We thank all reviewers for their insightful and constructive comments. We'll release all code and data.

(R1) GRU with pre-trained embeddings and BERT. We supplement extra experiments using GRU with pre-trained GloVe embeddings and BERT in Table A, as an extension to the Table 4 in the paper. Pre-trained embeddings from large language corpus indeed help baselines better predict the *synonym* relation between unseen concepts, but our VCML still outperforms them. We also observed that pretrained text embeddings do not improve the GRU baseline on the *instance-of* tests, whose weights are pretrained by language modeling on training questions.

Table A: Visual grounding helps predict metaconcepts between unseen concept pairs (evaluated in metaconcept QA accuracy).

CLEVR	Q. Type	GRU	GRU (GloVe)	BERT	VCML
Synonym	50.0	55.1	56.6	80.7	**86.3**
InstanceOf	25.0	58.9	42.5	44.6	**72.2**

GQA	Q. Type	GRU	GRU (GloVe)	BERT	VCML
Synonym	50.0	53.4	71.5	76.0	**94.5**
InstanceOf	12.5	**26.3**	14.6	14.9	19.9

(R1, R3) The 'instance-of' metaconcept in de-biasing (Sec. 4.3.2, Table 3). To verify that the *instance-of* metaconcept helps de-biasing, we perform an ablation study on VCML with and without *instance-of*. On CLEVR, VCML performs better with the instance-of metaconcept (55.6% *vs*. 43.3%). Meanwhile, we speculate that models on GQA de-biasing perform similarly since GQA concept categories are not well reflected in vision, especially for categories defined by functionality instead of appearance, such as vehicles. To evaluate VCML on interpreting metaconcepts better associated with visual appearance in natural images, we have supplemented results on CUB* (see below L26–L31).

(R1) Instance-of metaconcept generalization in Table 4. We agree with the review on that linguistic information helps more on the instance-of metaconcept in GQA, and the supplementary results in Table A also support this. Since VCML learns concept embeddings completely from visual data, it performs worse than the linguistic baselines.

(R1) Zero-shot compositional visual reasoning. Thanks for the suggestion. In this work we use manually generated datasets for two reasons. First, they enable controlled and diverse experiments such as de-biasing. Second, the extra metaconcept questions are essential: the de-biasing generalization will be otherwise ill-formed.

(R1) Technical details. The word embeddings in GRU-CNN is pretrained on the question set, same as in GRU. The semantic parsers used by NS-CL and our VCML are identical, both trained on question-program pairs. The fact that VCML outperforms the metaconcept-agnostic NS-CL suggests the importance of metaconcept learning.

(R2) Unsupervised discovery of metaconcepts. We agree that the unsupervised discovery of metaconcepts is a promising direction†. The main contribution of this paper is to incorporate metaconcepts into visual concept learning, in the form of supplementary question-answer pairs. The extra information enables learning from less and even biased data, which is ill-formed if no extra supervision (e.g., human-designed metaconcept-related questions) is present.

(R2, R3) Generalizing to new concepts and metaconcepts. We also apply VCML on the CUB dataset* to learn the *hypernym* metaconcept from visual data. Data are generated from the biological taxonomy of birds. We train different models on a partial set of the taxonomy, by providing the hypernym relationship of ∼74K pairs between 273 concepts, and evaluate them on ∼9K pairs between 93 concepts in the held-out set. Shown in Table B, our model outperforms both visual and linguistic baselines, which supports the generality of our VCML.

Table B: Metaconcept generalization evaluation of *hypernym* on CUB.

Model	Acc. (%)
Q. Type	50.0
GRU (Lang.)	74.3
BERT	73.1
GRU-CNN	76.7
NS-CL	54.3
VCML	**85.5**

(R3) Originality of metaconcept learning. This paper introduces a new approach to simultaneous learning of visual concepts and metaconcepts. Moreover, its applications such as de-biasing with metaconcepts have never been addressed before. Existing research on metaconcepts have been mostly restricted to linguistic domains. Two related topics are visual compositional learning and knowledge graph completion, both discussed in Section 2.

(R3) Generality of metaconcept operators. Our design of metaconcept operators is inspired by TransE‡, a framework for linguistic knowledge graph embeddings. It is a general operator for metaconcepts/relations between concepts.

(R3) Connection to Platanios et al. Thanks for suggesting the related work, which we will cite and discuss. In VCML, the projection embedding transforms the object embedding into a subspace, in which a cosine similarity is then computed to classify the object. This differs from the projection network of Platanios et al. which transforms a language embedding into a neural network parameter for encoding input sentences.

(R3) Application of concept embeddings to downstream tasks. We supplement extra results on the CLEVR visual reasoning challenge (Visual QA) and a referential expression task. The task of referential expression is to select out a specific object from a scene given a description (e.g., the red cube). We compare VCML with and without metaconcept information using the QA accuracy for visual reasoning and Recall@1 for referential expressions. Table C suggests that the metaconcept information significantly improves visual concept learning in low resource settings, using only 10K or even 1K visually grounded questions.

Table C: Evaluation of the learned concept embeddings on visual reasoning (in QA accuracy) and referential expression interpretation (in Recall@1).

	#Train	w/.	w/o.
Visual QA	10K	74.8	73.6
	1K	65.7	61.0
Ref. Expr.	10K	71.2	70.2
	1K	55.0	51.2

*Wah et al. The Caltech-UCSD Birds-200-2011 Dataset. TechReport, Caltech, 2011.

†Kemp and Tenenbaum. The Discovery of Structural Form. PNAS 2008

‡Bordes et al. Translating Embeddings for Modeling Multi-Relational Data. NeurIPS 2013