

# Supplementary Materials for "VES-RFT: Rewarding Visual Evidence Sensitivity to Mitigate Hallucinations in Large Vision–Language Models"

## Supplementary Material

### 1. Overview

In this supplementary material, we provide additional details for VES-RFT, our training-time framework that encourages vision–language models to rely more on visual evidence while keeping their answers factually correct. Intuitively, we treat the reduction in the entropy of a task-specific decision variable when conditioning on the image, compared to text alone, as a visual evidence sensitivity (VES) signal: when seeing the image makes the prediction both more confident and more accurate, we reward the model for grounding its answers in the visual input rather than in language priors.

Section 2 describes the datasets, preprocessing, and how we construct simple decision variables and rewards from existing annotations. Section 3 summarizes the reinforcement fine-tuning setup, including how we apply group-based policy optimization on top of pretrained backbones and the key hyperparameters used in our experiments. Section 4 presents additional experimental results and ablations, and Section 5 discusses limitations of our current design and outlines directions for future work.

### 2. Data, Metrics, and Implementation Details

In this section, we first summarize the benchmarks and evaluation protocols used in our experiments, then describe the shared data preprocessing pipeline, the construction of task-specific decision variables and their predictive distributions, and finally the implementation of verifiable rewards and the reinforcement fine-tuning configuration.

#### 2.1. Benchmarks and Protocols

We evaluate VES-RFT on four widely used benchmarks that focus on visual grounding and hallucination:

- **MS-COCO** is an object-centric captioning dataset with ground-truth annotations for 80 object categories. We adopt the standard Karpathy split and report results on the 5k test images. MS-COCO objects also serve as reference labels for our CHAIR and AMBER evaluations.
- **POPE** is a VQA benchmark for object hallucination. Each example asks a binary question of the form "Is there a *[object]* in the image?" with a balanced mixture of existent and non-existent objects. We follow the official Random, Popular, and Adversarial splits and report Accuracy, F1, and Yes ratio.
- **CHAIR** evaluates hallucination in image captioning by comparing captioned objects to ground-truth objects in

MS-COCO. We report the instance-level ( $\text{CHAIR}_i$ ) and sentence-level ( $\text{CHAIR}_s$ ) hallucination rates.

- **AMBER** is a hallucination-centric benchmark with detailed object annotations and a generative captioning track. We report CHAIR-style hallucination scores, object coverage, hallucination rate, and cognitive alignment using the official toolkit.

A concise summary of the benchmarks, splits, and metrics is provided in Table 1 of the supplementary material.

#### 2.2. Preprocessing Pipeline

We keep the data pipeline as close as possible to the official implementations of LLaVA-1.5 and Qwen2.5-VL to avoid confounding factors.

**MS-COCO.** For all experiments involving MS-COCO, we use the Karpathy split. Each image is resized so that the shorter side is at least  $S$  pixels and the longer side does not exceed  $L$  pixels while preserving aspect ratio (LLaVA-1.5 uses  $S = L = 336$ ; Qwen2.5-VL uses its default resolution). Images are then center-cropped and normalized with the same mean and variance as the original backbones. We do not apply additional augmentations such as random cropping or color jittering, so that measured uncertainty changes can be attributed to the actual image and text content.

**POPE.** For POPE, we use the official splits (Random, Popular, Adversarial) and the standard yes/no question template. Each example asks whether a specific object is present in the image. We directly reuse the provided image-question pairs and do not perform extra filtering. At evaluation time, we restrict the model output to the first non-special token and map it to a binary label (yes / no) using a small set of verbalizers (Sec. 2.3).

**CHAIR.** For CHAIR-based captioning evaluation on MS-COCO, we follow prior work and randomly sample 500 images from the COCO 2014 validation set. Each image is paired with the prompt "Please describe this image in detail." and we decode captions with a maximum length of 512 tokens to stress-test long-form hallucinations. Object-level hallucination statistics are computed using the official CHAIR evaluation scripts and the MS-COCO object annotations.

Table 1. Datasets and evaluation metrics used in our experiments. For each benchmark, we list the task type, usage, the base split or evaluation setting, and the primary metrics reported in the main paper.

Benchmark	Task Type	Use	Split / Base Data	Metrics
POPE	Object hallucination (VQA)	Eval only	Official Random / Popular / Adversarial splits (yes/no questions about object presence)	Accuracy, F1, Yes ratio
CHAIR	Caption hallucination	Eval only	500 images sampled from MS-COCO 2014 validation split, captioned with “Please describe this image in detail.”	CHAIR <sub>s</sub> , CHAIR <sub>i</sub>
AMBER	Caption hallucination / detection	Eval only	Official generative track with 1,004 images and MS-COCO-style object annotations	CHAIR, Coverage, HalRate, Cog
MS-COCO	Captioning / pretrain	Train / Eval	Karpathy split (train/val/test), 5k test images for evaluation; also used as the base dataset for CHAIR-style hallucination analysis	CIDEr, BLEU-4, SPICE, CHAIR-style scores

**AMBER.** For AMBER, we use the official set of 1,004 images and the generative-track evaluation toolkit. We generate one caption per image using the same prompt as in CHAIR, then run the official toolkit to obtain CHAIR, coverage, hallucination rate, and cognitive alignment scores.

### 2.3. Task-Specific Decision Variables and Verbalizers

VES is defined with respect to a low-dimensional, task-specific decision variable  $Z$  that is aligned with the supervision signal. We construct  $Z$  in a lightweight and dataset-specific manner.

**POPE verbalizers.** For POPE, the task is a binary presence decision. We set  $Z \in \{\text{yes}, \text{no}\}$  and derive a binary distribution from the probability mass assigned to small sets of verbalizers for each class (e.g., {"yes", "yeah", "correct"} versus {"no", "nope", "incorrect"}) at the first answer token. This construction is shared between the with-image and no-image branches.

**CHAIR & AMBER object vocabulary.** We overload the notation  $Z$  to also denote the multi-label object vector. For CHAIR and AMBER, the decision variable is a binary vector  $Z = (Z_o)_{o \in \mathcal{O}} \in \{0, 1\}^{|\mathcal{O}|}$  over the MS-COCO object vocabulary  $\mathcal{O}$  (80 categories). We maintain a dictionary that maps each object class  $o \in \mathcal{O}$  to a set of surface forms  $\mathcal{S}_o$  (e.g., for  $o = \text{bike}$ ,  $\mathcal{S}_o = \{\text{“bike”, “bicycle”}\}$ ), and handle multi-token names such as “traffic light” by grouping the corresponding token sequences. During evaluation, we match caption tokens against this dictionary to detect object mentions and define  $Z$  and its entropy accordingly. The full list of token-object mappings is released with our code.

### 2.4. Decision-Variable Distributions

We now formalize the predictive distributions  $p_\theta(z | x, v)$  used to compute VES for different benchmarks. In all cases,

the with-image and no-image distributions share the same parameters and text input; they differ only in the visual input.

#### 2.4.1. POPE: Binary Presence Decisions

For POPE, the decision variable  $Z \in \{\text{yes}, \text{no}\}$  records whether the queried object is present in the image. Given an image-question pair  $(v, x)$ , we follow the standard yes/no template and let the model generate the first answer token  $t_1$  under teacher forcing with temperature  $T = 1$  and no decoding constraints.

Let  $\mathcal{V}_{\text{yes}}$  and  $\mathcal{V}_{\text{no}}$  denote the verbalizer sets for the two classes, and let  $P_\theta(t_1 | x, v)$  be the token distribution at the first answer position. We construct a normalized binary distribution over  $Z$  by marginalizing over the verbalizers:

$$p_\theta(Z = \text{yes} | x, v) = \frac{\sum_{w \in \mathcal{V}_{\text{yes}}} P_\theta(t_1 = w | x, v)}{\sum_{w \in \mathcal{V}_{\text{yes}} \cup \mathcal{V}_{\text{no}}} P_\theta(t_1 = w | x, v)}, \quad (1)$$

and  $p_\theta(Z = \text{no} | x, v) = 1 - p_\theta(Z = \text{yes} | x, v)$ . The same construction is used for the no-image condition by replacing  $v$  with the empty image input.

#### 2.4.2. CHAIR and AMBER: Multi-Label Object Vectors

For CHAIR and AMBER, the decision variable is a multi-label object vector  $Z = (Z_o)_{o \in \mathcal{O}}$ , where each component  $Z_o \in \{0, 1\}$  indicates the presence of object  $o$  in the model’s caption. We treat these components as conditionally independent Bernoulli variables given  $(x, v)$  and factorize the distribution as

$$p_\theta(z | x, v) = \prod_{o \in \mathcal{O}} p_\theta(Z_o = z_o | x, v), \quad z = (z_o)_{o \in \mathcal{O}}. \quad (2)$$

In practice, instead of training a separate classifier for each object, we reuse the captioning logits and approximate  $p_\theta(Z_o = 1 | x, v)$  from the probability mass assigned to the surface forms  $\mathcal{S}_o$  across the generated sequence. Let

$y = (t_1, \dots, t_L)$  be the caption and  $P_\theta(t_\ell | t_{<\ell}, x, v)$  the token distribution at position  $\ell$ . We first compute

$$q_\theta(o | x, v, y) = \sum_{\ell=1}^L \sum_{w \in \mathcal{S}_o} P_\theta(t_\ell = w | t_{<\ell}, x, v), \quad (3)$$

and then squash  $q_\theta$  into  $[0, 1]$  via a logistic transform:

$$p_\theta(Z_o = 1 | x, v) = \sigma(\alpha q_\theta(o | x, v, y) + b), \quad (4)$$

where  $\sigma$  is the sigmoid function and  $(\alpha, b)$  are fixed scaling coefficients tuned on a small validation set. We set  $p_\theta(Z_o = 0 | x, v) = 1 - p_\theta(Z_o = 1 | x, v)$ , and use the same construction for the no-image branch.

## 2.5. Verifiable Rewards

The VES reward encourages the model to reduce uncertainty due to visual evidence, but by itself does not guarantee factual correctness. We therefore introduce a complementary verifiable reward  $r_{\text{verif}}(v, x, y)$ , implemented as a frozen, task-specific checker that maps a model answer  $y$  and ground-truth annotations to a calibrated score in  $[0, 1]$ . The same verifier is used during training and evaluation to mitigate reward hacking.

**POPE.** Each POPE example has a binary ground-truth label  $z^* \in \{\text{yes}, \text{no}\}$ . We map the model’s first non-special token to a predicted label  $\hat{z}$  using the verbalizers in Eq. (1). The verifier is a normalized exact match:

$$r_{\text{verif}}(v, x, y) = \mathbb{I}[\hat{z} = z^*], \quad (5)$$

where  $\mathbb{I}[\cdot]$  is the indicator function. We use this hard exact-match criterion for both training and evaluation.

**CHAIR.** For CHAIR, we use the official toolkit to extract objects  $O(y)$  mentioned in the generated caption  $y$  and compare them with the ground-truth object set  $O^*$  of the image. The instance- and sentence-level hallucination rates are stated in the main paper. We convert these into a reward by linearly mapping the hallucination rate to  $[0, 1]$ :

$$r_{\text{verif}}(v, x, y) = 1 - \lambda_i \text{CHAIR}_i(y) - \lambda_s \text{CHAIR}_s(y), \quad (6)$$

with non-negative weights  $\lambda_i, \lambda_s$ . In all reported experiments, we use  $\lambda_i = \lambda_s = 1$  and clip  $r_{\text{verif}}$  to  $[0, 1]$ .

**AMBER.** For AMBER, the official toolkit returns a normalized hallucination score  $\text{Hall}(y) \in [0, 1]$  (lower is better) and a cognitive alignment score  $\text{Align}(y) \in [0, 1]$  (higher is better). We define

$$r_{\text{verif}}(v, x, y) = \gamma_{\text{align}} \text{Align}(y) - \gamma_{\text{hall}} \text{Hall}(y), \quad (7)$$

with hyperparameters  $\gamma_{\text{align}}, \gamma_{\text{hall}} \geq 0$ . In our experiments we set  $\gamma_{\text{align}} = \gamma_{\text{hall}} = 1$  and rescale  $r_{\text{verif}}$  to  $[0, 1]$  using min-max normalization on the training set.

## 3. Reinforcement Fine-Tuning with GRPO

VES-RFT is applied as a lightweight reinforcement fine-tuning (RFT) stage on top of instruction-tuned vision-language checkpoints. We adopt a critic-free Group Relative Policy Optimization (GRPO) objective that uses relative rewards within small groups of sampled responses, allowing us to avoid training a separate value network while still obtaining a low-variance advantage signal.

**Backbones.** Unless otherwise specified, we initialize from the official LLaVA-1.5-7B and Qwen2.5-VL-7B checkpoints and keep both the vision and language encoder architectures unchanged. Only the unified VLM parameters are updated during RFT.

**Optimization.** We use AdamW with a cosine learning-rate schedule and linear warm-up. For both LLaVA-1.5 and Qwen2.5-VL we set the initial learning rate to  $2 \times 10^{-5}$ , use a warm-up ratio of 3%, and decay the learning rate to zero with cosine annealing. Training runs for  $T$  steps (typically 5k–10k iterations depending on the backbone) with a global effective batch size of 64–128 samples achieved via gradient accumulation. Weight decay is set to 0.1, and we enable mixed-precision training (bfloat16 or fp16 depending on hardware). Empirically, we find VES-RFT to be robust to a reasonable range of learning rates as long as KL regularization is present and the update horizon is kept short.

**Group sampling and rewards.** Let  $\pi_\theta$  denote the current VLM policy and  $\pi_{\text{ref}}$  a frozen reference policy (typically the supervised fine-tuned checkpoint). Given an image-text query  $(v, x)$ , we first construct a small group of  $G$  candidate responses by sampling from the current policy

$$\{y_i\}_{i=1}^G \sim \pi_\theta(\cdot | v, x), \quad (8)$$

using nucleus sampling (top- $p = 0.9$ , temperature 0.7) with a maximum length of  $L_{\text{max}}$  tokens.

For each candidate  $y_i$ , we compute a scalar dual reward

$$r_i = r(v, x, y_i) = r_{\text{verif}}(v, x, y_i) + \lambda r_{\text{ves}}(v, x, y_i), \quad (9)$$

where  $\lambda$  is the VES weight (set to 1 in our main experiments),  $r_{\text{verif}}$  is the task-specific verifiable reward described in Sec. 2.5, and  $r_{\text{ves}}$  is the shaped entropy-gap reward defined in the main paper. Rewards are standardized within each group before the GRPO update to stabilize optimization. The KL regularization coefficient  $\beta$  is chosen in the range  $[0.01, 0.05]$ , and the method is not overly sensitive within this range.

**Group-relative advantage and GRPO objective.** Instead of estimating a separate value function, GRPO defines a *group-relative* advantage by normalizing rewards within each sampled group. For a fixed query  $(v, x)$  with rewards  $\{r_i\}_{i=1}^G$ , we compute the group mean and variance

$$\mu_r = \frac{1}{G} \sum_{i=1}^G r_i, \quad (10)$$

$$\sigma_r^2 = \frac{1}{G} \sum_{i=1}^G (r_i - \mu_r)^2, \quad (11)$$

and obtain a normalized advantage for each response as

$$\tilde{r}_i = \frac{r_i - \mu_r}{\sigma_r + \varepsilon}, \quad (12)$$

where  $\varepsilon > 0$  is a small constant for numerical stability.

The GRPO objective for updating  $\theta$  can then be written as

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(v,x)} \left[ \frac{1}{G} \sum_{i=1}^G \tilde{r}_i \log \frac{\pi_\theta(y_i | v, x)}{\pi_{\text{ref}}(y_i | v, x)} \right. \\ \left. - \beta \text{KL} \left( \pi_\theta(\cdot | v, x) \parallel \pi_{\text{ref}}(\cdot | v, x) \right) \right], \quad (13) \end{aligned}$$

where  $\beta \geq 0$  controls how tightly the updated policy is regularized around the reference policy. In practice, the KL term is estimated from sampled trajectories and implemented as a token-level penalty, similar to PPO-style updates. During training we minimize the negative objective in Eq. (13) plus the KL term, as summarized in Algorithm 1.

**Preference pair construction.** Our RFT stage uses a relatively small number of preference pairs compared to prior RLHF or RFT work. We derive pairs from the same underlying datasets used for evaluation, namely MS-COCO (captioning and CHAIR) and POPE (object hallucination), and optionally include additional instruction-tuning image–text pairs from the base models. RFT pairs are constructed only from the training splits of MS-COCO. For each sampled image–prompt pair  $(v, x)$ , we draw  $K$  candidate responses  $\{y^{(k)}\}_{k=1}^K$ , score them with the task-specific verifier  $V$  to obtain

$$s^{(k)} = V(v, x, y^{(k)}), \quad (14)$$

and then form a preference pair  $(y^+, y^-)$  by selecting the best and worst candidates

$$y^+ = \arg \max_k s^{(k)}, \quad y^- = \arg \min_k s^{(k)}. \quad (15)$$

To avoid ambiguous supervision, we require a minimum margin  $s^+ - s^- \geq \delta$  for some  $\delta > 0$ ; otherwise the instance is discarded. The resulting mixture of POPE-style VQA prompts and MS-COCO-style captioning prompts helps the model retain general conversational ability while receiving strong anti-hallucination signals.

---

### Algorithm 1: VES-RFT training with GRPO

---

**Input:** Supervised checkpoint (reference policy)

$\pi_{\text{ref}}$ ,  
training dataset  $\mathcal{D}$  of image–text pairs  $(v, x)$ ,  
VES reward  $r_{\text{ves}}$ , verifiable reward  $r_{\text{verif}}$ ,  
group size  $G$ , KL weight  $\beta$ , VES weight  $\lambda$ .

**Output:** Fine-tuned policy  $\pi_\theta$ .

Initialize policy parameters  $\theta \leftarrow \theta_{\text{ref}}$  from  $\pi_{\text{ref}}$ ;

**while not converged do**

  Sample a mini-batch of queries

$\{(v^{(b)}, x^{(b)})\}_{b=1}^B$  from  $\mathcal{D}$ ;

**for**  $b = 1$  **to**  $B$  **do**

    Sample a group of  $G$  responses

$\{y_i^{(b)}\}_{i=1}^G \sim \pi_\theta(\cdot | v^{(b)}, x^{(b)})$ ;

**for**  $i = 1$  **to**  $G$  **do**

      Compute VES reward

$r_{\text{ves}}^{(b,i)} \leftarrow r_{\text{ves}}(v^{(b)}, x^{(b)}, y_i^{(b)})$ ;

      Compute verifiable reward

$r_{\text{verif}}^{(b,i)} \leftarrow r_{\text{verif}}(v^{(b)}, x^{(b)}, y_i^{(b)})$ ;

      Combine into dual reward

$r^{(b,i)} \leftarrow r_{\text{verif}}^{(b,i)} + \lambda r_{\text{ves}}^{(b,i)}$ ;

    Compute group mean and std:

$\mu_r^{(b)} \leftarrow \frac{1}{G} \sum_{i=1}^G r^{(b,i)}$ ,

$(\sigma_r^{(b)})^2 \leftarrow \frac{1}{G} \sum_{i=1}^G (r^{(b,i)} - \mu_r^{(b)})^2$ ;

**for**  $i = 1$  **to**  $G$  **do**

      Normalize reward to obtain  
      group-relative advantage:

$\tilde{r}^{(b,i)} \leftarrow \frac{r^{(b,i)} - \mu_r^{(b)}}{\sigma_r^{(b)} + \varepsilon}$ ;

  Estimate GRPO loss over the batch:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = - \frac{1}{BG} \sum_{b=1}^B \sum_{i=1}^G \tilde{r}^{(b,i)} \log \frac{\pi_\theta(y_i^{(b)} | v^{(b)}, x^{(b)})}{\pi_{\text{ref}}(y_i^{(b)} | v^{(b)}, x^{(b)})} \\ + \beta \text{KL}(\pi_\theta(\cdot | v^{(b)}, x^{(b)}) \parallel \pi_{\text{ref}}(\cdot | v^{(b)}, x^{(b)})) \end{aligned}$$

  Update  $\theta$  with one step of AdamW on

$\mathcal{L}_{\text{GRPO}}(\theta)$ ;

**return**  $\pi_\theta$ ;

---

**Connection to VES-RFT and test-time behavior.** In VES-RFT, the scalar reward  $r_i$  for each response  $y_i$  is exactly the dual reward in Eq. (9), where  $r_{\text{verif}}$  encourages semantic correctness and  $r_{\text{ves}}$  rewards reduction of uncertainty attributable to visual evidence. The group-relative normalization ensures that both components contribute stably to the advantage, even though  $r_{\text{verif}}$  can be nearly binary while  $r_{\text{ves}}$  may vary over a wider range. The KL regularization term prevents large policy drift away from the supervised checkpoint, which we find crucial for stability when using a relatively small number of preference pairs.

At test time, the GRPO machinery is entirely discarded:

we use the fine-tuned policy  $\pi_\theta$  for standard single-pass decoding conditioned on the image, without any additional critics, verifiers, or decoding-time heuristics.

## 4. Experiments

In this section, we evaluate VES-RFT on standard hallucination benchmarks and analyze its robustness and generalization. We first describe the experimental setup, and then present results on POPE (object hallucination in VQA), CHAIR and AMBER (caption hallucination), followed by ablations on the reward components and additional robustness and caption quality experiments.

### 4.1. Robustness and Generalization

We follow the evaluation protocols in Sec. 2.1 and Table 1 and focus here on robustness experiments.

Finally, we examine whether VES-RFT remains effective under noisy visual conditions and whether it preserves general caption quality on MS-COCO. We construct perturbed versions of the POPE and CHAIR datasets by applying mild Gaussian blur and random occlusions to each image, while keeping the prompts and decoding procedures unchanged. We then evaluate POPE F1 and Acc on both the clean and mildly degraded images.

*Perturbation strength.* To make the amount of visual degradation comparable to prior work, we calibrate our perturbations to roughly match the noise level corresponding to  $T = 800$  steps in the Gaussian forward-diffusion process of Visual Contrastive Decoding [1]. At this strength, the images still preserve the global scene layout and the identities of most salient objects, but fine-grained textures, edges, and low-contrast regions are visibly corrupted, which we regard as a mildly-to-moderately degraded regime.

Table 2 shows that VES-RFT maintains a consistent advantage over the base models in the clean and mild settings. All methods suffer performance drops under mild degradations, but VES-RFT consistently retains higher F1 and accuracy than the corresponding baselines. This is consistent with our design: when the visual evidence becomes unreliable, the entropy gap shrinks and the VES signal naturally weakens, preventing the model from over-committing to hallucinated objects.

## 5. Discussion

**Limitations.** While VES-RFT consistently reduces hallucinations across our evaluated settings, there are several aspects that bound the current scope. First, we primarily instantiate and test the method on two representative VLM families (LLaVA-1.5-7B and Qwen2.5-VL-7B) and on object-centric hallucination benchmarks such as POPE, CHAIR, and AMBER, where object presence is the main source of supervision. A broader evaluation on more di-

Table 2. Robustness to synthetic image degradations on POPE. We report POPE F1 ( $\uparrow$ ) and Acc ( $\uparrow$ ) for clean and mildly degraded images.

Model	POPE F1 ( $\uparrow$ )		POPE Acc ( $\uparrow$ )	
	Clean	Mild	Clean	Mild
LLaVA-1.5-7B	80.43	60.18	82.04	70.28
LLaVA-1.5-7B + VES-RFT	85.61	65.72	86.96	75.83
Qwen2.5-VL-7B	70.86	54.76	84.84	74.29
Qwen2.5-VL-7B + VES-RFT	85.57	60.48	88.93	77.36

verse multimodal tasks (e.g., open-domain dialog, compositional reasoning, or video) is therefore left to future work rather than a limitation of the framework itself. Second, our verifiable rewards rely on task-specific object annotations and lexical matching rules; although we find them stable enough for the benchmarks considered, they may under-reward some semantically correct paraphrases or rare expressions. Finally, computing the VES signal requires an additional no-image forward pass during reinforcement fine-tuning, which modestly increases the training cost; however, inference remains single-pass and does not incur extra overhead.

**Future Directions.** Our results suggest several directions to further develop VES-based training. One direction is to design richer verifiers that go beyond object presence and incorporate relations, attributes, and spatial or temporal consistency, for example by leveraging scene-graph or relation detectors so that VES-RFT can directly encourage visually grounded reasoning, not only object-level correctness. A second direction is to refine the uncertainty estimation underlying VES, exploring better-calibrated predictive distributions or ensemble-based estimates, which may yield more informative visual-evidence signals. It is also natural to combine VES-RFT with complementary alignment signals such as human preferences or rationale-level verifiers, encouraging the model to both rely on the image and explicitly justify its decisions in a visually consistent way. Finally, decoding-time methods such as contrastive or mutual-information-based reweighting are largely orthogonal to our training-time approach; jointly designing training objectives and inference strategies is a promising avenue for further reducing hallucinations while preserving general language ability.

## References

- [1] Sicong Leng, Hang Zhang, Guanzheng Chen, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024. 5

Q: Please describe this image in detail.



**LLaVA-1.5:** The image captures a baseball game in progress, with a baseball player wearing a blue hat and jersey walking across the field. The player is holding a baseball glove, ready to catch the ball. There are several other people on the field, including a player in the background and a few others scattered around the area. In the background, there is a **baseball bat visible**, indicating that the game is in full swing. The scene is lively, with players actively participating in the game.

**VES-RFT:** The photo shows an active baseball game. A player in a blue cap and uniform is moving across the field with a glove in hand, appearing prepared to make a play. The scene suggests live action on the field, with additional players visible in the background. Overall, it conveys the intensity of an ongoing match.

Q: Please describe this image in detail.



**LLaVA-1.5:** The image captures a heartwarming moment of a man and a woman holding their children in the air. **The man is on the left side of the scene, while the woman is on the right side.** The children are being lifted by their parents, creating a joyful atmosphere. There are two children in the scene, one near the center and the other towards the right side. The woman is wearing a black dress, and the man is wearing a plaid shirt. The scene appears to be set in a park, as there are trees in the background.

**VES-RFT:** The picture captures a touching family scene where a woman and a man are joyfully lifting their children high into the air. The mother stands on the left and the father on the right, each holding up one of the kids, creating a playful, uplifting moment between them. There are two children, one closer to the center and the other nearer to the right side. The woman is dressed in a black outfit, while the man sports a plaid shirt. The background shows trees and open space, suggesting they're spending time together in a park, adding to the warm and lively atmosphere.

Figure 1. Comparison of generated description with baseline.