We thank all the reviewers for their insightful and encouraging comments. We confirm our promise to release all the code and pre-trained model checkpoints upon paper acceptance.

To Reviewer 1, 3, and 4 **Q1**: ***How do perturbations in embedding space compare to those in pixel/token space?***

A1: Most previous work studies adding perturbations in the pixel space for image classification task, and the general observation is that robustness is often at odds with generalization, *i.e.*, adversarial training hurts the performance on clean data. When dealing with V+L tasks, we observe that adding perturbations in the embedding space actually boosts the performance on clean data. We hope this observation can inspire future work on adding feature perturbations for adversarial training in different tasks.

Unlike pixels, tokens are discrete in nature. It still remains a challenging task to effectively craft adversarial examples in the token space without changing the semantic meaning of the original text. Since we only care about the *end results* of adversarial training on downstream tasks, adding embedding perturbations is a natural way to circumvent this obstacle.

Our embedding-based adversary is generic and flexible, as it can make arbitrary manipulations on embeddings, which is not possible if in the pixel/token space.

Q2: ***What happens if adversarial perturbations are simultaneously added to both image and text domains?***

A2: Empirically, we observe that adding perturbations to both modalities simultaneously often reaches slightly worse performance than adding perturbations alternatively. This may be due to the fact that the injected noise is too much when perturbations are added simultaneously. Results are detailed in Table 1. More results on $\text{VILLA}_{\text{LARGE}}$ will also be included in the final version.

Method	VQA	VCR (val)		
	test-dev	Q→A	QA→R	Q→AR
$\text{VILLA}_{\text{BASE}}$ (simul.)	73.15	75.50	78.60	59.56
$\text{VILLA}_{\text{BASE}}$ (alter.)	**73.59**	**75.54**	**78.78**	**59.75**

Table 1: Adding perturbations simultaneously vs. alternatively.

Q3: ***Do the adversarial perturbations make the model more robust to adversarial attacks and paraphrases?***

A3: In order to test adversarial robustness, we need to perform adversarial attacks to existing V+L models. This V+L attack problem is largely unexplored in the literature. For example, how to reliably back-propagate the gradients from the multimodal Transformer to the CNN backbone to generate image adversaries is non-trivial. How to craft textual adversaries that align with the visual context is also challenging. In this paper, we mainly focus on improving model's generalization performance on clean data, leaving a more thorough investigation of adversarial attack and robustness as important future work. We are actively working on this.

Method	VQA Acc.	
	Ori.	Rep.
BAN	64.97	55.87
BAN + CC	65.87	56.59
$\text{UNITER}_{\text{BASE}}$	70.35	64.56
$\text{VILLA}_{\text{BASE}}$	**71.27**	**65.35**

Table 2: Results on VQA-Rephrasings.

To further address the reviewers' question, we have conducted additional experiments on the VQA-Rephrasings dataset to test the robustness of existing V+L models to paraphrases. For fair comparison, we have re-trained both UNITER and VILLA on the VQA training set only. Results are summarized in Table 2, where Ori. and Rep. denote the test set with original questions and their rephrasings. UNITER has already lifted up the performance by a large margin, and VILLA facilitates further performance boost.

To Reviewer 4 **Q4**: ***Unfair comparison with UNITER?***

A4: We would like to emphasize that the comparison with UNITER is a fair setting. As shown in L198-201 and Fig. 3(a), we carefully designed the experiments to make sure both UNITER and VILLA use the same number of optimization steps.

Q5: ***The training seems a little bit tricky to me.***

A5: We confirm that we use exactly the same model configuration as UNITER. No hidden trick was used for training VILLA, and our results can be easily reproduced. The experimental settings in the original UNITER paper are not updated. For example, in their arXiv v1 version, UNITER-large achieved 73.82 on VQA test-std. This has been lifted to 74.02 with better optimized hyper-parameters. We use the authors' most recent hyper-parameter setting to perform experiments on VILLA. Also, in the official UNITER repo, the training batch size is indeed set to 5120 (`https://github.com/ChenRocks/UNITER/blob/master/config/train-vqa-base-4gpu.json`).

To Reviewer 5 **Q6**: ***One slight concern is that visualizing only one example is not solid enough.***

A6: We agree that visualizing only one example is not enough, and have provided Table 4 as a more systematic evaluation, in which the learned alignment has been carefully examined on the dataset level, rather than cherry-picking one single example. Also, we have provided another two examples in Appendix as well. We will clarify this.