We thank all reviewers for finding our paper *novel*, *interesting*, and with *strong* performance (**R1**, **R2**, **R3**, **R4**). We apologize for missing the details of causal graph and some related references (**R2**). We will address all the concerns.

R1-Q1 Cause of Performance Drop. You are correct. We will rephrase to highlight the blame is on classifier.

R1-Q2 Coexistence of Causal Factors. In fact, the underlying assumption of using feature channels is that they are *Independent Mechanisms* (IM) [A], which generate X and Y (see details in **R2-Q6**) and there could be confounders across the subsets of channels. Fortunately, those confounders have no direct causal links to X and Y [B] and thus adjusting the channels can block the effect from the confounders (Markov factorization).

R1-Q3 Feature Selector. Feature selector c is a pre-defined mask for selecting feature channels (see line 193-195).

R1-Q4 More Convincing Experiments. Actually, we have provided the results in section A.6 of the supplementary material, where feature-wise and class-wise adjustment gain similar improvements on average.

R1-Q5 Other Causality-Based Strategies. The analysis in our paper can include Rubin's potential outcome framework, *e.g.*, using propensity score as another deconfounding approach, besides class-/feature-wise adjustment.

R2-Q1 Figure 2b. We want to clarify that in the backdoor adjustment, $do(X)$ does not make the "grass" feature disappear, *i.e.*, it is still used as a predictive signal but with its contribution adjusted by $P(\text{"}grass\text{"})$.

R2-Q2 Figure 2a. We will add the error bars. The dissimilarity is measured by query hardness defined in line 265.

R2-Q3 Formal Definition of Causal Graph. Sorry for the clarity issue. We omitted this as we intended to only offer a high-level concept for readers with CV/ML background. We will follow your suggestion to provide a formal and well-defined SCM in revision. Specifically, we model FSL as a Structural Causal Model $\mathcal{M}$ that consists of a collection $\mathcal{M} = (f_X, f_C, f_Y)$ of structural assignments $X := f_X(I, D), C := f_C(X, D), Y := f_Y(X, C)$. D is defined as the stratum set of pre-trained knowledge $D = \{d_1, \ldots, d_n\}$ learnt from large dataset $\mathcal{D}$, where d_i is a subset of feature channels in feature-wise adjustment (*FT*) or a pre-training class in class-wise adjustment (*CL*). The sample ID $I = \{1, \ldots, |\mathcal{S}|\}$ in training and $I = \{1, \ldots, |\mathcal{Q}|\}$ in testing, where $\mathcal{S}$ is support set and $\mathcal{Q}$ is query set ($\mathcal{S}, \mathcal{Q} \cap \mathcal{D} = \emptyset$). f_X uses deep network to obtain feature X for the image with ID I . f_C projects X on a stratum of knowledge $D = d_i$ to get the image-specific C representation (see line 193, 207). The classification logits Y is given by f_Y. We are sorry for the confusion on $X \to Y$ and we will highlight in revision that $X \to C \to Y$ is sufficient in *FT*, while in *CL*, $X \to Y$ is necessary as the class-based C might be an incomplete representation of X. The objective of FSL is $P(Y|do(X))$ and the parameters of f_Y is learnt in training. Our model is generally applicable to fine-tuning and meta-learning, where they differ in the parameterization of f_Y: θ in fine-tuning (see line 79) and an *additional* set of parameters ϕ in meta-learning (see line 83).

R2-Q4 Explicit Form of $[\mathbf{x}]_c$. $[\mathbf{x}]_c = \{x_i\}_{i \in c}$, where x_i is the value of feature vector $\mathbf{x}$ at i-th position.

R2-Q5 Backward Edge $X \to I$. For example, in the 1-shot extreme case of FSL, there is a 1-to-1 mapping between sample ID I and feature X, denoted as the bi-directed edge $I \leftrightarrow X$ in Figure 4(b). However, in MSL where training data is abundant, $X \to I$ is cut off because tracing the ID given feature X is practically impossible (see line 148).

R2-Q6 Causal or Anti-Causal. We will discuss your suggested related work as follows, *i.e.*, why do we adopt $X \to Y$ (causal) not $Y \to X$ (anti-causal) in FSL? Anti-causal learning [C] is based on the Independent Mechanisms (IM) or causal generative factors assumption [A][B], which states that the observations are generated from IM. Therefore, when label Y is simply disentangled enough to be IM (*e.g.*, 10 digits in MNIST [C]), $Y \to X$ establishes. However, in our FSL, when the label is much more complex, *e.g.*, the ImageNet labels "dog" and "cat" are semantically entangled such as "soft fur", we should consider the causal prediction $X \to Y$ as it is essentially a reasoning process , *e.g.*, there are recent empirical justifications of $X \to Y$ in complex CV tasks [D]). In this way, the IM becomes the D in our method, where D generates visual features X and $D \to Y$ emulates our human's naming process, *e.g.*, using "small, fur, four-legged" to name "meerkat". Note that, although each piece of knowledge in D is also complex, CNN has "engineered" them to be disentangled, such as the feature channels (feature-wise adj.) and softmax class responses (class-wise adj.). We will also explore the combination of anti-causal and causal predictions in future work, *e.g.*, following [E] when Y is not perfectly disentangled or entangled.

R3-Q1 Backbone Choice. Thanks, we will validate on weaker backbones following your suggestion.

R3-Q2 Why meta-learning suffers less? We totally agree with your opinion that meta-learning suffers less from the deficiency and actually we validated this intuition in our experiments (see line 283, 299). We have also discussed the potential reason that meta-learning is essentially a form of intervention (see line 284).

R3-Q3 Negative Transfer. We will revise and add empirical comparisons to negative transfer literature.

R3-Q4 Figure 2. Sorry for the confusion. Figure 2 targets fine-tuning and we will revise to highlight this.

R4-Q1 Preliminaries. Thanks for the suggestion. We will provide a more detailed introduction to preliminaries during revision, such as the formal definition of the casual graph in **R2-Q3**.

R4-Q2 Design Choices. The design choices for feature-/class-wise adjustment reflect the motivation discussed in line 179-185: *e.g.*, in class-wise adjustment, $g(\mathbf{x}, d)$ is the distilled pre-trained knowledge.

R4-Q3 Table 2 Clarity. Sorry for the clarity issue. We will revise the caption of Table 2 to highlight its purpose.

[A] Parascandolo et al. *Learning independent causal mechanisms.* ICML'18 [B] Suter et al. *Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness.* ICML'19 [C] Heinze-Deml et al. *Conditional Variance Penalties and Domain Shift Robustness.* arXiv [D] Qi et al. *Two causal principles for improving visual dialog.* CVPR'20 [E] Siddharth et al. *Learning disentangled representations with semi-supervised deep generative models.* NeurIPS'17