SEAL: Spatial Entity Alignment Learning for Reducing Hallucinations in Fashion Image Captioning

Anonymous ACL submission

Abstract

Fashion captions often contain local, attribute-bearing hallucinations: a mostly correct sentence may add unsupported entities such as *bow* or *skirt*. We present SEAL, a token-level spatial alignment framework that trains fashion VLMs to prefer visual claims grounded in corresponding garment regions. Given Fashionpedia-style masks, SEAL extracts generated visual chunks, normalizes them to garment categories, measures token-to-mask Attention Recall, and trains LoRA adapters with grouped GRPO. Its reward combines category coverage, hallucinated-category penalties, spatial gating, and group-level hallucination filtering. On 3,600 Fashionpedia test images, SEAL improves semantic category F-score from 0.4748 to 0.4994, reduces hallucinated-category count from 0.6006 to 0.4854, and improves grounded POPE-F1 from 0.5635 to 0.5846 over three seeds. Average Attention Recall remains stable, so we use attention as training and diagnostic feedback rather than as evidence of a global attention shift. A DPO-init control does not reduce hallucinated categories. Code will be available after paper acceptance.

1 Introduction

Vision-language models (VLMs) generate fluent image descriptions, but fluency can hide plausible yet absent content. In fashion search, cataloging, styling assistance, and accessibility, confusing a *dress* with a *skirt*, or adding an unsupported *bow*, changes the structured information that downstream systems and users rely on.

Fashion captioning is a useful stress test because descriptions are visually dense and linguistically specific. Prior work shows that accurate product descriptions require categories, parts, and attributes beyond generic objects (Yang et al., 2020). Fashionpedia exposes this structure through apparel categories, parts, accessories, attributes, and localization annotations (Jia et al., 2020). Yet VLMs may still lean on co-occurrence patterns, describing a *jacket* because an upper garment is present or adding a *bow* because it is stylistically likely.

Figure 1: **Fashion hallucinations targeted by SEAL.** SEAL addresses two common errors in fashion captioning: unsupported entity fabrication, such as hallucinating a "bow", and local category confusion, such as describing a dress as a skirt. The method encourages each generated fashion token to be grounded in its corresponding visual region.

As illustrated in Figure 1, we target two local hallucination patterns: unsupported entity fabrication and local category confusion. These errors are local enough that a caption can remain fluent and mostly plausible while still injecting incorrect structured fashion information.

A central difficulty is locality: one noun phrase can be wrong while the rest of the caption is useful. Recent entity-centric preference work shows that hallucination feedback should target factual units such as entities, attributes, and relations rather than only whole responses (Wu et al., 2025). Complementary intervention evidence links hallucinated text to attention heads that insufficiently attend to image tokens (Sarkar et al., 2025b). These findings motivate a fashion-specific question: when a model generates an attribute-bearing garment phrase, does the responsible token attend to the region that licenses that claim? Fine-grained work such as RLHF-V and HALVA uses segment- or

phrase-level feedback (Yu et al., 2024; Sarkar et al., 2025a); we ask whether fashion masks can provide this feedback automatically.

We answer with SEAL (Spatial Entity Alignment Learning), which aligns generated fashion tokens with segmentation masks. For each image, SEAL samples captions from Qwen2-VL-2B-Instruct, parses visual chunks such as *red puff shirt*, normalizes them to Fashionpedia categories, extracts token-to-vision attention, and computes *Attention Recall*: the probability mass assigned by a generated token to the mask of the mentioned category. A semantic-spatial reward combines this evidence with category coverage and hallucination penalties, and a GRPO-style update normalizes rewards within each image-conditioned group without a value model. Our claim is that mask-derived token feedback can discourage unsupported entity mentions, not that average attention mass must increase globally.

SEAL differs from response-level preference learning by tying supervision to the visual token that needs evidence. EMPO aligns multimodal preferences at the entity level, while SPIN suppresses image-inattentive heads during inference. In contrast, SEAL learns spatially grounded attention alignment during optimization: it rewards fashion entities and their modifiers when the generated tokens are region-supported. We do not treat attention as a complete causal explanation; we use token-to-mask attention as an operational signal that is available during training and checkable against localization labels.

Our contributions are as follows:

- We formulate fine-grained fashion hallucination reduction as token-level spatial alignment between generated visual chunks and Fashionpedia-style masks.
- We introduce a semantic-spatial reward that combines category coverage, hallucination penalties, attention-recall gating, and group-level hallucination filtering.
- We adapt GRPO to online fashion caption rollouts, updating a LoRA adapter with group-relative advantages and no value model.
- We evaluate on 3,600 Fashionpedia test images with three seeds, grounded POPE queries, and an adapter-only DPO-init control.

2 Related Work

Fashion image captioning. Fashion captioning requires garment categories, parts, materials, colors, and style attributes. Semantic-reward and attribute-aware captioning methods improve fine-grained accuracy beyond maximum-likelihood training (Yang et al., 2020; Li et al., 2021; Cai et al., 2024), and FACap highlights the need for fashion-domain language in composed image retrieval (Gardères et al., 2025). SEAL builds on this tradition but adds a spatial requirement: a category mention should be grounded in its region, and unsupported categories should receive token-level penalties.

Object hallucination in VLMs. Hallucination has been measured with object-level metrics such as CHAIR (Rohrbach et al., 2018) and POPE-style polling over present and absent objects (Li et al., 2023). Training-free methods such as VCD and OPERA reduce language-prior reliance through visual contrast or decoding penalties (Leng et al., 2024; Huang et al., 2024). They are broadly applicable, but they do not use masks to assign local reward to generated fashion tokens.

Attention grounding and interpretability. Attention-based captioning made word-image alignment visible (Xu et al., 2015), while grounded-captioning work showed that fluent captions can still have poor word-region grounding (Zhou et al., 2020). Transformer explainability studies caution that cross-modal attention requires model-aware interpretation (Chefer et al., 2021). SPIN strengthens the link between attention and hallucination through intervention: it identifies image-inattentive heads for each query token and suppresses them during inference to reduce hallucinations (Sarkar et al., 2025b). SEAL uses this insight at a different granularity and stage: instead of pruning heads at decoding time, it rewards generated fashion tokens when their visual attention falls inside the relevant category mask during training.

Fine-grained hallucination feedback. Recent methods show that hallucination mitigation benefits from localized supervision. RLHF-V uses segment-level human corrections, HALVA aligns hallucinated and correct phrase pairs, and region-level captioning constrains generation with region or control-word signals (Yu et al., 2024; Sarkar et al., 2025a; Zhao et al., 2025). SEAL follows this

fine-grained view but derives token feedback from dataset masks and normalized fashion categories.

Preference optimization and RL alignment. RLHF often uses PPO-style objectives (Ouyang et al., 2022), while DPO optimizes over chosen and rejected responses (Rafailov et al., 2023). VLM alignment variants include HA-DPO, RLAIF-V, V-DPO, and EMPO (Zhao et al., 2024; Yu et al., 2025; Xie et al., 2024; Wu et al., 2025). EMPO aligns multimodal preferences at the entity level by constructing image-, instruction-, and response-side negatives around entities, attributes, and relations, showing the value of focusing preference supervision on hallucination-relevant factual units. SEAL adopts this fine-grained spirit but changes both the supervision source and the optimization signal: it derives spatial rewards from masks and token-to-region attention, then applies grouped online GRPO rather than response-pair DPO. GRPO estimates advantages within sampled groups without a separate critic (Shao et al., 2024).

3 Method

3.1 Problem Formulation

Let I be a fashion image and $\mathcal{M} = \{(c, m_c)\}$ its annotations, where c is a canonical Fashionpedia category and $m_c \in \{0,1\}^{H\times W}$ is a mask. A VLM policy π_θ generates a caption $y = (y_1, \ldots, y_T)$ conditioned on I. We aim to improve visual faithfulness while preserving coverage; a generated category is hallucinated if its normalized category is absent from the ground-truth set G.

SEAL assumes that some generated tokens can be aligned with visual categories. When y_t belongs to a chunk mapped to c, the token should place sufficient visual attention mass inside m_c. Figure 2 shows this process as a training loop from Fashionpedia masks and grouped caption rollouts to visual-chunk parsing, token–mask evidence, semantic–spatial reward construction, and GRPO-based LoRA updates.

3.2 Visual Chunk Extraction and Category Normalization

SEAL parses each sampled caption into visual chunks: a fashion noun and its visually meaningful modifiers. In *red puff shirt*, for example, *red*, *puff*, and *shirt* are bound to the shirt category. Dependency parsing identifies nouns and modifiers, and character spans are mapped back to autoregressive token positions through progressive tokenizer decoding.

For comparison with Fashionpedia annotations, a curated normalization layer maps aliases such as *trousers/jeans* to pants, *gown* to dress, and *handbag/purse* to bag/wallet. Ambiguous aliases are resolved against ground-truth categories when possible and reused for scoring, hallucination counting, and token-mask matching.

3.3 Attention Recall over Category Masks

We use generated-token-to-visual-token attention as an operational grounding signal. For each token y_t aligned to category c, we extract its attention distribution over the visual-token span. Because Qwen2-VL places visual tokens in the decoder sequence, this is cross-modal self-attention. We infer the post-merge visual grid, reshape the attention vector, and upsample it to the mask resolution. Let $m_c \in \{0,1\}^{H\times W}$ be the binary mask for category c, and let $\bar{a}_t(i,j)$ be the normalized attention probability at pixel (i,j) on the same $H \times W$ grid. The token-level score is:

$$\mathrm{AttnRec}(t,c) = \sum_{i=1}^{H}\sum_{j=1}^{W} \bar{a}_t(i,j) m_c(i,j). \quad (1)$$

This probability mass, not a thresholded pixel IoU, measures how much token attention falls inside the supporting region.

For split words, SEAL averages consecutive subword attention before scoring. The main configuration averages selected mid-late decoder layers to reduce single-layer noise.

3.4 Semantic and Hallucination Reward

Let $P_y = P(y)$ be the set of normalized categories mentioned in caption y, and let G be the ground-truth category set. We use a recall-weighted semantic score:

$$\mathrm{SemF}(P_y, G) = \frac{5\,\mathrm{Prec}(P_y,G)\,\mathrm{Rec}(P_y,G)}{4\,\mathrm{Prec}(P_y,G) + \mathrm{Rec}(P_y,G)}. \quad (2)$$

where $\mathrm{Prec}(P_y,G) = |P_y \cap G| / \max(|P_y|, 1)$ and $\mathrm{Rec}(P_y,G) = |P_y \cap G| / \max(|G|, 1)$. This F2-style form discourages avoiding hallucination by mentioning very few garments. Hallucinated categories are:

$$\mathrm{Hall}(y,G) = |P_y \setminus G|. \quad (3)$$

Here $\setminus$ denotes set difference. The semantic component is:

$$S_{\mathrm{sem}}(y) = \mathrm{SemF}(P_y,G) - \lambda_h \,\mathrm{Hall}(y,G). \quad (4)$$

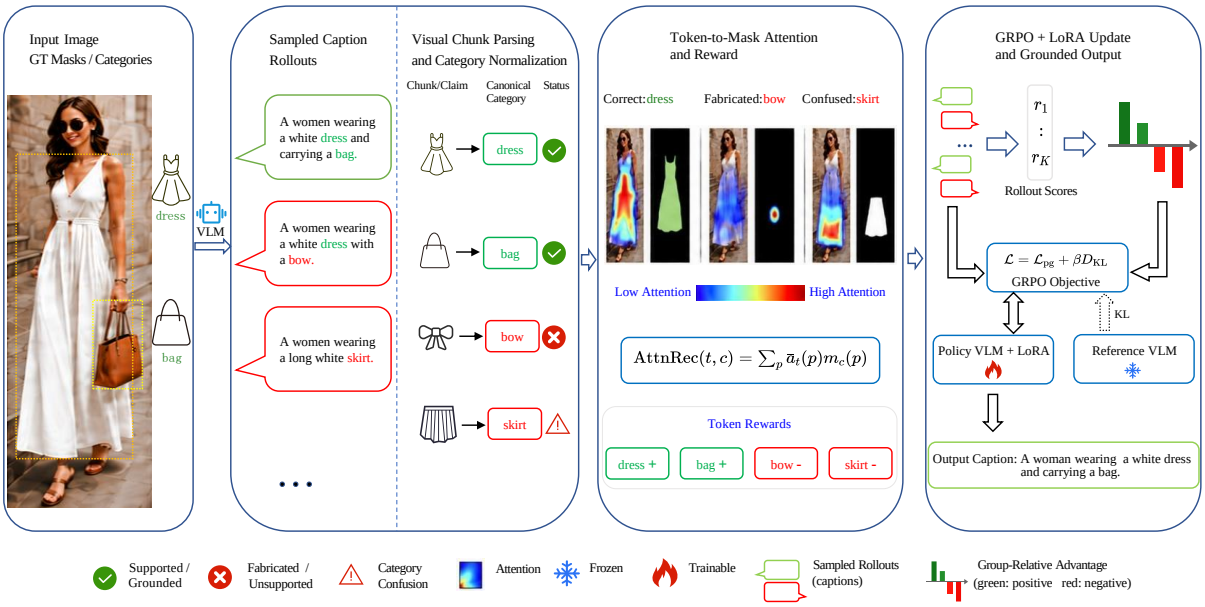

Figure 2: **SEAL framework.** The figure illustrates the left-to-right training loop for spatially grounded fashion caption alignment. Given a fashion image and Fashionpedia masks, SEAL samples grouped captions, extracts attribute-bearing visual chunks, normalizes them to garment categories, and measures token-to-mask attention evidence. Category coverage, hallucination penalties, spatial attention gating, and group-level filtering are then combined into token rewards for KL-constrained GRPO updates of the LoRA policy, with grounded metrics and heat-map diagnostics used for evaluation.

We use $\lambda_h = 1.10$, making hallucination suppression a primary reward term rather than a lightweight post-hoc adjustment.

3.5 Spatially Gated Token Reward

For each generated token t aligned to category c, the reward combines semantic and spatial evidence. If $S_{\mathrm{sem}}(y) > 0$, the positive semantic signal is gated by attention recall:

$$r_t = S_{\mathrm{sem}}(y) \cdot \eta \max(0, \mathrm{AttnRec}(t,c) - \tau) - \lambda_a \mathbb{I}[\mathrm{AttnRec}(t,c) < \tau], \tag{5}$$

where $\tau = 0.032$ is the attention floor, $\eta = 10$ is a scaling factor, $\lambda_a = 0.14$ penalizes low-attention tokens, and $\mathbb{I}[\cdot]$ denotes the indicator function. If $S_{\mathrm{sem}}(y) \leq 0$, the negative semantic signal is retained. If no visual chunk is matched, SEAL distributes a bounded fallback reward, with stronger negatives when hallucinated categories are present.

3.6 Hallucination Hard Gate

GRPO ranks rollouts within each image-conditioned group. Because a high-recall rollout can still contain unsupported categories, SEAL applies a group-level hard gate before advantage normalization. For rollouts $\{y^{(i)}\}_{i=1}^{K}$, where $y^{(i)}$ is the caption sampled in rollout i, define $h_i = \mathrm{Hall}(y^{(i)}, G)$ and let $\tilde{h} = \mathrm{median}(\{h_i\}_{i=1}^{K})$. Token rewards in rollouts with $h_i > \tilde{h}$ are clamped to non-positive values and penalized by $\lambda_g = 0.38$:

$$r_t^{(i)} \leftarrow \min(r_t^{(i)}, 0) - \lambda_g. \tag{6}$$

This keeps high-hallucination rollouts from winning the relative ranking through reward artifacts.

3.7 Token-Level GRPO Objective

For each image, SEAL samples $K = 4$ rollouts, with T_i denoting the token length of rollout i. Token rewards are aggregated into $R_i = \frac{1}{T_i}\sum_t r_t^{(i)}$ and normalized within the group:

$$A_i = \mathrm{clip}\left(\frac{R_i - \mu_R}{\sigma_R + \epsilon}, -2, 2\right), \tag{7}$$

where μ_R and σ_R are computed over the group, ϵ is a small numerical stabilizer, and the sequence advantage is broadcast to its tokens.

The adapter is updated with a clipped policy objective and a KL penalty to a reference policy, implemented by disabling the active LoRA adapter:

$$\mathcal{L}_{\mathrm{pg}} = -\mathbb{E}_{i,t}\Big[\min\Big(\rho_{i,t}A_i, \mathrm{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c)A_i\Big)\Big], \tag{8}$$

where $\mathbb{E}_{i,t}$ averages over rollout tokens, π_θ is the current LoRA policy, π_{old} is the rollout policy snapshot, $\rho_{i,t} = \exp(\log \pi_\theta(y_t^{(i)}) - \log \pi_{\text{old}}(y_t^{(i)}))$, and $\epsilon_c = 0.2$. We use the GRPO KL estimator against the frozen reference policy π_{ref}:

$$D_{\text{KL}} = \exp(\log \pi_{\text{ref}} - \log \pi_\theta) - (\log \pi_{\text{ref}} - \log \pi_\theta) - 1. \tag{9}$$

The final objective is $\mathcal{L} = \mathcal{L}_{\text{pg}} + \beta D_{\text{KL}}$ with $\beta = 0.10$.

3.8 Category Curriculum

Fashionpedia contains large garments, accessories, and small parts. To stabilize reward assignment, SEAL schedules the mask set: the first 260 steps use main apparel categories, steps 260–560 add accessories, and later steps use all categories. This exposes the policy to reliable large regions before noisier small objects and parts. Algorithm 1 summarizes the full training procedure.

4 Experimental Setup

Dataset. We use Fashionpedia, the ECCV 2020 dataset of Jia et al. (2020), which introduces a fashion ontology, segmentation annotations, and attribute localization for fine-grained fashion understanding. Fashionpedia provides apparel categories, garment parts, accessories, segmentation masks, and localized attributes. Our implementation uses its 46-category category ontology and converts the available segmentation annotations or bounding boxes into category masks for token-region rewards and grounded evaluation. From the 45,585-image Fashionpedia index used in this work, we fix an image-level split with random seed 42: 39,985 images for training, 2,000 images for validation and sanity checks, and 3,600 held-out images for final testing. The validation and test splits are excluded from SEAL training and preference-data construction.

Model and training. The base VLM is Qwen2-VL-2B-Instruct (Wang et al., 2024). Main SEAL adapters start directly from the no-adapter policy; no DPO or catalog-captioning adapter is used. We train LoRA adapters with rank 16, alpha 32, dropout 0.05, and attention/MLP projection targets. Training uses BF16, no QLoRA, $K = 4$, learning rate 10^{-5}, maximum generation length 128, temperature 1.0, top-$p = 0.9$, and 3,000 update steps on a single RTX 5090. The available DPO adapter is evaluated only as an adapter-only initialization control.

Evaluation metrics. We report semantic category F-score, hallucinated-category count per sample, attention recall, attention coverage, and grounded POPE. The semantic metric follows Eq. 2; hallucination rate is the average number of normalized predicted categories absent from G. Grounded POPE uses random, popular, and hard negative category queries with $k = 12$ negatives per mode; a positive answer must mention the category and satisfy an attention threshold of 0.02.

Baselines and reporting. The primary baseline is no-adapter Qwen2-VL-2B-Instruct under the same fixed-grid, native-resolution, chunk-level protocol. SEAL starts from this policy and applies the spatial GRPO stage. The adapter-only DPO control uses the same evaluation without SEAL updates, testing whether a preference-tuned starting point accounts for the gains. We leave full VCD, OPERA, SPIN, HA-DPO, V-DPO, and EMPO reproductions to future work because adapting decoding-time head suppression and response-pair preference methods to the same mask-grounded reward and grounded-POPE protocol is nontrivial; our claim is narrower: under a shared base VLM and evaluation pipeline, token-level spatial GRPO improves hallucination and grounded factuality over the no-adapter model.

5 Results

Table 1 shows the main results. The DPO init control improves semantic F-score and grounded POPE-F1, but it increases hallucinated-category count from 0.6006 to 0.6203 and lowers Attention Recall from 0.2692 to 0.2245. Across three seeds, SEAL reduces hallucinated categories by 0.1152 per sample, a 19.2% relative reduction over the no-adapter baseline. Semantic category F-score improves by 0.0246, grounded POPE-F1 by 0.0210, false positive rate decreases by 0.0020, and POPE accuracy improves by 0.0075.

Because the main SEAL adapters start from no-adapter Qwen2-VL-2B-Instruct and the DPO control does not reduce hallucination, we attribute the gains to the SEAL reward and token-level GRPO update.

Attention recall remains approximately stable: the mean change is +0.0003 with seed-level variation in both directions. Thus SEAL's main effect

Algorithm 1: Token-Level Spatially Grounded GRPO

Input: image I, masks $\mathcal{M}$, policy π_θ, reference π_{ref}, rollouts K, coefficients $\tau, \lambda_h, \lambda_a, \lambda_g, \beta$
for each training image I **do**
  Sample captions $\{y^{(k)}\}_{k=1}^{K} \sim \pi_\theta(\cdot|I)$ # Grouped rollouts
  for each rollout k **do**
    Extract visual chunks and normalize mentions $P(y^{(k)})$ # Entity claims
    Map chunk spans to generated token positions
    for each aligned token-category pair (t, c) **do**
      Extract token-to-vision attention $\bar{a}_t$
      $\text{AttnRec}(t, c) = \sum_p \bar{a}_t(p) m_c(p)$ # Token–mask evidence
    end for
    $S_{\text{sem}} = \text{SemF}(P(y^{(k)}), G) - \lambda_h \, \text{Hall}(y^{(k)}, G)$ # Coverage and hallucination
    Assign token rewards r_t with the attention floor τ and fallback rule # Spatial gate
  end for
  Apply group hallucination gate using median $\text{Hall}(y^{(k)}, G)$
  Aggregate token rewards into rollout scores R_k
  Normalize $A_k = \text{clip}((R_k - \mu_R)/(\sigma_R + \epsilon), -2, 2)$ # Group-relative advantage
  Update LoRA parameters with clipped GRPO objective plus βD_{KL}
end for
Return: LoRA-adapted SEAL policy

Method	Sem.F ↑	Hall. ↓	AttnR ↑	G-POPE F1 ↑	FPR ↓	Acc. ↑
Qwen2-VL-2B	0.4748	0.6006	0.2692	0.5635	0.0209	0.8035
DPO init only	0.4956	0.6203	0.2245	0.5809	0.0214	0.8088
SEAL seed 123	**0.5034**	0.4775	0.2695	0.5866	0.0186	0.8118
SEAL seed 777	0.4923	0.5083	**0.2717**	0.5779	0.0198	0.8085
SEAL seed 2026	0.5026	**0.4703**	0.2673	**0.5892**	**0.0183**	**0.8127**
SEAL mean ± std	0.4994±0.0062	0.4854±0.0202	0.2695±0.0022	0.5846±0.0059	0.0189±0.0008	0.8110±0.0022
Δ vs. baseline	+0.0246	-0.1152	+0.0003	+0.0210	-0.0020	+0.0075

Table 1: Main results on 3,600 Fashionpedia test images. Hall. is average hallucinated-category count per sample. DPO init only evaluates the available DPO adapter without SEAL updates. SEAL starts from the no-adapter policy and reduces hallucination while improving grounded POPE.

is semantic hallucination reduction under stable token-mask attention coverage, not a large global shift in visual attention mass.

Category logs give a concrete view: SEAL seed 2026 reduces hallucinated *skirt* from 421 to 253 and *glasses* from 129 to 37 while retaining similar major-garment coverage.

6 Ablation Study

Table 2 isolates two reward mechanisms and one attention choice. Removing the hallucination gate raises hallucinated-category count from 0.4775 to 0.5239 while leaving semantic F-score nearly unchanged, supporting high-hallucination filtering before group-relative normalization. Removing the attention floor slightly improves semantic F-score and POPE-F1 but weakens hallucination reduction. Replacing fused mid/late attention with last-layer attention lowers Attention Recall from 0.2744 to 0.2691 and grounded POPE-F1 from 0.5844 to 0.5808.

Extra-category damping is treated as an implementation detail. Under $\lambda_h = 1.10$, over-predicted categories already make S_{sem} non-positive, so the positive-only damping branch is effectively inactive. Removing it was numerically indistinguishable from Full SEAL in a single-seed check.

7 Analysis

Why token-level feedback matters. Response-level rewards blur local errors: a caption that correctly describes a shirt but adds an unsupported bow should not receive the same signal as an entirely wrong caption. SEAL assigns spatial reward only to aligned visual chunks and counts hallucinated categories explicitly, using masks as automatic supervision.

Relation to DPO initialization. SEAL is not a DPO variant and does not require a preference-tuned adapter. The DPO adapter reaches Sem.F 0.4956 and G-POPE-F1 0.5809, but its hallucinated-category count is 0.6203, higher than the no-adapter baseline. This supports the interpretation that SEAL's hallucination reduction comes from its token-level reward and GRPO update.

Configuration	Sem. F ↑	Hall. ↓	AttnRec ↑	POPE-F1 ↑	POPE-FPR ↓	POPE-Acc ↑
Matched baseline	0.4748	0.6006	0.2692	0.5635	0.0209	0.8035
Full SEAL	0.4996	**0.4775**	0.2744	0.5844	**0.0186**	0.8111
w/o hallucination gate	0.4980	0.5239	0.2728	0.5836	0.0200	0.8102
w/o attention floor	**0.5031**	0.4856	**0.2753**	**0.5888**	0.0187	**0.8125**
last-layer attention	0.4952	0.4919	0.2691	0.5808	0.0191	0.8097

Table 2: Single-seed ablations under the 3,000-step protocol. Removing the hallucination gate weakens hallucination reduction. Removing the attention floor gives back part of the hallucination gain, and last-layer attention lowers grounding and POPE metrics.

Grounded POPE behavior. Grounded POPE combines semantic yes/no judgments with an attention threshold. SEAL's F1 gain comes mainly from higher recall under low false positive rates: it increases supported positive category judgments without making the model less conservative.

Relation to attention interventions. SPIN shows that suppressing image-inattentive heads at inference can reduce hallucination. SEAL's ablations are consistent with this attention-alignment view at the token-region level: the attention floor helps hallucination reduction, and fused mid/late attention outperforms last-layer attention on grounding and POPE. The difference is procedural: SEAL learns a training-time preference signal for where attribute-bearing tokens should look, rather than applying a decoding-time head mask.

Attention recall stability. Stable average Attention Recall should not be read as a failure of token-level grounding. The reward operates on matched chunks, while the reported average spans all matched entity tokens and mask sizes. Improvements can come from suppressing unsupported mentions, resolving aliases, or avoiding poorly grounded captions without shifting the global mean. Bbox analysis can further decompose target, distractor, and background mass.

Qualitative grounding visualization. Figure 3 shows token-to-image heat maps selected when the compared token maps to a Fashionpedia category with a ground-truth box and SEAL assigns more target-box attention than the baseline. We treat these as diagnostics, not independent evidence. They illustrate local token-region differences that aggregate Attention Recall can hide.

Qualitative patterns. Manual inspection suggests that SEAL often removes unsupported accessories, replaces over-specific garment nouns with supported categories, and preserves concise attribute phrases when the head noun is grounded. Remaining failures involve ambiguous lower-body categories, small parts, and attributes without explicit annotations.

8 Conclusion

We introduced SEAL, a token-level spatial alignment framework for reducing fine-grained hallucinations in fashion VLM captions. SEAL links generated visual chunks to Fashionpedia-style masks, computes attention recall for entity tokens, and optimizes grouped online rollouts with semantic-spatial rewards and hallucination-aware filtering. On 3,600 test images, SEAL reduces hallucinated category mentions, improves semantic category coverage, and improves grounded POPE while keeping average attention recall stable. Compared with entity-level preference optimization and inference-time head suppression, SEAL moves hallucination mitigation to trainable token-region preference signals. When localization annotations are available, hallucination mitigation can move from whole-response preferences to the tokens where visual claims are made.

Limitations

SEAL requires spatial annotations during training. This makes it suitable for domains such as fashion, retail, medical imaging, and remote sensing when masks or boxes are available, but it is less directly applicable to web-scale caption corpora without localization labels. The method depends on category normalization and dependency-based chunk parsing; parser errors and ambiguous fashion aliases can affect reward assignment. The current reward focuses on category-level hallucination and does not fully validate fine-grained attributes such as material, texture, or style. Attention recall is an operational grounding signal rather than a proof of causal visual reasoning, and our aggregate results show stable rather than substantially improved average

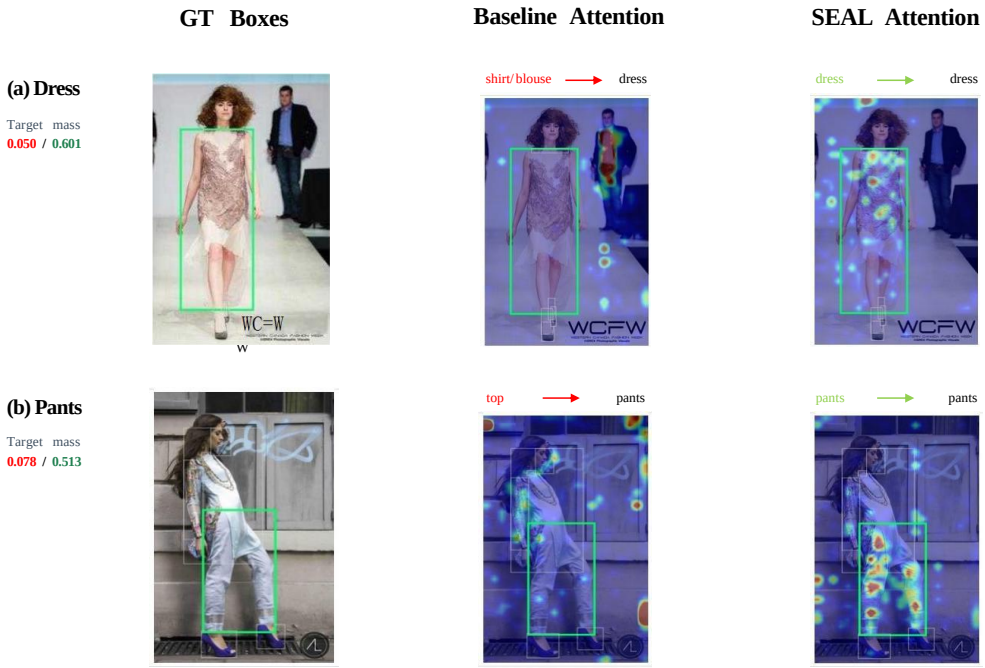

Figure 3: **Token-level heat-map comparisons selected by bbox attention mass.** Rows show diagnostic cases with target boxes, baseline heat maps, and SEAL heat maps. Green boxes mark target categories, gray boxes mark other annotated regions, and left labels report target-box attention mass. The figure is a diagnostic complement to the quantitative results.

Attention Recall. Our baseline suite is also limited: the main comparison is against the no-adapter base model under a unified grounded protocol, while the DPO adapter is evaluated only as an adapter-only initialization control and not as a full reimplementation of response-level DPO. We also do not fully re-implement decoding-time hallucination methods under the same mask-grounded evaluation. Finally, experiments use one base VLM and one fashion dataset, and extracting token-to-vision attention at native resolution makes SEAL more expensive than purely text-level preference tuning.

Ethical Considerations

Reducing hallucinations can make VLM-generated descriptions more useful, but it can also create unwarranted trust if users assume all generated details are verified. SEAL should be presented as a mitigation method, not a guarantee of correctness. Fashion datasets may encode biases in body presentation, gender expression, cultural dress, and product categories. Systems built on these data should include uncertainty reporting, user feedback channels, and additional validation before deployment in accessibility or commerce settings where incorrect descriptions can affect user decisions.

References

Chen Cai, Kim-Hui Yap, and Suchen Wang. 2024. Towards attribute-controlled fashion image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–18.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

François Gardères, Shizhe Chen, Camille-Sovanneary Gauthier, and Jean Ponce. 2025. FACap: A large-scale fashion dataset for fine-grained composed image retrieval. *arXiv preprint arXiv:2507.07135*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. 2020. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision – ECCV 2020*, pages 316–332.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*, pages 13872–13882.

Xianrui Li, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. 2021. Clothes image caption generation with attribute detection and visual attention model. *Pattern Recognition Letters*, 141:68–74.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. 2025a. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, volume 2025, pages 73796–73846.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. 2025b. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12481–12500. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. 2025. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19445–19461. Association for Computational Linguistics.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision*, pages 1–17.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2025. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19985–19995.

Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. 2025. ControlCap: Controllable region-level captioning. In *European Conference on Computer Vision*, pages 21–38.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4777–4786.