First Logit Boosting: Visual Grounding Method to Mitigate Object Hallucination in Large Vision-Language Models

Supplementary Material

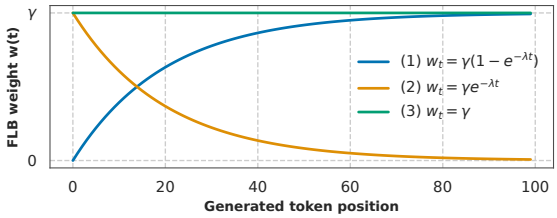


Figure 8. **FLB weight function candidates.** Comparison of three w_t formulations for weight optimization across token positions.

A. Detailed experiment result

In all experiments, we used three hyperparameters, β, γ, and λ, which were set to 0.1, 0.3, and 0.05, respectively. These values were chosen based on hyperparameter optimization experiments and were fixed throughout all evaluations. In this supplementary section, we provide optimization results beginning with those presented in Sec. B.

B. Hyperparameter optimization

B.1. Weight function and hyperparameter optimization

To find the best weight function in Eq. (5), we conducted an optimization experiment on different formulations of the weight equation using the AMBER generative task. We evaluated three settings: exponentially (1) increasing, (2) decreasing, and (3) constant weight. Fig. 8 illustrates the variation of the weight values across decoding steps in each setting. We introduce a new metric to evaluate the performance of each weight defined as:

$$object\_score = 0.5((1 - CHAIR) + Cover), \quad (8)$$

which combines the CHAIR and Cover scores, representing the proportion of hallucinated and ground truth object words in the generated sequences of the AMBER benchmark. When evaluating performance according to this metric, $w_t = \gamma(1 - e^{\lambda t})$ achieved the best performance as shown in Tab. 12. This aligns with our intuition that as generated token position increases, the long-term decay problem becomes more severe, and hence a stronger suppression of long-term decay is required. Consequently, we adopted this formula as the weight function of FLB.

Furthermore, to identify the most effective values for γ and λ, we conducted hyperparameter optimization. We used the same evaluation metric as in the weight function optimization. According to Tab. 13 and Tab. 14, both LLaVA1.5 and InstructBLIP models achieved the best performance when $\gamma = 0.3$ and $\lambda = 0.05$, which were used in the experiments. Based on this result, we adopted these values as the hyperparameter settings for all subsequent experiments.

Table 12. Best object score in Eq. (8) at each weight function setting.

w_t	(1) inc.	(2) dec.	(3) const.
object score	**72.3**	72.2	71.75

Table 13. Object score in Eq. (8) at each hyperparameter setting on LLaVA1.5.

λ / γ	0.1	0.3	0.5
0.005	71.15	71.2	71.4
0.02	71.65	71.8	72.1
0.05	71.8	**72.3**	71.7

Table 14. Object score in Eq. (8) at each weight setting on InstructBLIP.

λ / γ	0.1	0.3	0.5
0.005	72.4	72.3	72.0
0.02	72.3	72.4	72.3
0.05	72.2	**72.6**	72.1

B.2. Effect of β in adaptive plausible constraint

To understand how the magnitude of adaptive plausible constraint affects the performance of FLB, we conducted a hyperparameter optimization over β. Tab. 15 shows the results of the hyperparameter optimization for β. A larger value of β leads to more aggressive truncation, allowing only tokens with higher probabilities to be selected. We conducted this optimization on LLaVA-1.5 using the AMBER benchmark. As shown in Tab. 15, both excessively large and excessively small values of β limit the performance improvement of FLB. This indicates that overly aggressive or insufficient truncation can negatively affect the model's ability to suppress hallucinations while preserving correct information.

In addition, when the adaptive plausible constraint is not applied (i.e., $\beta = 0$), we observed unnatural token selections, such as spurious insertions of tokens like "The" in the middle of a sentence, as illustrated in Fig. 9. When the constraint is applied (e.g., $\beta = 0.1$), this erroneous behavior disappears, suggesting that the adaptive plausible constraint plays an important role in stabilizing token selection. Based

on these results, we selected 0.1 as the value of β, which achieves the highest object score and provides the best balance between reducing hallucinations and maintaining correct object coverage.

B.3. Effect of decoding strategy

In addition to the sampling-based decoding used in our main experiments, we also evaluated the performance under greedy decoding, where the token with the highest probability is selected at each step. As shown in Tab. 16, FLB consistently mitigates hallucination even when the decoding strategy is switched to greedy decoding, demonstrating that its effectiveness is robust to the choice of decoding method.

However, we also observed a slight decrease in Cover score under greedy decoding. This suggests that the deterministic nature of greedy decoding may restrict the FLB's ability to fully describe ground truth objects in mentions, indicating that additional tuning or decoding adjustments may be necessary to further improve Cover score.

C. Ratio of "The" in responses after applying FLB

Since FLB adds first logit, which assigns a high probability to "The," to the logits of subsequent tokens, we compare the proportion of "The" in the original responses and in the FLB applied responses using the 1,004 queries of the AMBER benchmark. As shown in Tab. 17, the proportion of "The" among all sentence-initial tokens rises by 21.9%p after applying FLB with $\gamma = 0.3$ and $\lambda = 0.05$, which is our optimized setting. This raises concerns that FLB might reduce sentence diversity or lead to more monotonous phrasing. However, as demonstrated in Sec. 6.2, FLB does not degrade the overall quality of the generated sentences.

To further examine this effect, we conducted an additional GPT-4V based evaluation comparing baseline LLaVA-1.5 and FLB outputs across different γ values. Because γ determines the maximum strength of the weighting function, we evaluated FLB with $\gamma \in \{0.1, 0.3, 0.5, 0.7\}$, where $\gamma = 0.3$ corresponds to our optimized setting. The prompt format used for this evaluation is shown in Fig. 12. As shown in Tab. 17, FLB consistently achieves high GPT-4V evaluation scores across all tested γ values. Notably, FLB also outperforms the baseline in the newly introduced *Expression Diversity* metric, demonstrating that the increase in "The" does not negatively impact stylistic variety or expression richness. If an application requires reducing the proportion of "The," this can be easily controlled by lowering the value of γ. Overall, these results indicate that FLB does not harm linguistic diversity or sentence quality despite influencing the distribution of sentence-initial tokens.

D. Other backbone models

In the main paper, we evaluate FLB using LLaVA1.5 and InstructBLIP as LVLM backbones. To further assess its generalization across different models, we additionally conduct experiments on mPLUG-Owl2 using the AMBER benchmark. All hyperparameters are kept identical to those used in the main experiments. As shown in Tab. 18, consistent with the results in the main paper, FLB outperforms both the baseline and VCD across hallucination-related metrics. These results further demonstrate the robustness and generalizability of FLB across different LVLM backbones.

E. Results on discriminative task

FLB is designed to mitigate hallucination that becomes more pronounced in longer generated sequences. Therefore, its impact is expected to be limited in discriminative tasks with short outputs. We evaluate FLB on two representative object hallucination benchmarks, POPE [17] and MME [9], using LLaVA1.5 as the backbone. We compare FLB with baseline decoding and a β-only setting, which applies only the adaptive plausibility constraint without logit boosting. Other experimental settings are identical to those in the main paper. As shown in Tab. 19, applying FLB yields the same performance as β-only decoding, indicating minimal effect in this setting. This suggests that FLB is particularly beneficial for generative settings with longer outputs.

F. Prompt of GPT-4V evaluation

To assess the quality of responses after applying FLB, we use GPT-4V to evaluate both accuracy and level of detail in the generated outputs. The specific prompt format used for this evaluation is shown in Fig. 10, and an example of the resulting evaluation is provided in Fig. 11. For the analysis in Sec. C, we additionally include the corresponding prompt format and evaluation example in Fig. 12 and Fig. 13, respectively.

Table 15. Hyperparameter optimization on β on AMBER benchmark.

β	CHAIR↓	Cover↑	Hal↓	Cog ↓	Object score↑
0	7.8 ($\pm$ 0.17)	50.2 ($\pm$ 0.41)	39.7 ($\pm$ 1.58)	2.9 ($\pm$ 0.08)	71.2 ($\pm$ 0.12)
0.01	7.5 ($\pm$ 0.25)	50.3 ($\pm$ 0.5)	37.5 ($\pm$ 0.64)	3.3 ($\pm$ 0.17)	71.4 ($\pm$ 0.13)
0.05	6.8 ($\pm$ 0.25)	50.4 ($\pm$ 0.19)	33.3 ($\pm$ 0.70)	2.8 ($\pm$ 0.17)	71.8 ($\pm$ 0.21)
0.1	6.1 ($\pm$ 0.37)	50.4 ($\pm$ 0.22)	31.4 ($\pm$ 1.20)	2.7 ($\pm$ 0.24)	**72.1** ($\pm$ 0.12)
0.2	6.6 ($\pm$ 0.21)	50.2 ($\pm$ 0.21)	31.9 ($\pm$ 1.07)	2.9 ($\pm$ 0.29)	71.8 ($\pm$ 0.18)
0.4	6.5 ($\pm$ 0.08)	50.5 ($\pm$ 0.17)	30.4 ($\pm$ 0.52)	3.3 ($\pm$ 0.22)	72.0 ($\pm$ 0.10)

Table 16. Performance comparison of LLaVA1.5 and greedy sampling on AMBER generative tasks. The **highest** scores are marked in **bold**.

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
Baseline	7.1	50.5	32.4	3.8
VCD	8.2	52.2	38.0	4.0
ICD	6.4	51.0	30.6	3.2
M3ID	7.0	**55.8**	37.5	2.8
FLB (Ours)	**4.9**	48.8	**25.2**	**2.3**

Table 17. Comparison of the frequency of "The" appearing as the sentence-initial token with and without applying FLB.

Method	Proportion of "The" among sentence-initial tokens	CHAIR↓	Cover↑	Accuracy↑	Detailedness↑	Expression Diversity↑
Baseline	67.4%	11.9	49.6	4.83	4.65	5.36
FLB ($\gamma = 0.1$)	83.1%	7.5	51.1	6.21	5.80	6.23
FLB ($\gamma = 0.3$)	89.4%	5.7	50.3	6.47	5.91	6.34
FLB ($\gamma = 0.5$)	91.3%	5.8	49.2	6.26	5.77	6.10
FLB ($\gamma = 0.7$)	92.1%	5.3	48.7	6.29	5.85	6.22

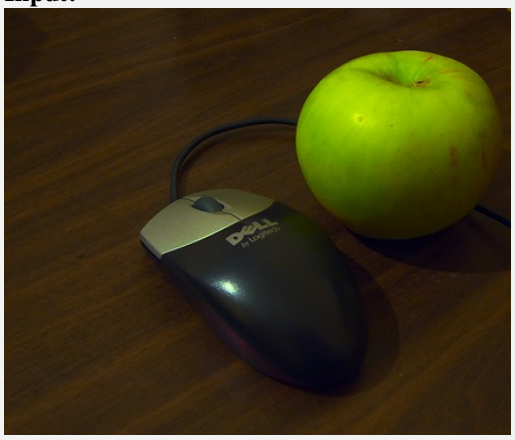
Table 18. Performance comparison of mPLUGOwl2 on AMBER. The **highest** scores are marked in **bold**.

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
Baseline	12.5	51.2	50.8	5.2
VCD	11.3	**53.1**	46.4	5.5
FLB	**7.1**	51.6	**33.0**	**2.9**

Table 19. Performance comparison of LLaVA1.5 on discriminative tasks.

Benchmark	POPE						MME
Dataset	Random		Popular		Adversarial		
Metric	Acc.	F1	Acc.	F1	Acc.	F1	MME score
Baseline	0.829	0.808	0.811	0.792	0.786	0.771	114.20
β-only	0.846	0.826	0.827	0.809	0.801	0.786	115.88
FLB (Ours)	0.846	0.826	0.827	0.809	0.801	0.786	115.88

Input:



[Prompt]
Please describe this image in detail.

[First Logit Boosting with $\beta = 0$]
The scene features an Apple computer mouse and a green apple on a wooden table, with the mouse placed next to the apple on the right side. The apple is Apple's trademark in the image. The flat in the background indicates the mantle of the scene. The table's surface here is reflective and gray. The Sub In The Over The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The The ...

[First Logit Boosting with $\beta = 0.1$]
The image features a black and silver computer mouse placed next to a shiny green apple. The mouse is positioned on the left side of the table, while the apple is located on the right side, both sitting on the same wooden surface. This setting suggests a casual and comfortable workspace.

Figure 9. Example outputs from LLaVA-1.5 with FLB applied using $\beta = 0$ and $\beta = 0.1$, illustrating that removing the adaptive plausible constraint ($\beta = 0$) leads to unnatural outputs.

Description:
You are required to score the performance of two AI assistants in describing a given image.
Pay special attention to hallucination — parts of descriptions inconsistent with the image content
(e.g., objects not present, wrong counts, colors, or positions).

Rate each response on a scale of 1 to 10 for:
1. Accuracy — whether the response matches the image content.
2. Detailedness — whether the response is rich in necessary details (but not hallucinated).

Input format:

[Assistant 1]
{Response 1}
[End of Assistant 1]

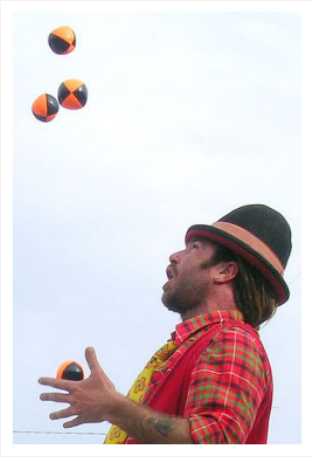
[Assistant 2]
{Response 2}
[End of Assistant 2]

Output format:

Accuracy: [score1] [score2]
Reason:
Detailedness: [score1] [score2]
Reason:

Figure 10. **Prompt template for the GPT-4V evaluation**. Assistant 1 and Assistant 2 indicate the response of baseline method and FLB method, respectively. {Response 1} is the baseline response, and {Response 2} is the response after applying FLB.

Input:



[Assistant 1]
In the image, a man dressed in a top hat and scarf is balancing four orange balls on his fingers.
He appears to be performing a daring and skillful act of keeping the balls in place while holding only one finger.
The man seems to be confident in his presentation, capturing the attention of those watching.
The four balls are placed at different positions on his fingers, with each ball spaced apart from each other.
The arrangement challenges his dexterity and hand-eye coordination, highlighting his abilities as an entertainer.
[End of Assistant 1]

[Assistant 2]
The image features a man wearing a top hat and a red and yellow plaid shirt, standing in the middle of the scene.
He is in the midst of juggling three orange and black balls in the air, showcasing his skill and coordination.
The man is the main focus of the image, and the scene captures the excitement and concentration of the juggling act.
[End of Assistant 2]

Output:

Accuracy: 1 9
- Assistant 1: Incorrectly describes the man as balancing the balls on his fingers, which is not the case. He is juggling four balls.
- Assistant 2: Accurately describes the man juggling, but mistakenly mentions three balls when there are actually four.

Detailedness: 3 7
- Assistant 1: Provides incorrect information with unnecessary details about balancing, which reduces relevance.
- Assistant 2: Offers relevant details about the attire and the juggling act, despite the minor error in the number of balls.

Figure 11. **Example of a GPT-4V evaluation**. Assistant 1 and Assistant 2 indicates the response of baseline methd and FLB method, respectively. The response generated with FLB obtains a higher evaluation score.

Description:
You are required to score the performance of two AI assistants in describing a given image.
Pay special attention to hallucination — parts of descriptions inconsistent with the image content
(e.g., objects not present, wrong counts, colors, or positions).

Rate each response on a scale of 1 to 10 for:
1. Accuracy — whether the response matches the image content.
2. Detailedness — whether the response is rich in necessary details (but not hallucinated).
3. Expression Diversity — whether the wording is varied and natural, avoiding repetitive or formulaic phrasing while remaining clear.

Input format:

[Assistant 1]
{Response 1}
[End of Assistant 1]

[Assistant 2]
{Response 2}
[End of Assistant 2]

[Assistant 3]
{Response 3}
[End of Assistant 3]

[Assistant 4]
{Response 4}
[End of Assistant 4]

[Assistant 5]
{Response 5}
[End of Assistant 5]

Output format:

Accuracy: [score1] [score2] [score3] [score4] [score5]
Reason:
Detailedness: [score1] [score2] [score3] [score4] [score5]
Reason:
Expression Diversity: [score1] [score2] [score3] [score4] [score5]
Reason:

Figure 12. **Prompt template for the additional GPT-4V evaluation**. Assistant 1 indicates the baseline response, while Assistant 2–5 correspond to FLB responses with $\gamma \in 0.1, 0.3, 0.5, 0.7$, respectively. {Response 1} denotes the baseline response, while {Response 2}, {Response 3}, {Response 4}, and {Response 5} correspond to the responses generated with FLB under different γ settings.

Input:

[Assistant 1]
The image depicts a serene beach scene with benches and a sunset as the main focal point.
Two wooden benches face the ocean, one closer to the left side of the frame and the other towards the right.
The sun is setting, casting a warm glow on the scene.
In addition to the benches, there are palm trees visible in the background, making the location seem more relaxing and tropical.
They frame the view while a person can be seen standing a bit further back on the beach, gazing towards the ocean and enjoying the picturesque atmosphere.
[End of Assistant 1]

[Assistant 2]
The image features a picturesque beach scene with the ocean in the background.
A bench is placed near the shoreline, offering a view of the water.
Two benches are visible, one positioned closer to the left side of the frame and the other closer to the right side.
They are arranged to face the water, creating a comfortable spot for beachgoers to relax and enjoy the view.
The benches appear to be empty in this scene.
[End of Assistant 2]

[Assistant 3]
The image features a beach with a bench facing the ocean, creating a serene atmosphere.
The bench is positioned under the shade of a palm tree, providing a comfortable spot for visitors to relax and enjoy the beautiful view.
The ocean in the background is calm, with waves gently lapping on the shore.
The setting sun adds a touch of warmth and tranquility to the scene, making it a perfect location for a peaceful day by the sea.
[End of Assistant 3]

[Assistant 4]
The image showcases a serene beach scene at sunset with a beautiful ocean view.
In the foreground, a bench is situated near the water, overlooking the ocean and the sunset.

The bench is located on a wooden walkway that extends along the beach, providing a picturesque view of the ocean and the sun setting in the background.
The bench is the focal point of the scene, inviting visitors to sit and enjoy the captivating beauty of the sunset and the tranquility of the beach.
[End of Assistant 4]

[Assistant 5]
The image depicts a serene beach scene with a bench situated near the water's edge, overlooking the ocean.
The bench is placed on the beach and is located in the center of the image.
The beach is lined with a few trees, and the sun is shining on the scene, creating a peaceful and inviting atmosphere.
The bench provides a comfortable spot for visitors to sit and enjoy the ocean view.
[End of Assistant 5]

Output:

Accuracy: 2 7 9 5 4
- Assistant 1: Contains inaccuracies, like stating there are two benches and mentioning a person, which is not visible.
- Assistant 2: Correctly identifies one bench but incorrectly describes two benches when only one is visible.
- Assistant 3: Accurately describes a single bench and palm trees but doesn't mention any incorrect elements.
- Assistant 4: Incorrectly describes a wooden walkway and implies proximity to water, which isn't accurate.
- Assistant 5: Correct in describing the bench and ocean, but inaccurately mentions trees lining the beach.

Detailedness: 3 6 8 6 5
- Assistant 1: Includes imaginary elements (second bench, person), compromising detail accuracy.
- Assistant 2: Provides a moderate level of detail but errs in number of benches.
- Assistant 3: Offers rich and correct details about the scene and environment.
- Assistant 4: Attempts detailed storytelling but uses elements not present in the image (wooden walkway).
- Assistant 5: Lacks some specific details and includes inaccuracies (tree lineup).

Expression Diversity: 4 6 8 7 6
- Assistant 1: Uses some varied language but hallucinates too much, reducing quality.
- Assistant 2: Contains some varied expressions but redundant in mistaken mention of two benches.
- Assistant 3: Well-varied expressions with descriptive phrases effectively painted a vivid scene.
- Assistant 4: Although somewhat expressive, incorrect details detracted from overall diversity.
- Assistant 5: Fair use of diverse expressions, lessened by inclusion of inaccurate content.

Figure 13. **Example of an additional GPT-4V evaluation**. Responses generated with FLB under different γ values receive different evaluation scores. Assistant 1 shows the baseline response, while Assistant 2–5 correspond to FLB outputs with $\gamma \in 0.1, 0.3, 0.5, 0.7$, respectively. **Bold** highlights occurrence of "The". After applying FLB, the frequency of "The" increases, but the overall response quality does not decrease.