Thanks again for your response and highly constructive suggestions! We will provide further clarification on the remaining issues:

Response to New Question (1):

There seems to be some misunderstanding regarding the experimental setup. First, we would like to clarify that in 1gd8.pdf, the term "aligned image token" is synonymous with "projected image token." Additionally, the expressions "pre-aligned image tokens" and "pre-alignment visual tokens" are both used to describe the image tokens before projection. We apologize for any confusion resulting from our imprecise terminology and have made the necessary corrections in 1gd8.pdf.

Regarding Experiment (ex1): We compared the distributions of image tokens before projection (outputs from the CLIP encoder, represented by blue points in Figure ex1), image tokens after projection (outputs from the MLP projection layer of LLaVA-Next, represented by red points in Figure ex1), and text tokens (generated by the tokenizer and embedding layer of the LLM, represented by green points in Figure ex1) in a 2D space after dimensionality reduction. Our findings clearly indicate that the projected image tokens exhibit higher distributional similarity to the text tokens.

We understand that the reviewer may interpret such distributional similarity as an indication of some degree of semantic similarity, given the inherent ambiguity in defining "semantics." However, in this paper, we prefer to interpret the results of Experiment (ex1) solely as evidence that the projection layer establishes alignment between images and text at the distributional level. In our understanding, assessing the semantics of an image necessitates a thorough consideration of the interrelationships among all image tokens rather than evaluating individual image tokens in isolation.

Similarly, to more thoroughly evaluate the semantics represented by each image token within its respective image, we introduce the concept of the Attribute Vector for image tokens in Experiment (ex2). This is defined as a collection of each image token's cosine similarity to all other image tokens derived from the same image. For a formal mathematical definition, please refer to 1gd8.pdf. We then assess the influence of the projection layer on image semantics by calculating the cosine similarity between the Attribute Vectors of each image token before and after the projection. As depicted by the blue curve in Figure ex2, all Attribute Vectors of the projected image tokens exhibit high similarity to their counterpart before projection. This observation suggests that the overall semantics of the image remain largely consistent. Consequently, the projection layer fails to achieve effective semantic alignment, as successful semantic alignment should produce significant alterations in the Attribute Vectors.

We hope that the explanations provided above clarify our understanding of the projection layer's role for the reviewer. We will refine the description in line 94 to eliminate any potential misunderstandings and ensure that the discussions mentioned are incorporated into the revised manuscript. Nonetheless, even if a consensus cannot be reached with the reviewer regarding these perspectives on the projection layer's function, it does not diminish the effectiveness of our method in enhancing image-text alignment and improving model performance. In the following sections, in accordance with

the reviewer's suggestions, we will conduct additional experiments to further validate the efficacy of our approach.

Response to New Question (2):

As noted by the reviewer, when comparing (a) the cosine similarity between projected image tokens and image tokens before projection with (b) the cosine similarity between projected image tokens and text tokens, it is indeed expected that (b) would yield significantly higher values due to the stronger distributional consistency. However, in Experiment (ex2), we employ the Attribute Vector of image tokens to represent their semantics within the image. The blue curve in Figure ex2 illustrates the cosine similarity between the Attribute Vectors of each image token both before and after projection, while the orange curve represents the maximum cosine similarity between the projected Attribute Vectors and each text token in their corresponding image caption.

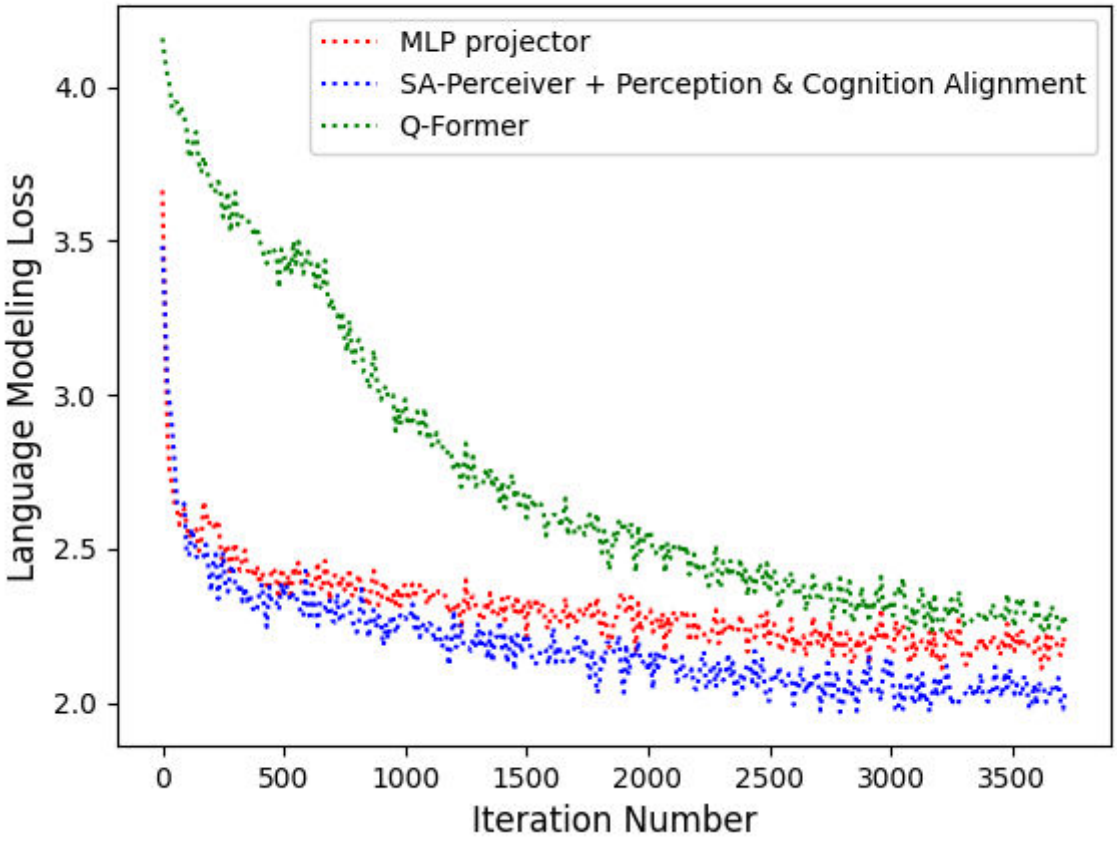
Supplementary experiments to substantiate the enhanced alignment during the training process：

We illustrate that our method attains faster and better alignment during training when compared to the standard MLP projector and Q-former. This conclusion is backed by our analysis of performance variations across three aspects throughout the training process: (a) the convergence of language modeling loss, (b) two quantitative assessments of image-text alignment, and (c) benchmark metrics.

(a) The convergence of language modeling loss

We visualized the language modeling loss of three variants during the image-text alignment pretraining (Stage 1 of VLSA) on the image captioning dataset. As illustrated in Figure ex18, our method exhibits faster convergence and achieves better final loss than the other two variants. Given that the performance of image captioning tasks directly depends on the degree of alignment, it demonstrates that our method enhances both the speed and effectiveness of alignment.

(Figure ex18) Comparisons on the convergence of language modeling loss.



(b) Two quantitative assessments of image-text alignment

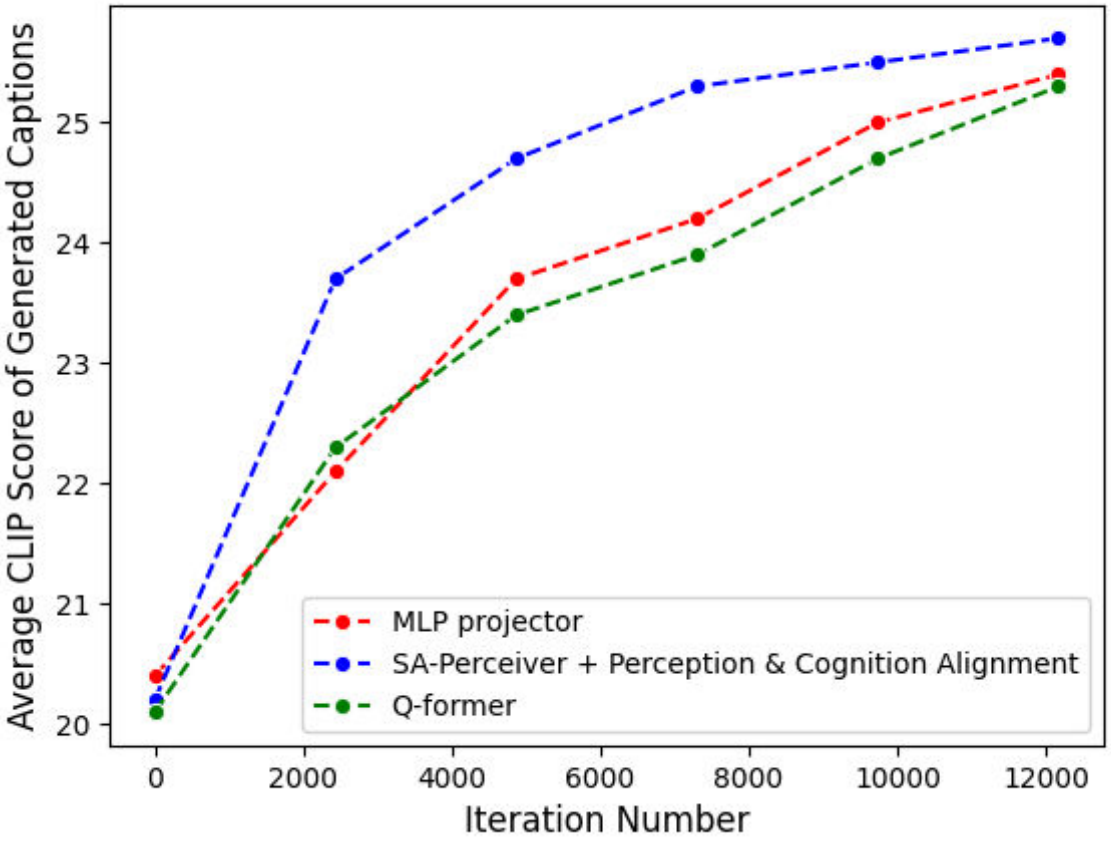
We have implemented two metrics to quantitatively evaluate the effectiveness of alignment. Specifically, we trained three variants using our 980k visual instruction tuning dataset, saving a checkpoint every 2435 iterations (noting that one epoch consists of 12,176 iterations). We randomly selected 1,000 (or 300) images for each checkpoint and prompted the model to generate detailed captions. We then assessed the quality of these generated captions using the following two metrics, with higher values indicating better alignment:

(1) **CLIP Score**: For each checkpoint, we employed the CLIP model to evaluate the correlation logits between the 1,000 input images and their corresponding generated captions. The final score was determined by calculating the average of these logits.

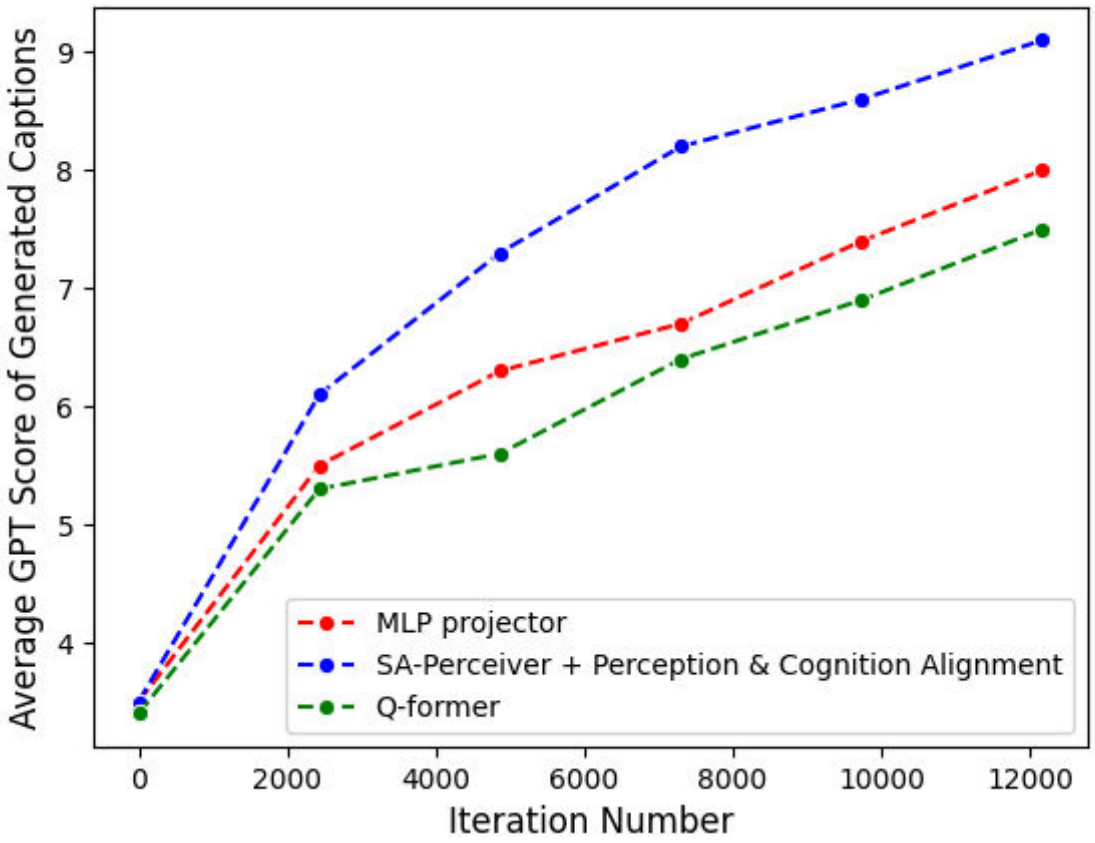
(2) **GPT Score**: For each checkpoint, we created a series of prompts to guide the GPT-4o model in evaluating the quality of captions generated from 300 images, using these images as clues. Ratings were provided on a scale from 1 to 10, and the average of these ratings was recorded as the GPT Score.

The specific prompts used for generating detailed captions at each checkpoint, as well as those instructing GPT-4o to score the captions, will be included in the revised manuscript.

(Figure ex19) Comparisons on CLIP Score.



(Figure ex20) Comparisons on GPT Score.



The results illustrated in Figures ex19 and ex20 clearly demonstrate that our method enhances the speed of achieving alignment, ultimately leading to superior alignment performance.

(c) Benchmark metrics

Alongside the findings in section (b), we assess the performance of the five checkpoints from the experiments detailed in (b) using the GQA benchmark. This evaluation enables us to validate the impact of our proposed method on modality alignment by analyzing the variations in model performance throughout the training process.

(Figure ex21) Variations in model performance during training