Match me if you can: Semi-Supervised Semantic Correspondence Learning with Unpaired Images – Appendix –

In this appendix, we provide additional information that complements the materials in our main paper and various aspects of our research. Specifically, we include the following items:

- **Novel Robustness Evaluation Benchmark.** We introduce a new robustness benchmark for semantic correspondence (dubbed **SPair-C**). To our knowledge, this is the first benchmark for evaluating the robustness of semantic correspondence learning methods.
- **Further Analyses.** We present an ablation study and further analyses to understand the effectiveness of our method.
- **Further Training Details.** We present further training details that elaborate on our method of leveraging unlabeled data to boost performance.
- **Visualizations.** We showcase qualitative results by comparing our method with state-of-the-art methods.

A Robustness Evaluation Benchmark

We introduce a new corruption benchmark to verify the robustness of dense correspondence learning methods. This benchmark is dubbed SPair-C (*i.e.*, a corrupted SPair-71k [9]), which was mentioned in the Robustness evaluation section of the main paper. Its purpose is to complement the existing dense correspondence learning task by providing a more challenging dataset for evaluating the robustness of models.

Dataset details. Since hardly corrupted or noise-injected images have been used to evaluate dense correspondence learning, we construct a new corruption robustness benchmark for semantic correspondence following the existing regime [5]. The future applicability of a model can be determined by evaluating its robustness against corrupted images involving frequently observed corruption occurring in the wild.

Fig. A shows 15 types of corruptions in the SPair-C dataset, selected among corruptions [5] used for measuring robustness. We choose appropriate corruptions for pixel-level prediction from noise, blurred weather, and digital categories, which do not affect significant point changes after corruptions. The finalized categories are noise (Gaussian, shot, impulse, and speckle), blur (defocus and Gaussian), weather (snow, frost, fog, and spatter), and digital categories (brightness, contrast, saturate, pixelated, and JPEG). We apply the five levels of the severity s per each type of corruption exemplified in Fig. B. Therefore, a set of 75 common visual corruptions, which enable models to be fooled by small

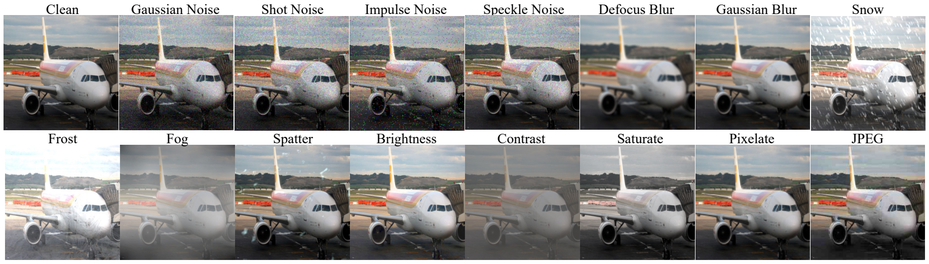


Fig. A: Visualization of corrupted images in SPair-C. The corrupted images of one sample consist of types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions.

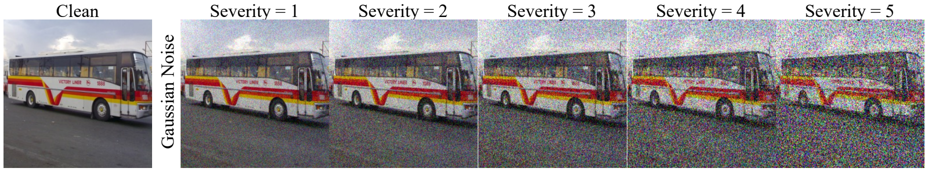


Fig. B: Corrupted images with different severities. We visualize a chosen image from SPair-71k [9] with severity from 1 to 5. The images get noisier as the severity increases.

changes in the original image, are used for one test image. We use the codebase[1] to build the benchmark.

B Further Analyses

In this section, we first ablate our model regarding the confidence threshold. Then, we analyze PCK concerning the tolerance threshold α. Finally, We examine our proposed method `MatchMe` by training it, varying the ratio of unlabeled images to labeled training images.

Ablation study on confidence threshold. The confidence threshold τ in Eq.(8) is a critical hyper-parameter that determines the quantity and quality of machine-annotated data used for training. To investigate the relationship between the confidence threshold and model performance, we conduct an ablation study with the threshold at intervals of 0.2 across a wide range of confidence thresholds from 0.1 to 0.9. We use CATs [1] for this study.

As shown in Tab. A, while the threshold value of 0.7 produces the highest PCK value, other values do not significantly degrade performance and still produce higher PCK values than the baseline's PCK 49.9, which do not use our method. The PCK trend suggests that the quantity of machine-annotated data

[1] https://github.com/hendrycks/robustness

Table A: Impact of machine-annotated data. We study the sensitivity of PCK as the quantity and quality of machine-annotated data change by the confidence threshold τ. We adjust τ from 0.1 to 0.9 and confirm how PCK changes. CATs [1] is used for this study. We observe the model can yield the maximal PCK with the tuned τ but shows insensitive PCKs to τ.

Confidence threshold (τ)	PCK
0.1	52.7
0.3	52.6
0.5	52.6
0.7	53.0
0.9	52.3

Table B: PCK comparison with state-the-art methods under varying tolerance (α). We report PCK for different α of the state-of-the-art methods on SPair-71k [9]. Numbers in bold denote the best. `MatchMe` outperforms all the methods by a large margin.

α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
CATs [1]	1.93	7.0	13.8	20.9	27.7	33.6	38.6	43.0	46.8	49.9
CATs++ [1]	4.3	14.6	25.0	33.9	40.8	46.4	50.8	54.3	57.3	59.8
SemiMatch [7]	2.1	7.7	15.0	22.4	29.0	34.8	39.7	43.9	47.6	50.7
SCORRSAN [6]	3.6	12.1	21.3	29.3	35.8	41.0	45.2	48.8	51.7	55.3
`MatchMe` (ours)	**6.1**	**18.5**	**29.9**	**38.6**	**45.1**	**50.0**	**53.9**	**57.0**	**59.7**	**62.0**

is more critical than its quality, as demonstrated by the lower PCK value at the higher confidence threshold (*e.g.*, $\tau = 0.9$) compared to that at the lower threshold (*e.g.*, $\tau \leq 0.3$). Moreover, the low sensitivity to threshold values suggests that our novel data itself contributes to improving model performance.

PCK analysis. We report PCK results with the various tolerance thresholds from 0.1 to 0.01 on SPair-71k in comparison between ours and the competitive baselines [2,3,6,7] in Tab. B. The results demonstrate that our expanded keypoint correspondences, amplified at both pixel-level and image-level, enable the trained model to more accurately estimate the correspondences, as evidenced by the significant PCK gaps with much stricter tolerance criteria (smaller α values), compared to other methods. Moreover, the results show that as the tolerance threshold of PCK (α) decreases, the gap with the baseline [2] does not narrow but widens. For example, at $\alpha = 0.1$, the performance gap is 2.2, but at $\alpha = 0.05$, it is 4.3, showing a gap approximately twice as large. Therefore, this demonstrates that our predicted point correspondences closely approximate the GT point correspondences.

Analysis on using labeled/unlabeled data. We conduct further experiments by varying the ratios of unlabeled and existing labeled training data ratios.

Table C: PCK of MatchMe trained with a varying fraction of labeled images. We observe that MatchMe is not heavily dependent on a large fraction of labeled images. It achieves high PCK even when trained with only 20% labeled data from the entire labeled images, which amounts to approximately 0.04% of the unlabeled data.

PCK with fraction of labeled data (baseline: 49.9)				
0.2	0.4	0.6	0.8	1.0
50.2 (+0.3)	50.5 (+0.5)	50.7 (+0.8)	51.1 (+1.2)	52.0 (+2.1)

We aim to determine the necessary amount of labeled data for performance and understand the dependency on labeled data. Table. C offers the following observations: 1) the minimum amount of labeled data required is quite low, and 2) there is a low reliance on labeled data. For example, when using 20% images of the entire labeled data, which corresponds to approximately 0.04% of the unlabeled data, MatchMe outperforms the baseline (*i.e.*, 49.9), trained with the whole labeled data in a supervised manner, on SPair-71k.

C Further Details on Training

Unlabeled images for training We use the labeled data in the training set of SPair-71k [9] along with the unlabeled data in PASCAL VOC 2012 [4], which is the source of the SPair-71k. We assign images for each object class according to the corresponding classification labels. Only non-overlapping images in the validation and test set of SPair-71k are used for training to avoid cheating. Furthermore, images in the 'dining table' and 'sofa' classes are not included in the same way that the dining table and sofa categories are not used as in SPair-71k. Since SPair-71k was built to have similar numbers of labeled data for each class, we balance the number of unlabeled data for each class to ensure that their distribution matches that of the labeled data. We use different batches of unlabeled data in every iteration to diversify the training sample.

D Visualizations

Handling unseen data. We qualitatively verify the generalizability of our method by examining the labels generated by MatchMe for newly captured data from the ImageNet dataset [10], which were not included in our training dataset. As shown in Fig. C, our model produces high-quality correspondences for both data from known classes (*e.g.*, Cat) and even previously unseen classes (*e.g.*, Hen). This demonstrates the high potential to extend our method by incorporating newly captured data into the existing unlabeled data for training or evaluation.

Comparison with State-of-the-arts. Alongside the qualitative results shown in Fig.3 in the main paper, we offer further visualizations of example pairs with

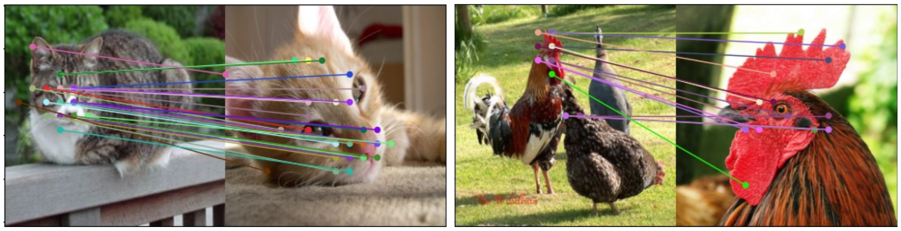
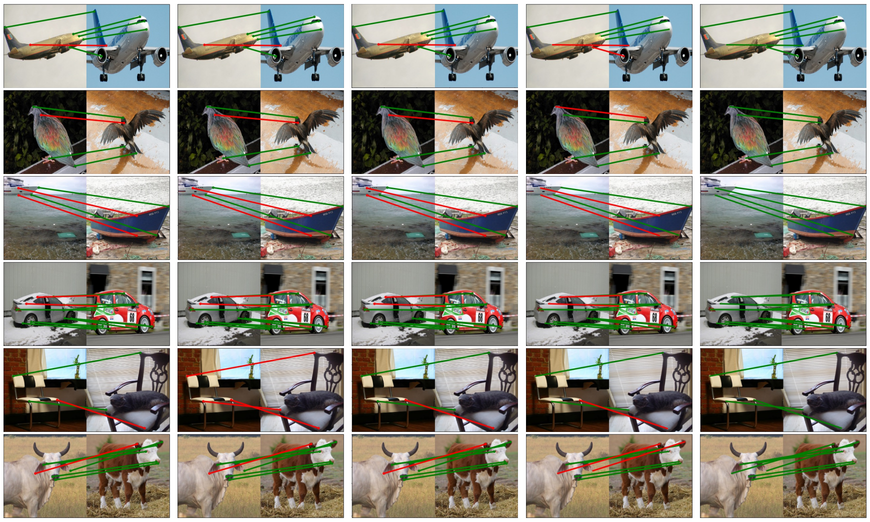


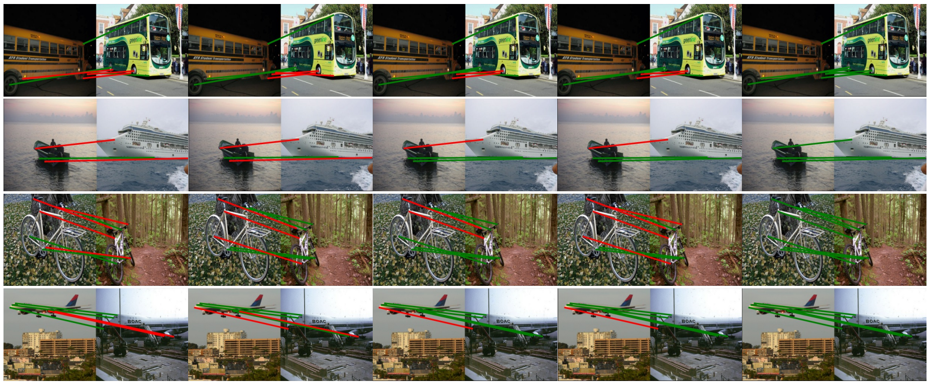
Fig. C: Qualitative example of handling unseen images. We demonstrate the applicability of `MatchMe` using newly acquired data beyond the training dataset. This includes images for both existing and unseen classes, with the images sourced from ImageNet [10]. The top two images display matching results for cat images, while the bottom images feature the unseen class, Hen. The results indicate the strong applicability of our method, as evidenced by the high accuracy of the correspondence in both cases.



(a) CATs **(b)** SemiMatch **(c)** SCORRSAN **(d)** CATs++ **(e)** `MatchMe`

Fig. D: Qualitative results on SPair-71k in comparison with SOTA methods (cont'd) The point-to-point matches are drawn by linking key point pairs with line segments. Green and red lines denote correct and incorrect predictions with respect to the ground-truth pairs, respectively. We observe that ours performs much better compared with the counterparts across all the sample image pairs.

their predicted matches for `MatchMe` and the highly competitive methods in both the supervised and semi-supervised regimes: CATs [1], CATs++ [2], SemiMatch [7], SCORRSAN [6]. As shown in Fig. D, our approach produces more accurate estimations of correspondences between image pairs across various object classes and differences in variation factors compared with other methods.

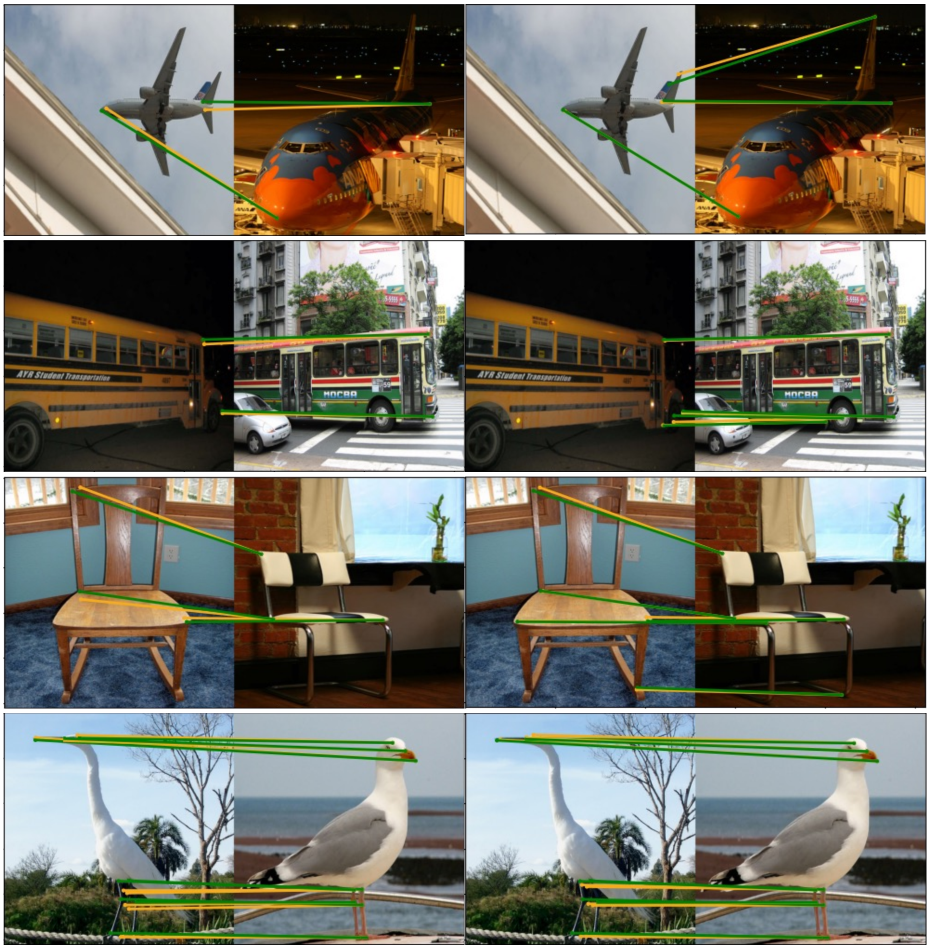


(a) CATs++ (Baseline) **(b)** (a) + CNNGeoU **(c)** (a) + PWarpC **(d)** (a) + SCORRSAN **(e)** (a)+ `MatchMe` (Ours)

Fig. E: Qualitative results on SPair-71k [9] in comparison with other semi-supervised methods: For a fair comparison, we use (a) the fixed baseline CATs++ [2] for all semi-supervised methods, (b) + CNNGeoU [8], (c) + PWarpC [11], (d) + SCORRSAN [6], and (e) + `MatchMe`. The point-to-point matches are drawn by linking key point pairs with line segments. Green and red lines denote correct and incorrect predictions with respect to the ground-truth pairs, respectively. We observe that ours performs much better compared with the counterparts across all the sample image pairs.

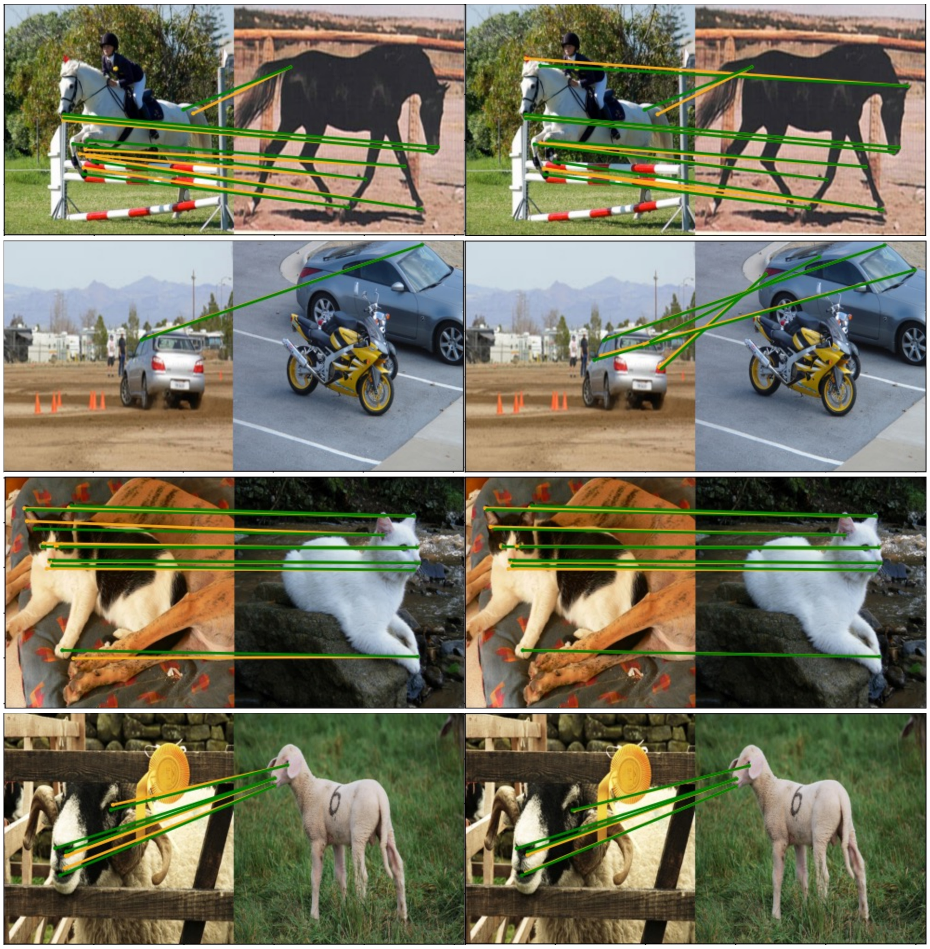
Comparison with semi-supervised methods. We show qualitative results to complement the aspect of the controlled experiments for the learning methods section of the main paper. For a fair comparison, all methods are trained under the same network architecture, suggested in CATs++ [2]. As shown in Fig. E, we predict correct correspondences, even in challenging samples that exhibit significant differences in scale and viewpoint between image pairs, unlike other methods, which tend to produce incorrect predictions for such samples.

Qualitative PCK analysis. In addition to quantifying model performance through the analysis of PCK values, we demonstrate the superiority of our method by visualizing its predictive quality in Fig. F and Fig. G. We compare the differences between the correctly predicted and the ground truth (GT) point correspondences at $\alpha = 0.1$, as indicated by yellow and green colors, respectively. This visualization illustrates how many more points our model predicts correctly as well as how closely our model's predictions align with the correct GT key points, even at the extreme points compared to the baseline [2]. Specifically, the example of the sheep class, having the most minor categorical PCK difference compared to the baseline, illustrates that, even though the PCK achieved by our method is similar to that of the baseline, the quality of the predicted correspondence is superior. This outstanding performance can be attributed to the fact that our method generates machine-annotated point correspondences, providing diverse and rare supervisions that are difficult to obtain through the limited amount of manually annotated GT key points.



(a) Baseline **(b)** `MatchMe`

Fig. F: Visualization of the difference between correctly predicted points and ground truth (GT) points on SPair-71k. The GT points in the left images corresponding to the GT points in the right images for each image pair are marked in green lines, and the predicted point correspondences are marked in yellow lines. The closer the predicted correspondence to the GT correspondence is, the more accurate the prediction. Notice that if only the green line is visible, the predicted and GT point correspondences are perfectly matched.



(a) Baseline **(b)** `MatchMe`

Fig. G: Visualization of the difference between correctly predicted points and ground truth (GT) points on SPair-71k (cont'd). The GT points in the left images corresponding to the GT points in the right images for each image pair are marked in green lines, and the predicted point correspondences are marked in yellow lines. The closer the predicted correspondence to the GT correspondence is, the more accurate the prediction. Notice that if only the green line is visible, the predicted and GT point correspondences are perfectly matched.

References

1. Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Semantic correspondence with transformers. arXiv preprint arXiv:2106.02520 (2021) 2, 3, 5
2. Cho, S., Hong, S., Kim, S.: Cats++: Boosting cost aggregation with convolutions and transformers. arXiv preprint arXiv:2202.06817 (2022) 3, 5, 6
3. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS (2011) 3
4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015) 4
5. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019) 1
6. Huang, S., Yang, L., He, B., Zhang, S., He, X., Shrivastava, A.: Learning semantic correspondence with sparse annotations. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV. pp. 267–284. Springer (2022) 3, 5, 6
7. Kim, J., Ryoo, K., Seo, J., Lee, G., Kim, D., Cho, H., Kim, S.: Semi-supervised learning of semantic correspondence with pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19699–19709 (2022) 3, 5
8. Laskar, Z., Kannala, J.: Semi-supervised semantic matching. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 6
9. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019) 1, 2, 3, 4, 6
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015) 4, 5
11. Truong, P., Danelljan, M., Yu, F., Van Gool, L.: Probabilistic warp consistency for weakly-supervised semantic correspondences. arXiv preprint arXiv:2203.04279 (2022) 6