 (Liu et al., 2023b), InstructBLIP (Dai et al., 2024) and Qwen-VL (Bai et al., 2023). Specifically, we use pretrained LLaVA-1.5 and InstructBLIP with Vicuna (Chiang et al., 2023) 13B language decoder, and Qwen-VL with Qwen 7B backbone. Ablation studies on model size can be found in Appendix D.

Algorithm 1 `VACoDe`: Visual-Augmented Contrastive Decoding

```
Input: Image and question pair $(v, q)$, target sequence length $T$, Augmentation set $\mathcal{A}$, # of VA $N$, Distance function $D(\cdot)$, Amplification coefficient $\alpha$, plausibility constraint parameter $\beta$
for $t = 1...T$ do
    if $T = 1$ then                                                      ▷ Determine contrast augmentation for the entire decoding process
        $z_t \leftarrow f(y_t|v, q, y_{<t})$ and $\tilde{z}_{t,i} \leftarrow f(y_t|\mathcal{O}_o(v), q, y_{<t}), \quad \forall o \in \mathcal{A}$      ▷ Generate logits
        $p_t \leftarrow \texttt{SOFTMAX}(z_t)$ and $\tilde{p}_{t,i} \leftarrow \texttt{SOFTMAX}(\tilde{z}_{t,i}), \quad \forall o \in \mathcal{A}$      ▷ Compute probability
        $\hat{o} \leftarrow \arg\max_{o \in \mathcal{A}}(D(p_t, \tilde{p}_{t,i}))$      ▷ Select the most constrastive augmentation
    else
        $z_t \leftarrow f(y_t|v, q, y_{<t})$ and $\tilde{z}_{t,\hat{o}} \leftarrow f(y_t|\mathcal{O}_{\hat{o}}(v), q, y_{<t})$      ▷ Generate logits
        $p_t \leftarrow \texttt{SOFTMAX}(z_t)$ and $\tilde{p}_{t,\hat{o}} \leftarrow \texttt{SOFTMAX}(\tilde{z}_{t,\hat{o}})$      ▷ Compute probability
    end if
    $p_{\texttt{VACoDe},t} = (1 + \alpha) \cdot p_t - \alpha \cdot \tilde{p}_{t,\hat{o}}$      ▷ Compute VACoDe probability
    $V_{\text{cand}}(y_{<t}) \leftarrow \{y_t \in V : p_t(y_t|v, q, y_{<t}) \geq \beta \max_w p_t(w|v, q, y_{<t})\}$      ▷ Candidate Set
    $p_{\texttt{VACoDe},t}(y) = 0$, if $y \notin V_{\text{cand}}(y_{<t})$      ▷ Discard not-candidate words
    $y_t = \texttt{SAMPLING}_y(p_{\texttt{VACoDe},t})$      ▷ Sampling next word
end for
```

Table 2. `MME` performance on perception task by using LLaVA-1.5 13B. The best and second-best performances are reported using **bold** and underline formatting, respectively.

Method	Aug.	existence	count	position	color	posters	celebrity	scene	landmark	artwork	OCR	Total
Regular	-	182.00	125.33	110.33	154.67	128.57	123.00	153.05	131.30	108.30	111.00	$1327.55_{\pm 16.2}$
VCD	noise	185.00	122.33	125.00	151.67	137.62	133.12	151.15	139.10	110.85	98.50	$1354.34_{\pm 24.5}$
Single	color	182.00	134.00	129.33	160.00	142.86	142.24	154.60	143.40	112.60	113.50	$1414.53_{\pm 9.56}$
	edge	185.00	146.00	125.00	157.67	141.70	142.24	152.95	139.50	113.15	121.00	$1424.20_{\pm 22.0}$
	sharp	182.00	113.33	130.00	156.33	136.46	130.76	156.90	137.10	109.85	109.00	$1361.74_{\pm 20.3}$
	crop	187.00	110.33	138.33	147.67	149.80	146.65	156.70	146.65	105.75	103.50	$1392.38_{\pm 24.7}$
	erase	185.00	126.67	116.33	144.67	147.55	128.29	156.60	132.85	110.95	117.00	$1365.91_{\pm 22.5}$
	flip	183.00	122.00	129.00	155.00	143.61	132.12	151.45	133.90	109.55	115.00	$1374.62_{\pm 14.9}$
`VACoDe`	*all*	184.00	138.67	134.00	167.00	146.80	144.29	149.35	145.30	114.65	119.00	$\underline{1443.06}_{\pm 6.80}$
	selection	183.00	140.33	132.00	165.33	146.46	143.71	149.80	145.05	114.45	123.00	$\mathbf{1443.14}_{\pm 9.99}$

Augmentations. We use 7 augmentations in Figure 2. Each single augmentation is used as a baseline and note that a single noise addition augmentation is equivalent to the VCD (Leng et al., 2023) method. When applying `VACoDe`, we employ both *all* and *selection* strategies. The selected augmentations vary depending on the models or datasets. For example, on the `MME` benchmark, the LLaVA-13B model utilizes four specific augmentations: color, edge, crop, and flip for *selection*.

4.2. Experiment Results

In this section, we analyze the main result. More discussions on our study can be found in Appendix B.

Result on each category. Table 2 shows the MME score of CD using different augmentations on the `MME` dataset for each perception category using the LLaVA-1.5 13B model. As mentioned in Section 3.1, if each single visual augmentation corresponds to a contrastive augmentation on the given question, it improves the CD performance. Although it does not improve the performance in all other question categories, the total MME score increases. This means that LVLMs are likely to provide incorrect answers when image augmentations are applied. For instance, in the case of questions about recognizing celebrities or landmarks, humans can answer the corresponding labels even if the color information is distorted. However, in the case of LVLMs, when a color-distorted visual image is given, the LVLMs fail to perceive the object, and the contrast increases significantly. Through these observations, we can indirectly figure out some impacts of augmentations on the LVLMs.

Using `VACoDe` results in better performance compared to using a single augmentation. As shown in Figure 4 and Table 2, when using a single augmentation for contrastive decoding of LVLMs, it is challenging to gain distinguished performance across all types of questions. However, `VACoDe` automatically selects the candidate expected to have high contrast based on the given task among the candidate augmentations and uses it for CD. Selecting an appropriate visual augmentation based on a given question and image shows outstanding performance improvement across overall question categories compared to using single visual augmentations.

Table 3. MME, VQAv2, and MMBench performance on different LVLMs. LV, QV, and IB denote the LLaVA-1.5 13B, Qwen-VL 7B, and InstructBLIP 13B, respectively.

Method	Aug.	MME			VQAv2			MMBench		
		LV	QV	IB	LV	QV	IB	LV	QV	IB
Regular	-	1327.55	1355.32	1151.45	67.54	75.38	61.82	73.74	64.49	43.75
VCD	noise	1354.34	1406.15	1208.44	71.29	75.54	66.64	74.55	68.53	48.80
Single	color	1414.53	1422.69	1237.71	71.94	76.26	67.26	75.42	68.95	48.06
	edge	1424.20	1393.32	1220.63	71.88	75.92	67.51	74.77	69.07	49.76
	sharp	1361.74	1395.14	1164.32	71.35	76.18	66.45	74.69	68.17	47.12
	crop	1392.38	1396.83	1205.55	71.22	76.06	66.03	74.67	68.33	47.59
	erase	1365.91	1385.33	1185.32	71.66	76.10	66.64	74.86	68.36	47.17
	flip	1374.62	1425.20	1213.61	71.76	75.54	66.81	75.34	69.38	48.69
VACoDe	*all*	1443.06	1406.05	1248.30	**72.53**	76.06	67.97	75.49	69.89	50.49
	selection	**1443.14**	**1426.43**	**1256.09**	72.46	**76.29**	**67.99**	**75.57**	**70.01**	**50.67**

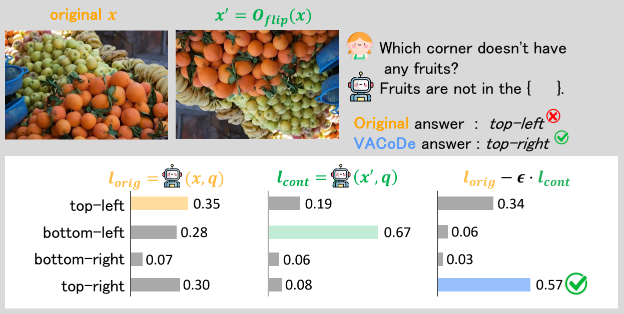


Figure 7. The example of MMBench shows how LLaVA-13B utilizes VACoDe to correct the answer.

Results across more datasets and models. Table 3 presents the results for the MME, VQAv2, and MMBench datasets using LLaVA-1.5 13B, Qwen-VL 7B, and InstructBLIP 13B models. Performance on MME is measured by the MME score, while accuracy is used for the other benchmarks. The performance of the MMBench dataset is evaluated using the CircularEval strategy. A notable observation is that VACoDe shows a significant performance improvement in each setting, regardless of the dataset or model used. This indicates the robustness of VACoDe in improving the accuracy and reliability of LVLM outputs.

Moreover, *selection* shows better performance than *all* in most experiments. This indicates that our approach to eliminating noisy augmentations is effective and highlights the importance of using only the most effective augmentations to achieve better performance. This approach not only proves its efficacy but also provides users with guidance on choosing the optimal subset of augmentations from various options.

4.3. Case Study

In this section, we discuss examples of using VACoDe in MMBench with LLaVA-13B as illustrated in Figure 7. On the position-type question "which corner doesn't have any fruits?", the original prediction answers 'top-left', which is incorrect. After flipping the image, the empty space moves to the bottom-left, and the model correctly identifies it. It's important to note that 'top-left' has a high probability in both images, indicating that the LVLM may have a bias to assign high probability to 'top-left' given question. In this case, CD successfully eliminates this bias, resulting in the correct output.

5. Conclusion

In this paper, we introduce VACoDe for utilizing multiple augmentations by adaptively choosing contrastive decoding. Initially, we examined the effects of various augmentations and found that their effectiveness depends on the type of question. Specifically, each query has key features that act as clues for answers, and contrastive augmentations can modify these features. Therefore, selecting the contrastive augmentation that creates a significant contrast is essential for improving CD. Based on this, we propose an algorithm called VACoDe, which selects augmentation by the largest distance D. Experiments show that VACoDe outperforms other methods across different datasets and underscores the importance of selecting appropriate augmentations.

Limitation. Our method selects the appropriate contrastive augmentation among augmentation candidates. No matter how properly VACoDe works and the appropriate augmentation is selected for the given task, if there is no sufficient contrastive augmentation for the task among the candidates, it is difficult to expect a significant performance gain.

Future work. Future work includes implementing an automatic search for candidate augmentation sets suitable for the target task. Additionally, investigating the relationship between visual contrast and language contrast on LVLMs suggests a further direction for expanding this study.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST), 10%) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 90%).

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Zhang, H., and Gan, C. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023.

Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., and Zhou, J. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Kim, J. M., Koepke, A., Schmid, C., and Akata, Z. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2584–2594, 2023.

Kim, T., Kim, J., Lee, G., and Yun, S.-Y. Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions.

Kumar Singh, K. and Jae Lee, Y. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 3524–3533, 2017.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

Lin, W., Chen, J., Mei, J., Coca, A., and Byrne, B. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. Mmbench: Is your multi-modal model an all-around player?, 2024b.

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

Rose, D., Himakunthala, V., Ouyang, A., He, R., Mei, A., Lu, Y., Saxon, M., Sonar, C., Mirza, D., and Wang, W. Y. Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

Surís, D., Menon, S., and Vondrick, C. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

Wan, D., Cho, J., Stengel-Eskin, E., and Bansal, M. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024.

Wang, X., Pan, J., Ding, L., and Biemann, C. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yang, L., Wang, Y., Li, X., Wang, X., and Yang, J. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, L., Ji, D., Chen, T., Xu, P., Ye, J., and Liu, J. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.

-Supplementary Material-

VACoDe: Visual Augmented Contrastive Decoding

A. Experiment Details

A.1. Descriptions of the augmentations

In this paper, we employ the operations $\mathcal{A} = \{\texttt{color}, \texttt{flip}, \texttt{random crop}, \texttt{random erase}, \texttt{sharp}, \texttt{edge}, \texttt{noise}\}$. Examples are illustrated in Figure 2. The descriptions of the augmentations are: (1) color: color inversion, (2) flip: horizontal flip followed by vertical flip, (3) crop: cropping a random part of the image, (4) erase: randomly erasing part of the image, (5) sharp: adjusting image sharpness, (6) edge: extracting edge textures, and (7) noise: adding diffusion noise. Note that we use the default noise setting from VCD (Leng et al., 2023).

A.2. Details of datasets and evaluation metrics

- MME is a LVLM evaluation dataset with granular question categories, including 10 categories from the perception tasks and 4 from the cognition tasks. The labels consist of ‘Yes’ or ‘No,’ and performance is measured by MME score, which is derived from accuracy. In this paper, we evaluate the perception category as our method focuses on observation ability.

- MMBench is a dataset of image-question pairs from 20 categories to validate how skillfully LVLM performs on various vision-language tasks with option labels. For evaluation, we incorporate SingleEval and CircularEval. SingleEval provides a score based on fixed labels, while CircularEval rotates the positions of the possible option labels in a circular manner.

- VQAv2 is a dataset containing open-ended questions paired with images. This allows a proper evaluation of how expertly the model can utilize the given visual information rather than simply using the learned language priors. We randomly select $30,000$ samples from the VQAv2 evaluation dataset to validate our method.

For the reliability of the results, we report performance using the average of the results of 5 different seed runs for MME and MMBench, and a single run for VQAv2.

A.3. Experimental setting details for investigating the effect of VA on LVLM decoding

We hypothesize that there exists a specific set of VAs for each query, termed that contrastive augmentation set. This set includes VAs that cause the LVLM to produce incorrect answers. Essentially, these VAs alter the input image, causing it to lose key features necessary to answer the query correctly. For instance, if the query pertains to color, color-related augmentation, such as color inversion, can lead to an incorrect response. In this section, we describe the MME benchmark (Fu et al., 2024), which is primarily used to explore our hypothesis and present our findings based on that dataset.

Experimental setting. In this section, we provide a summary of the MME dataset (Fu et al., 2024) and the experimental settings in detail for investigating the effect of VA on LVLM decoding. MME benchmark categorizes question types into 14 groups, such as color, count, position, existence, and more. We concentrate on questions related to color, existence, and position to thoroughly investigate the influence of VAs. We manually select contrastive augmentation for each query type based on our hypothesis that they can produce incorrect outputs. Specifically, depending on the query type, we choose the contrastive augmentation pair as outlined in Table 1. Note that the remaining two augmentations for each are considered persistent augmentations. In subsequent experiments, we use these three augmentations and three query types to evaluate the augmentation effect by assessing softmax outputs.

A.4. Implementation details

For the main experiment, we choose $\alpha = 1$ and $\beta = 0.1$ for the VACoDe. Additionally, we use $T = 1$ and $p = 1$ for the sampling strategy, which employs the softmax distribution for the next token generation. In the absence of prior knowledge, applying all data augmentation operations can be beneficial. However, having access to the ground truth labels for a subset allows us to use this information to identify a more optimal subset of candidate augmentations. We refer to these two

Table 4. MME performance of VACoDe with different candidate combinations on LLaVA-1.5 13B. We evaluate the performance of the candidate set $\mathcal{A}$ by excluding each candidate one by one.

Method	Aug	existence	count	position	color	posters	celebrity	scene	landmark	artwork	OCR	Total
Regular	-	182.00	125.33	110.33	154.67	128.57	123.00	153.05	131.30	108.30	111.00	1327.55
Single	color	182.00	134.00	129.33	160.00	142.86	142.24	154.60	143.40	112.60	113.50	1414.53
	crop	187.00	110.33	138.33	147.67	149.80	146.65	156.70	146.65	105.75	103.50	1392.38
	flip	183.00	122.00	129.00	155.00	143.61	132.12	151.45	133.90	109.55	115.00	1374.62
VACoDe (subset)	color+crop	186.00	116.67	132.33	160.00	150.27	149.82	155.70	153.35	108.75	108.00	1420.90
	color+flip	181.00	138.33	136.33	161.67	145.10	141.41	150.10	141.00	113.55	108.00	1416.50
	crop+flip	184.00	116.00	133.33	150.67	148.57	147.94	155.55	151.50	107.80	103.50	1398.86
VACoDe	color+crop+flip	183.00	120.33	133.33	161.00	150.07	149.94	155.70	155.70	109.20	108.00	**1426.28**

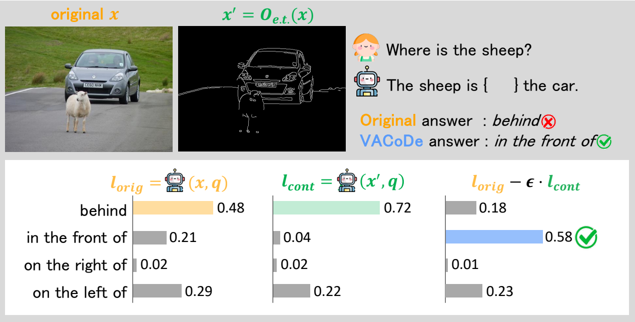


Figure 8. Another case study example of MMBench. VACoDe successfully corrects the answer. The edge augmentation is selected as the contrastive augmentation.

scenarios as *all*, which incorporates all available augmentations, and *selection*, which leverages a validation set to choose a subset of augmentations. The refinement strategy is elaborated upon in Appendix F.

A.5. Experiment computation resource

In this paper, all reported our experiment used LVLM models can run on a single 48 GB NVIDIA RTX A6000. In the process of applying VACoDe, our model requires inference as the number of VAs used in the first step only, and each subsequent generation step requires twice token generation stages.

B. Further Discussion

B.1. Analysis on the Combination of Visual Augmentations

In this section, we evaluate different combinations of augmentations to estimate the impact of each augmentation. For simplicity, we limit the augmentation set to $\{\texttt{color}, \texttt{flip}, \texttt{random crop}\}$. Table 4 shows the effect of using all augmentation candidates in the set and the impact of excluding each one individually. According to the results, VACoDe performance using all three augmentations, color, crop and flip, shows higher performance than other sub-combinations. Specifically, when color or flip augmentation is removed from the augmentation set, performance in the color and position categories significantly decreases. Considering each augmentation has a different contrastive effect, the results confirm that selecting an appropriate combination of VAs can provide proper contrast for a given task.

B.2. Qualitative Study

In this section, we discuss another example of using VACoDe in MMBench with LLAVA-13B as illustrated in Figure 8. The example demonstrates an instance where LVLM incorrectly predicts as “sheep is behind the car.” When edge augmentation is applied, it exacerbates LVLM’s confusion, increasing the likelihood of an incorrect answer. However, CD corrects this by addressing the disadvantage on it and generates the correct answer.

C. Ablation on Distance metric D

In this section, we examine comprehensive several additional ablation experiments that are considerable in the environment in which VACoDe is applied. Based on these ablation results, we expect VACoDe to have universally high robustness and be able to perform various tasks, models, and inferences.

We perform experiments using several common distance measures to define our distance function D that VACoDe uses to select which VA will produce high contrast. The experiment is performed in the MME dataset using the LLaVA-1.5 13B model. Also, we use the average softmax Gain directly to check the effect. In detail, softmax Gain on the correct answer label obtained when applying the distance measure candidate D_i used in the experiment and the VAs used in Figure 2 to VACoDe for all samples. In order to control the variables of VAs that contain randomness, each experiment performs a total of 5 experiments with different seeds on the entire MME dataset and then measures softmax Gain through the average.

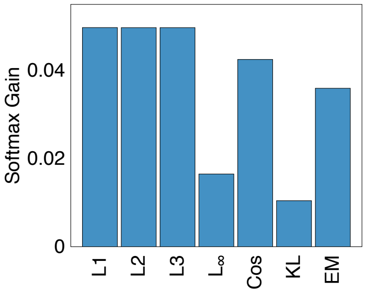


Figure 9. Average softmax gain by different distance metrics.

Figure 9 shows the result of affectness of different distance D functions. In this experiment, we use L_1, L_2, L_3, L_∞, Cosine similarity, Kullback-Leibler Divergence (KL divergence), and Earth mover's distance (EM distance) as distance candidates. The x-axis of the results in Figure 9 means the candidate distance names used, and the y-axis means the average softmax gain improved compared to regular decoding obtained through VACoDe when each distance is used as a measurement. From the results, we can check that L_1, L_2, and L_3 norms show high performance improvement almost no difference overall. This means that any of these can be used in the algorithm as a distance function at a similar level. However, in the case of L_∞ and KL divergence, it can be seen that the actual performance improvement is much smaller compared to others. These show very low-performance improvement compared to the L_2 distance, which we used in the main experiment, meaning they are improper measurements for estimating the expected contrast of VAs. The other two distances, cosine similarity and EM distance, performed higher than KL divergence but did not perform higher than L_2 norm for the entire MME dataset. Based on this result, we empirically confirmed that using L_2 norm as our main VACoDe distance D is a meaningful standard through experiments with these distance measures and the results shown throughout our main experiments.

D. Analysis of Different Model Sizes

We showed that VACoDe is proper for general LVLMs and has a significant effect on performance by experimenting with three different models LLaVA-1.5, InstructBLIP, and Qwen-VL on various types of datasets at the Section 4. In this ablation, we conduct an experiment using LLaVA-1.5 7B, 13B and InstructBLIP 7B, 13B to check the effect of the model size on VACoDe. MME dataset is used for this experiment. We measured the performance for the perception category and the total performance for each model.

Table 5 shows the performance of VACoDe on each model and size for the MME dataset. From the result, we can confirm even if the model size and model used are different, the softmax gain obtained when each VA is used in VACoDe is robust to the type and size of the model and shows a tendency to be dependent on the given task. Throughout the experimental results, the single VA edge and color show very high performance. On the other hand, we can see that single VA sharp and erase have an overall low-performance gain. For different models, the performance gain shown by each VA shows an overall similar trend, and it can be seen that there is a higher performance improvement compared to the original regular decoding.

Furthermore, for different model sizes, we can see that there is a significant performance gain when applying our algorithm

Table 5. MME performance by different model sizes.

Method	Aug	LLaVA-1.5 7B	LLaVA-1.5 13B	InstructBLIP 7B	InstructBLIP 13B
Regular	-	1272.22	1327.55	1155.26	1151.45
VCD	noise mask	1323.44	1354.34	1218.90	1208.44
Single	color	1347.24	1414.53	1224.26	1237.71
	edgetexture	1350.68	1424.20	1221.15	1220.63
	sharpness	1323.60	1361.74	1177.84	1164.32
	randcrop	1338.50	1384.65	1194.13	1205.55
	randerase	1310.89	1365.91	1195.27	1185.32
	flip	1344.75	1374.62	1222.34	1213.61
VACoDe	all	**1368.89**	1443.06	1249.56	1248.30
	selection	1364.36	**1443.14**	**1254.16**	**1256.09**

Table 6. MME Performance by different sampling strategies.

Method	Aug	Top P	Top K	Temperature	
		$p = 0.9$	$k = 50, T = 0.7$	$T = 0.7$	$T = 1.5$
Regular	-	1352.87	1399.33	1403.99	1169.71
VCD	noise mask	1370.47	1425.60	1429.52	1316.95
Single	color	1405.90	1443.27	1445.19	1349.20
	edge	1434.14	1433.86	1420.64	1364.72
	sharp	1381.00	1415.88	1416.63	1294.08
	crop	1391.01	1413.09	1422.15	1342.43
	erase	1374.47	1404.27	1399.67	1315.08
	flip	1404.76	1426.97	1425.54	1340.88
VACoDe	all	**1462.67**	1456.03	1454.32	**1389.03**
	selection	1462.58	**1457.10**	**1458.73**	1377.47

VACoDe. VACoDe using all of the VAs specified in Figure 2 shows a higher performance improvement than using each single VA. This indicates that, regardless of model and size, each application has the highest performance in the entire perception category and total performance.

E. Effect of Different Sampling Strategies

We perform analysis studies on different sampling strategies to see how VACoDe is affected by sampling methods other than basic regular decoding. In this experiment, 4 sampling techniques are applied: (1) Top P sampling (specifically, $p = 0.9$), (2) Top K sampling (specifically, $k = 50$), (3) Temperature sampling (specifically, $T = 0.7/1.5$). Top P sampling is a method in which the only token candidates in the distribution on cumulative probability p can be selected as the next token. This has the effect of preventing noise samples with too low a probability to be extracted from candidates. Top K sampling uses only the top k candidates from the highest probability for sampling. In temperature sampling, temperature scaling is applied to the softmax to calculate the next token logits. When temperature T is low, the possibility of selecting a high-probability candidate group increases, and the possibility of choosing low-probability candidates decreases. It has the effect of increasing the probability of more static responses. Conversely, when the temperature T is large, the chance of choosing among the high-probability candidates decreases, and the low-probability candidates increases. It has the effect of increasing the possibility of making more diverse responses.

Table 6 show the experiment result of VACoDe with different sampling strategies. From the table, we can check that VACoDe gives us a high performance in various types of samplings. This is not only for regular decoding, but it also shows higher performance compared to single VA in the Top P sampling and Top K sampling. A notable observation is that VACoDe shows high performance in both cases where the temperature scale gets higher or lower. In the case of high temperature, the model has a higher probability of generation more diverse, and the explanations and representations are getting richer. However, in this case, there is a potential problem that the entire output is inaccurate while in generation. In particular, if specific information for a given image must be utilized rather than using inherent prior knowledge, however, there is a possibility that incorrect output may lose correlation with visual information on LVLMs. Our results show that using VACoDe in this situation can be expected to have the effect of concentrating the model to intentionally utilize visual information by contrastive decoding the output through contrast VA. As can be seen from the results, in situations where the temperature scale is large, CD through VA produces a more significant performance gain. Additionally, the magnitude of contrastiveness produced by each VA is different in the task so that we can see a considerable performance difference between single VA CDs. In this situation, `VACoDe`, which automatically selects and applies the appropriate VA for a given task, can be used more appropriately and robustly to the given scenario. Furthermore, it shows that `VACoDe` has the highest performance improvement.

Our algorithm can be also used at the low temperature scale scinarios, which grows the sampling possibility of high probabiltiy token being chosen as next token. In this scenario, the original model's high logits become more extensive than usual by temperature scaling, increasing the probability of being selected as the next token. When the correct answer logit does not have a high value, the possibility of being selected as the next token is crucially dropped. For a low-temperature scale, once the model starts generation with an incorrect token, it is more likely to continue generating incorrect responses. As mentioned in CD, in the case of high confidence in high logit sampling methods in a generation, a wrong token selection can significantly impact the quality of future responses. In this situation, using `VACoDe` can increase the likelihood that a low correct answer token will be selected as the correct answer through CD using contrast VA. As a result, it shows high robustness against the temperature sampling scale and increases the likelihood of providing an appropriate response.

F. Selection Strategy

Removing noisy augmentations via acceptence threshold. Using the distance D, we expect to select a VA that shows high-performance improvement when used on CD. However, there may exist cases where some VAs cannot be appropriate contrastive augmentation for a specific task overall. In this case, these VAs contribute less to performance improvement than other VAs on average and can sometimes become noise that prevents other VAs from being used as contrast. We use the Acceptance Threshold, a simple baseline that eliminates the noise VAs. To discover the suitableness of VAs for the target task, in the sample sub-dataset, we utilize the LVLM's first token generation distance by `VACoDe` for each VA. Let c_i be the number of times that VA$_i$ selected as contrast VA among a total of M VAs. For the N data samples and acceptance threshold τ, candidate VAs with $c_i < \tau\frac{N}{M}$ are treated as unsuitable for this task and removed. Throughout the main experiments, we used the acceptance threshold of $\tau = 0.5$.

G. Related Works

Large vision language models (LVLMs). LVLMs are among the most prominent multi-modality models. They process pairs of input image v and text (*e.g.,* question) q, denoted as (v, q), and generate answers by utilizing the visual information within v. This paper primarily focuses on generative LVLMs, producing words one at a time in a sequence similar to LLMs. In this paper, we primarily focus on generative LVLMs, (rather than CLIP-like (Radford et al., 2021) models), and similar to LLMs, LVLMs produce words in an autoregressive manner. The mathematical expression for this process is:

$$y_t \sim p(y_t|v, q, y_{<t}).$$

Here, $p(\cdot)$ represents the softmax of the output of the vocabulary set, and $y_{<t}$ denotes the words generated up to but not including the timestamp t. Like LLMs, note that LVLMs are also prone to hallucination (Li et al., 2023c; Liu et al., 2023a; Tong et al., 2024).

Visual augmentation. In the computer vision domain, visual augmentation has been studied to give variance to the image, thereby multiplying image information to avoid overfitting and ensuring stable training of the model. Traditional augmentations include changes in color, cropping, and flipping. Additionally, there are more advanced techniques of erasing (Kumar Singh & Jae Lee, 2017; DeVries & Taylor, 2017; Zhong et al., 2020), and other techniques such as mixup (Zhang et al., 2017) and CutMix (Yun et al., 2019). Furthermore, the automatic application of multiple augmentations has been explored (Cubuk et al., 2019; Lim et al., 2019).

Some studies in LVLMs employ VA to achieve the desired output in various methods. FGVP (Yang et al., 2024) adds blur to the background of the image, leaving the main object clear to emphasize it. To focus on each object in the image, (Chen et al., 2023; Surís et al., 2023; Lin et al., 2024) use multiple cropped images, each focusing on a single object to generate the desired output, while (Kim et al., 2023) uses inpating to erase objects to measure the correlation between objects.

Contrastive decoding. CD (Li et al., 2023b) was introduced in the NLP domain using two differently sized language models. It leverages contrastive output by subtracting the small model's probability from the larger model's to retain the strengths of the large model whilie eliminating the weaknesses that are evident in the small model. There are variants like DOLA (Chuang et al., 2023) which utilizes contrast in layer-level outputs and Instructive Decoding (Kim et al.) uses two contrastive instructions to generate an output opposite to the original output.

Recently, similar approaches have been applied in LVLMs, utilizing contrastive inputs to guide the model in generating accurate text, mainly focusing on reducing hallucination in LVLM (Li et al., 2023c; Liu et al., 2023a; Tong et al., 2024). VCD (Leng et al., 2023) demonstrates that adding noise to the image can elevate the hallucination inherent in LVLMs, subsequently applying CD to manage the hallucination. Another work CRG (Wan et al., 2024) employs a black bounding box to conceal the object relevant to the question, amplifying hallucination, while HALC (Chen et al., 2024) uses multiple different cropped images and explores multiple pairs of cropped images to find pairs that amplify the information in the cropped image. These works address methods to manage the hallucination in LVLMs using a single type of augmentation, which has limitations in generating enough contrast for various types of questions. There are other works that do not use additional image inputs. IBD (Zhu et al., 2024) fine-tunes an additional image-biased model to mitigate the text bias of VLM, and ICD (Wang et al., 2024) introduces using opposing instructions to generate incorrect output as ID. Unlike previous studies, VACoDe explores multiple augmentations and selects the most effective one to answer the question. Moreover, it does not require additional training or an external model, providing direct perturbation to the image.