

# Attention-space Contrastive Guidance for Efficient Hallucination Mitigation in LVLMs

## Supplementary Material

### A. Additional Experimental Details

#### A.1. Models.

We evaluate ACG on three open-source large vision–language models (LVLMs) with diverse language backbones and vision–language connectors: LLaVA-1.5, MiniGPT-4, and Qwen-VL-Chat.

**LLaVA-1.5.** [5] LLaVA-1.5 adopts a CLIP ViT-L/336px vision encoder and a Vicuna language model built on the LLaMA architecture, connected by a fully connected MLP-based vision–language projector. The CLIP encoder produces 576 visual tokens per image, which are mapped into the language model token embedding space by a two-layer MLP and then concatenated with text tokens. The model is trained in a two-stage pipeline: vision–language alignment on image–text pairs, followed by visual instruction tuning on conversational multimodal data.

**MiniGPT-4.** [9] MiniGPT-4 uses the visual frontend of BLIP-2: a ViT-G/14 visual encoder from EVA-CLIP followed by a Q-Former that compresses dense image features into a small set of visual tokens. The Q-Former employs a fixed set of learnable queries (32 in our setup), so each image is represented as 32 visual tokens. These Q-Former outputs are passed through a single linear projection layer to align them with the Vicuna language model embedding space, and the projected visual tokens are then fed into Vicuna as a soft prompt for generation.

**Qwen-VL-Chat.** [1] Qwen-VL-Chat builds on the Qwen language model and a ViT-bigG visual encoder from OpenCLIP. Image features are first extracted by the ViT encoder and then compressed by a position-aware vision–language adapter: a single-layer cross-attention module with learnable query embeddings. We use the default configuration with 256 queries, yielding 256 visual tokens per image, which are fed into the language model as a fixed-length visual token block.

#### A.2. Benchmarks.

**POPE.** [3] POPE (Precision-based Object Probing Evaluation) evaluates object hallucination through binary object-presence queries. For each image, the model answers questions of the form “Is there a <object> in the image?” with a balanced mixture of present and absent objects. POPE provides three complementary evalu-

ation sets: (1) *Random*: object categories are sampled uniformly from the vocabulary, reflecting unbiased hallucination performance; (2) *Popular*: focuses on frequently occurring objects in large-scale training corpora, testing whether the model over-relies on language priors; (3) *Adversarial*: selects semantically or visually confusable objects (e.g., querying “cat” for dog images), stressing context-induced hallucination.

**Metrics.** POPE reports Accuracy, Precision, Recall, and F1.

**CHAIR.** [7] CHAIR (Caption Hallucination Assessment with Image Re-annotation) measures hallucination in image captioning by aligning object mentions in generated captions with COCO ground-truth annotations. Any object mentioned in the caption but absent in the image is regarded as hallucinated.

**Metrics.** CHAIR provides two metrics:

$$\text{CHAIR}_i = \frac{\#\text{hallucinated object instances}}{\#\text{all mentioned objects}},$$
$$\text{CHAIR}_s = \frac{\#\text{captions containing hallucination}}{\#\text{total captions}}.$$

$\text{CHAIR}_i$  captures object-level hallucination frequency, while  $\text{CHAIR}_s$  measures how often a caption contains any hallucination.

**Why F1 is Reported Alongside  $\text{CHAIR}_s$  and  $\text{CHAIR}_i$ .**  $\text{CHAIR}_i$  and  $\text{CHAIR}_s$  measure hallucination from the perspective of “how often” hallucinated objects appear in a caption, but they do not consider whether the model successfully mentions objects that actually exist in the image. In contrast, the object-level F1 score captures the balance between avoiding hallucinated objects (precision) and correctly mentioning ground-truth objects (recall). A model may achieve a low CHAIR score simply by producing overly conservative captions that omit many valid objects, which results in low recall. Therefore, reporting F1 alongside  $\text{CHAIR}_i$  and  $\text{CHAIR}_s$  provides a more complete view of caption quality, distinguishing models that truly reduce hallucination from those that merely under-describe the image.

**MMHal-Bench.** [8] MMHal-Bench is a hallucination-centric benchmark specifically designed to diagnose the visual grounding reliability of large vision–language models (LVLMs). The dataset contains images paired with care-

fully constructed natural-language queries that target reasoning types known to induce hallucination, such as object attributes, spatial relations, counting, and adversarially misleading premises. Unlike captioning-based or binary object-probing benchmarks, MMHal-Bench evaluates open-ended, reasoning-intensive responses, where hallucinations arise not only in object mentions but also in relational, numerical, or contextual inferences. Each query expects a concise, grounded answer that can be automatically judged for hallucination and informativeness using an LLM-as-a-judge protocol.

The benchmark evaluates hallucination across several reasoning dimensions that are known to induce failure in LVLMs:

- **ATTR (Object Attributes):** questions about appearance attributes such as color, shape, texture, or material (e.g., “What color is the man’s jacket?”).
- **ADV (Adversarial Objects):** queries intentionally designed to include objects not present in the image (e.g., “What is the dog holding in its hand?”). This category tests the model’s robustness to adversarial wording and its ability to reject false premises.
- **COMP (Comparisons):** relational comparisons of size, number, or attributes between two or more objects (e.g., “Which cup is larger?”).
- **COUNT (Counting):** numerical reasoning about the number of instances.
- **SPAT (Spatial Relations):** reasoning about object positions or geometric relations (e.g., “Where is the bicycle relative to the car?”).
- **ENV (Environmental / Scene Inference):** global contextual reasoning about the environment, scene type, or high-level situational cues (e.g., “Is this an indoor or outdoor scene?”).

**Evaluation Protocol.** MMHal adopts an LLM-as-a-judge evaluation pipeline in which GPT-4 scores each model response along two dimensions: *informativeness* and *hallucination*. For every image–query pair, GPT-4 is prompted to evaluate (1) how informative the response is on a 0–6 scale, where 0 indicates an unhelpful or irrelevant answer and 6 indicates a fully grounded, complete, and contextually appropriate answer; and (2) whether the response contains any hallucinated visual content (binary judgment: hallucinated / not hallucinated). This protocol enables the benchmark to assess both the usefulness and the visual faithfulness of a model’s answer.

### A.3. Baseline Methods and Hyperparameters

To demonstrate our method’s generability among tasks, we set  $\gamma$  to 2.4 for LLaVA-1.5 [5], 0.3 for MiniGPT-4 [9], 1.4 for Qwen-VL [1] for all experiments unless explicitly stated otherwise. For a consistent and comparative analysis, we set greedy decoding as default.

Table 1. **ACG quantitative validation.** (LLaVA-1.5, POPE random split)

(a) **Text-only path approximation quality.** We compare  $\tilde{U}$  to the true text-only trajectory  $U$  using attention distributions (Attn) and logits.

	Attn cos $\uparrow$	Attn KL $\downarrow$	Logits cos $\uparrow$	Top-10 overlap $\uparrow$
$\tilde{U}$ vs. $U$	0.935	0.195	0.909	0.705

(b) **Orthogonalization effect.**  $z_C$  conditional logits,  $z_U$  true text-only logits, and  $z_{\tilde{U}}$  logits from the single-pass masked-and-propagated trajectory.

	$\Delta_{\text{mask}} (= z_C - z_{\tilde{U}})$	ACG no-orth	ACG + orth
<b>Mechanism (attention-output level, <math>\Delta O</math>)</b>			
proj_ratio( $\Delta O \rightarrow \text{text}$ ) $\downarrow$	–	0.407	$2 \times 10^{-4}$
<b>Behavior (logits level, <math>\Delta z</math>)</b>			
cos( $\Delta z, \Delta_{\text{true}}$ ) $\uparrow$	0.649	0.469	0.592
proj_ratio( $\Delta z \rightarrow z_U$ ) $\downarrow$	0.591	0.790	0.577

For baseline comparisons, we evaluated VCD [2], PAI [6], and VISTA [4]. All baselines were reproduced using their official code repositories, and all experiments were conducted under a unified greedy decoding setting. For hyperparameters, we followed the configurations reported in the original papers unless otherwise noted. An exception is VISTA: as prior reports indicate that the official hyperparameters do not reproduce the reported results, we conducted a parameter search and set `vsv-lambda` to 0.01 for POPE and 0.15 for CHAIR.

## B. Quantitative Validation of ACG

### B.1. Justifying the Masked Unconditional Path.

We quantitatively validate whether our **single-pass** masked text-only path matches a true image-absent forward pass. On 500 samples from a random POPE split with LLaVA-1.5, We form  $\tilde{U}$  by running one image-conditioned forward pass, masking *image keys* for the *last generated token* at each layer, and propagating the masked outputs across layers. We compute the true text-only trajectory  $U$  via a separate forward pass without images. Using attention distributions and logits,  $\tilde{U}$  closely tracks  $U$  throughout the network (Table 1a), suggesting that any residual leakage/redistribution remains bounded and does not induce a divergent text-only path.

### B.2. Effect of textual orthogonalization.

Textual Orthogonalization removes the text-direction component of the layerwise steering signal, and we verify this mechanism yields improved steering behavior at the logits level (Table 1b). At the attention-output level, orthogonalization effectively eliminates the text-direction projection. At the logits level, it increases alignment to the true steering direction and reduces text-only leakage measured by  $\text{proj\_ratio}(\Delta z \rightarrow z_U)$ . Overall, orthogonalization acts as

Table 2. **Effect of guidance scale  $\gamma$  on CHAIR (max 128) for LLaVA-1.5.** We report sentence-level hallucination (CHAIR<sub>s</sub>), instance-level hallucination (CHAIR<sub>i</sub>), F1, and average caption length (Len). We choose  $\gamma = 2.4$  as our operating point.

$\gamma$	CHAIR <sub>s</sub> (↓)	CHAIR <sub>i</sub> (↓)	F1 (↑)	Len
1.0	47.4	12.8	77.8	91.6
1.3	48.6	12.9	78.1	89.4
1.5	44.8	11.6	78.1	88.2
1.7	42.4	10.3	79.1	84.9
1.9	36.4	8.6	79.3	83.5
2.0	33.0	8.0	79.0	82.1
2.1	34.2	7.6	77.6	80.8
2.2	27.6	6.4	76.7	77.5
2.3	27.8	6.6	75.5	74.9
<b>2.4</b>	<b>21.0</b>	<b>4.8</b>	<b>74.4</b>	72.4
2.5	19.2	4.8	72.4	67.2
2.6	15.4	4.8	68.0	58.7
2.7	12.8	4.9	64.7	51.0
2.8	9.0	4.8	60.4	42.0
2.9	7.2	6.0	56.1	33.9
3.0	6.2	4.2	51.8	25.8

Table 3. **Effect of guidance scale  $\gamma$  on CHAIR (max 128) for MiniGPT-4.** We again report CHAIR<sub>s</sub>, CHAIR<sub>i</sub>, F1, and average caption length (Len). We choose  $\gamma = 0.3$  as our operating point.

$\gamma$	CHAIR <sub>s</sub> (↓)	CHAIR <sub>i</sub> (↓)	F1 (↑)	Len
0.10	27.6	9.0	70.6	72.9
0.15	24.4	7.9	71.0	69.4
0.20	21.0	6.8	70.3	63.5
0.25	16.6	5.2	69.5	59.1
<b>0.30</b>	<b>10.8</b>	<b>3.3</b>	<b>68.0</b>	66.3
0.35	6.2	2.5	63.2	73.6
0.40	2.6	1.6	54.3	94.9

a targeted correction of text-bias in the steering signal that empirically restores the desired steering behavior.

## C. Behavior Across Guidance Scale and Depth

### C.1. Guidance Scale Selection.

For each LVLM, we select the guidance scale  $\gamma$  on the CHAIR (max 128 tokens) benchmark by sweeping  $\gamma$  and monitoring the trade-off between hallucination and caption quality. Concretely, we measure sentence-level hallucination (CHAIR<sub>s</sub>), instance-level hallucination (CHAIR<sub>i</sub>), F1, and the average caption length (Len) under greedy decoding, and choose an operating point that (i) substantially reduces CHAIR<sub>i</sub> compared to the greedy baseline, while (ii) keeping F1 within roughly 5% of the baseline and (iii) avoiding degenerate overly short captions. The selected  $\gamma$

is then reused for all CHAIR(both max tokens 64 and 128) and POPE experiments on the corresponding model.

This protocol is consistent with prior training-free hallucination mitigation methods. For instance, PAI [6] tunes its scaling parameters by sweeping on the CHAIR benchmark itself, jointly considering CHAIR and F1, without introducing a separate validation split.

Table 2 shows the sweep for LLaVA-1.5. As  $\gamma$  increases from 1.0 to 2.4, CHAIR<sub>i</sub> consistently decreases from 12.8 to 4.8, while F1 only drops from 77.8 to 74.4 and the average length remains moderate (72.4 tokens). Beyond  $\gamma = 2.4$ , CHAIR<sub>i</sub> continues to decrease but F1 and length collapse sharply (e.g., F1 = 51.8 and Len = 25.8 at  $\gamma = 3.0$ ), indicating an over-aggressive regime. We therefore choose  $\gamma = 2.4$  as the canonical operating point for LLaVA-1.5. A similar trend is observed for MiniGPT-4 in Table 3. We select  $\gamma = 0.3$ , while avoiding the more unstable behavior at larger  $\gamma$ .

### C.2. Layer-block configurations.

We also study where to apply ACG inside the LLaVA-1.5 decoder by restricting the guidance to different layer blocks. Table 4 reports CHAIR<sub>s</sub>, CHAIR<sub>i</sub>, F1, and average caption length (Len) on CHAIR (max 128) when we apply ACG only to early (layers 1–8), early–mid (9–16), mid–late (17–24), or late (25–32) blocks. We denote the early-only variant (layers 1–8) as *ACG-Fast*, which offers a good trade-off between hallucination reduction and efficiency.

**Interpretation.** Early-layer guidance is notably more efficient than mid-to-late guidance: it achieves strong hallucination reduction with smaller  $\gamma$ , whereas deeper blocks require substantially larger coefficients for comparable effects. This trend is consistent with the view that early layers are a favorable intervention point for lightweight guidance, motivating our ACG-Fast variant.

## D. Limitations and Future Work

We highlight two limitations of our method and several directions for future work.

**Architecture-specific assumptions.** ACG assumes a standard LVLM design in which visual tokens appear as a contiguous block in the decoder input. However, architectures such as InstructBLIP employ Q-Former-based encoders or cross-attention modules that interleave visual information more tightly with the text stream. In such settings, simply masking visual-key positions may not fully remove visual information, because contextualized text embeddings can already contain fused vision features. Future work could develop architecture-aware masking strategies or alternative constructions of the text-only path that better match each model’s multimodal fusion mechanism.

**Depth-dependent sensitivity.** ACG also exhibits noticeable depth-dependent behavior: early layers respond

Table 4. **Layer-block ablation on LLaVA-1.5 (CHAIR, max 128)**. We apply ACG only to a given layer block and report CHAIR<sub>s</sub>, CHAIR<sub>i</sub>, F1, and average caption length (Len).

(a) Early (layers 1–8, ACG-Fast)					(b) Early–mid (layers 9–16)				
$\gamma$	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Len	$\gamma$	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Len
0.5	47.8	12.6	77.1	93.4	0.5	49.2	14.0	76.6	94.3
1.0	46.0	13.1	77.0	91.9	1.0	49.0	14.2	76.4	93.7
1.5	43.2	12.2	77.6	88.3	2.0	53.6	14.4	75.9	92.4
2.0	35.8	10.2	77.4	82.2	2.5	51.0	13.9	76.3	92.0
2.5	28.0	7.1	77.5	77.6	3.0	53.6	15.3	74.8	90.7
3.0	19.0	5.3	69.8	68.0	6.0	38.0	11.1	74.8	79.2
6.0	0.2	2.9	13.2	15.6	8.0	14.2	5.7	56.6	45.9
10.0	0.0	0.0	0.2	1.2	10.0	2.2	2.1	22.5	16.3

(c) Mid–late (layers 17–24)					(d) Late (layers 25–32)				
$\gamma$	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Len	$\gamma$	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Len
0.5	46.4	12.6	77.6	94.4	0.5	44.0	12.8	76.9	93.5
1.0	46.4	12.5	78.3	93.4	1.0	43.8	12.2	77.4	93.7
1.5	43.8	12.5	78.2	94.6	1.5	44.2	12.1	77.7	93.1
2.5	47.2	12.7	78.1	95.6	2.0	41.8	11.4	78.4	93.6
4.0	51.8	12.7	76.6	95.9	2.5	40.4	10.7	78.6	93.0
6.0	42.6	10.7	77.4	95.6	4.0	39.4	10.2	78.9	92.5
8.0	40.2	9.2	76.3	95.0	6.0	37.8	9.7	78.2	93.4
10.0	35.8	7.0	73.4	91.8	10.0	36.0	8.8	75.7	90.7

Table 5. **Category-wise informativeness and hallucination rate on MMHal.**

$\gamma$	Avg	ATTR	ADV	COMP	COUNT	SPAT	ENV	HOL	OTH	Hallucination Rate
0 (Vanilla)	1.94	2.25	1.42	2.67	1.50	1.92	2.92	1.75	1.08	0.59
1	2.06	2.25	1.33	2.17	1.33	2.75	3.25	2.08	1.33	0.56
1.5	2.07	2.50	1.08	1.75	1.92	2.00	3.25	2.08	2.00	0.57
2	2.12	2.50	1.67	1.92	1.58	2.25	3.25	2.00	1.83	0.53
2.4	2.01	2.58	2.00	1.83	1.17	2.25	3.25	1.17	1.83	0.56

strongly to guidance, whereas mid-to-late layers require significantly larger  $\gamma$  to achieve comparable hallucination reduction. This suggests that a uniform guidance scale may under-correct or over-correct depending on the layer. More adaptive designs—such as depth-specific scaling, head-wise weighting, or learned guidance schedules—may further stabilize ACG across layers and improve hallucination mitigation.

## E. Additional Experimental Results and Qualitative Examples

### E.1. Detailed MMHal-Bench Scores.

In addition to the MMHal-Bench scores reported in main paper, we also provide results across multiple values of  $\gamma$ , ranging from 1.0 to 2.4. As shown in Table 5, our method

consistently achieves higher scores and lower hallucination rates than the vanilla model, achieving highest score at  $\gamma = 2$ .

### E.2. Additional examples on MMHal-Bench.

Figure 1 provides additional examples from MMHal-Bench. For the environmental reasoning query, vanilla LLaVA-1.5 describes the scene as **sunny** and well-lit, following language priors, whereas our method correctly answers that the weather appears **dark** because the cabin is dimly lit by indoor lights. For the counting-style question, vanilla misreads the fastest runner’s bib number as **1019**, but our method outputs the correct number **1097**, aligned with the ground-truth. These examples support our quantitative findings that ACG improves informativeness while reducing hallucination on MMHal-Bench.

Visual Input :



Query :  
What was the weather like when this photo was taken?

[Ground Truth Answer]

When this photo was taken, it was already night and the sky was dark.

[LLaVA-1.5]

The weather in the photo appears to be **sunny**, as the room is well-lit and the **sunlight** is shining through the windows.

[Ours]

The weather in the photo was **dark**, as the room is dimly lit, and the cabin is lit up by a light.

Visual Input :



Query :  
What number is the fastest girl wearing in the picture?\*

[Ground Truth Answer]

The fastest girl in the picture is wearing number 1097.

[LLaVA-1.5]

The fastest girl in the picture is wearing the number **1019**.

[Ours]

**1097**

Figure 1. **Success examples on MMHal-Bench.** In the first example, vanilla LLaVA-1.5 hallucinates a bright and **sunny** environment, whereas our method correctly infers that the cabin is **dark** and lit by artificial light. In the second example, vanilla misreads the runner’s bib number as **1019**, while our method outputs the correct number **1097**, matching the ground-truth answer.

### E.3. Qualitative CHAIR, VQA Examples.

Figure 2 shows CHAIR-style captioning examples. In the toaster image (top), vanilla LLaVA-1.5 and the PAI baseline hallucinate background objects such as a **sink**, **cup**, or even misclassify the appliance as a **toaster oven**, while our method only mentions the grounded **toaster** and its visible attributes. In the second example (bottom), both baselines repeatedly refer to a **table** and a **knife** that are not clearly present in the image. ACG instead concentrates on the truly visible entities—**glasses**, **paper**, and **scissors**—showing that attention-space guidance effectively suppresses spurious background objects while preserving the core scene seman-

tics. We also present qualitative results from MMMU and MathVista in Fig. 3 and Fig. 4.

### References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1, 2
- [2] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

Visual Input :



Query :  
Please help me describe the image in detail.

[LLaVA-1.5]

The image features a silver **toaster** sitting on a tiled countertop.(...) In the background, there is a **sink** visible, suggesting that the kitchen is well-equipped for cooking and cleaning tasks. Additionally, there is a **cup** placed on the countertop, possibly for holding a beverage or other items.

[PAI]

The image features a silver **toaster** oven sitting on a tiled countertop. The toaster **oven** is placed on a tile countertop, and it appears to be a Cuisin brand toaster oven. The toaster oven is turned on, and it is placed on a tile countertop.

[Ours]

The image features a **toaster** on a green tiled countertop. The toaster is a Charger toaster, and it is sitting on a tile counter. The toaster is silver and has a black cord plugged into it.

Visual Input :



Query :  
Please help me describe the image in detail.

[LLaVA-1.5]

The image features a man wearing **glasses**, sitting at a **table** and cutting a piece of **paper** with a pair of **scissors**. (...) The man is using a **knife** to cut the paper, which is placed on the table in front of him.

[PAI]

The image features a man wearing **glasses**, who is cutting a piece of **paper** with a pair of **scissors**. He is focused on the task at hand, and the paper is placed on a **table**. (...)The man is focused on his task, and the paper is placed on a table, which is located in the foreground of the image.

[Ours]

The man is wearing **glasses** and cutting a piece of **paper** with a pair of **scissors**. He is holding the paper with his fingers and is looking down at it. The man is wearing a white shirt and glasses.

Figure 2. **Qualitative CHAIR examples comparing vanilla LLaVA-1.5, PAI, and our ACG method.** For each caption, we highlight object tokens: **blue** tokens denote objects that are grounded in the image, while **red** tokens indicate hallucinated objects (e.g., **sink**, **cup**, **table**, **knife**). Compared to the baselines, ACG removes spurious background objects and focuses on the truly visible entities (e.g., the toaster, glasses, paper, and scissors).

pages 13872–13882, 2024. 2

- [3] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 1
- [4] Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 35799–35819. PMLR, 2025. 2
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 1, 2
- [6] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in llms, 2024. 2, 3
- [7] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045,

Year	Project A	Project B
0	-\$20,000	-\$24,000
1	13,200	14,100
2	8,300	9,800
3	3,200	7,600

**Q:** Maxwell Software, Inc., has the following mutually exclusive projects. Suppose the company uses the NPV rule to rank these two projects. Which project should be chosen if the appropriate discount rate is 15 percent?

**Choices :** A : Project A , B : Project B

**GT:** B

**Baseline:** A

**Ours:** B

Sample No.	Moisture Content	Bulk Density (g/cc)
1	4.8	2.164
2	7.5	2.322
3	7.8	2.351
4	8.9	2.345
5	9.7	2.327

**Q:** The results of a compaction test on samples of soil that are to be used for an embankment on a highway project are listed below. Determine the optimum moisture content.

**Choices:** A : 10%. , B : 8%. , C : 9%.

**GT:** B

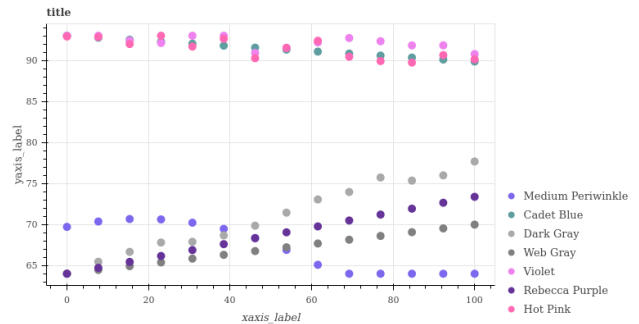
**Baseline:** C

**Ours:** B

Figure 3. Qualitative comparison on MMMU benchmark.

Brussels, Belgium, 2018. Association for Computational Linguistics. [1](#)

- [8] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand, 2024. Association for Computational Linguistics. [1](#)
- [9] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models, 2023. [1](#), [2](#)



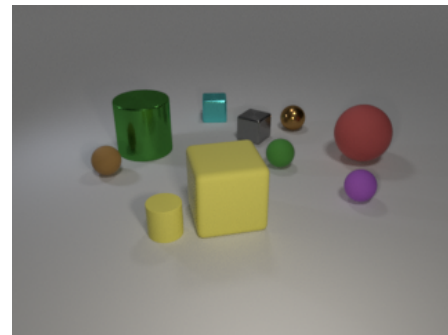
**Q:** Is Medium Periwinkle the smoothest?

**Choices :** yes, no

**GT:** no

**Baseline:** yes

**Ours:** no



**Q:** Subtract all tiny balls. Subtract all green metallic things. How many objects are left?

**GT:** 5

**Baseline:** 4

**Ours:** 5

Figure 4. Qualitative comparison on Mathvista benchmark.