A Appendix

In this section, we present arrays of ablation studies to understand crucial properties of the source domains. All experiments in this section are performed in the fully-correlated target domains.

A.1 DiagVib Dataset Configurations

As mentioned in the experiments in section 4, we use datasets based on the DiagVib framework which allows generation of synthetic datasets with custom configurations of basic visual factors. We consider five factors whose number of possible values are listed according to Table 4. It should be noted that DiagVib-Caltech and DiagVib-Animal have different number of available shapes.

Table 4: Different visual factors, which can be configured in the DiagVib framework

Factor	**Description**	**No. of Classes**
Shape	Object boundary defined by a silhouette	Caltech: 50 Animal: 10
Color	Hue value in HSV space	12
Lightness	Lighting condition (e.g., bright, dark)	4
Texture	Pattern drawn inside the object (e.g. wooden, checkerboard)	5
Background	Background color	3

A.2 Ablation Studies on the Source Domain

Table 5: Accuracies of FactorSRC-IL in the target domain (DiagVib-Animal) with variations of source domains to demonstrate the impact of their uncorrelated factors.

Source Setting	Images from	Correlated Factors	Target HM Acc.
Uncorrelated	DiagVib-Caltech	False	**33.5 ± 1.0**
Correlated	DiagVib-Caltech	True	2.5 ± 0.7
Target	DiagVib-Animal	True	1.7 ± 0.4

Impact of Uncorrelation of Factors In this study, we aim to investigate whether the improvement in generalization performance after incorporating the

source domain stems from uncorrelating visual factors. We compare the following source dataset settings: *a*) Uncorrelated: all factor combinations are available *b*) Correlated: shape and color factors are one-to-one correlated *c*) Target: use correlated data sampled from the target distribution (DiagVib-Animal) for training. For a fair comparison, the number of target-associated factors (shape/color) are reduced to 10 for Uncorrelated and Correlated settings, so as to match the Target setting. Results in Table 5 indicate that the Uncorrelated setting yields significantly higher accuracy compared to others. This empirically shows that this improvement in OOD generalization is indeed due to the uncorrelated nature of the source dataset and not just a mere result of the increased dataset size.

Impact of Shape Variations We conduct another experiments to understand if the complexity of the shapes provided in the source domain affects accuracies in the target domain. We modify the DiagVib-Caltech source domain to use MNIST shapes and compare it to the original setting with Caltech shapes (we use 10 shapes in both cases to be comparable). Table 6 shows that the setting with MNIST shapes has lower accuracies. We believe that this is due to the fact that MNIST has less intra-class shape variation compared to Caltech. For example, the shape of the number ones are not much different across different samples. This degrades the generality of the learned shape representation. This experiment suggests that a primary concern when constructing a source dataset should be intra-class variability of each factor.

Table 6: Accuracies of FactorSRC-IL with variations of shape in the source domain.

Shape	DiagVib-Animal HM. Acc	Color-Fruit HM. Acc	AO-CLEVR HM. Acc
Caltech	**33.5 ± 1.0**	**56.0 ± 2.7**	**36.4 ± 1.8**
MNIST	31.9 ± 0.7	46.9 ± 3.2	29.0 ± 1.4

Impact of Available Factors In this experiment, we study the effects of varying the number of basic visual factors represented in the source domain. According to the result in Table 7, while we find that increasing the number visual factors yields better performance overall, for some factors, the effect on different target domains is different. For instance, with DiagVib-Animal as a target, including the background as a factor in the source domain improves performance significantly, due to the fact that the target domain has variable background colors. In contrast, this effect is not observed on Color-Fruit, whose images have a constant background. Instead, learning lightness and texture can improve generalization performance since these two factors have high variation in this target

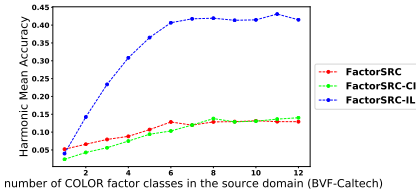
domain (DiagVib-Animal doesn't have variations of lightness and texture). We can infer from this result that the performance in the target domain tends to be better if the source domain captures basic factors which are represented in the target domain.

Table 7: HM Accuracies from FactorSRC-IL approach on DiagVib-Caltech source domain with different presence of factors (S, C, L, T, B correspond to Shape Color, Lightness, Texture and Background respectively).

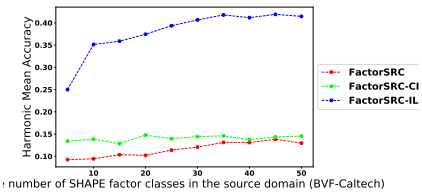
Factors	DV.-Animal	Color-Fruit
S/C	28.6 ± 1.2	49.5 ± 5.0
S/C/L	29.3 ± 1.1	53.2 ± 3.5
S/C/L/T	31.5 ± 0.5	**57.7±3.3**
S/C/L/T/B	**41.3±1.6**	57.0 ± 4.7

In summary, we have performed an array of ablation studies to analyze properties of the source domain which encourage better generalization in target domains. Firstly, we showed that visual factors in the source domain should be uncorrelated. This facilitates disentanglement of visual factors' representations, which in turn leads to less shortcut vulnerability. Secondly, we demonstrated that intra-factor variability is crucial in order for deep networks to learn generalizable representations. Lastly, visual factors encoded in the source domain should cover as many predictive features in the target domain as possible. We believe that these three aspects are among the most important criteria, which should guide practitioners towards choosing better source domains for augmenting biased training datasets.

Impact of Variations of the Number of Factor Classes From our experiments, we showed that FactorSRC-IL can learn factor representations from the source domain, which improve compositional generalization in several target domains. In this section, we would like to investigate how the number of factor values in the source dataset affects generalization performance in the target domain. For this purpose, we vary the number of factor values associated to each target label (shape and color in our setting) and measure compositional generalization in the DiagVib-Animal target domain. Results are shown in Figure 6 and indicate that a higher number of factor values generally leads to the better performance. This is intuitive since a higher number of classes should encourage networks to learn more general factor representations. An interesting observation is that the network needs only around 8 color classes to be close to optimal performance while around 35 classes are needed in the case of shape. We believe this is due to the fact that shape, as a basic visual factor, is more high-dimensional and thus more difficult to model than color.



(a) Varying the number of colors



(b) Varying the number of shapes

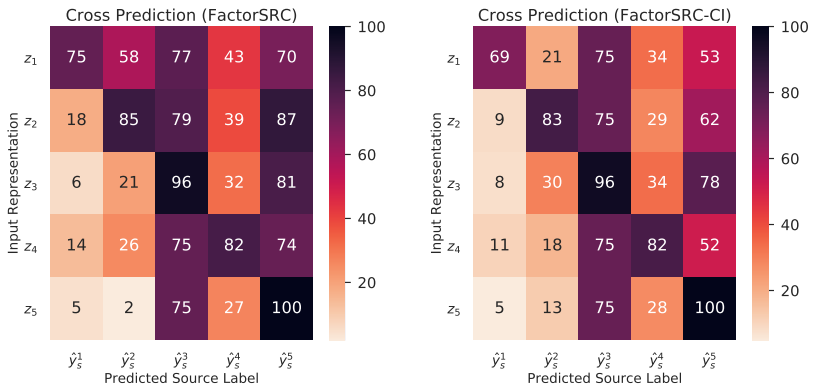
Fig. 6: Accuracies of FactorSRC-IL on the DiagVib-Animal with different number of factor classes (color and shape) while maintaining the same configuration for the other factors on the DiagVib-Caltech source domain.

A.3 Effect of the CI Constraint on the Source Domain

In our experiment section, we stated that the Cross-Factor Independence Constraint (CI) promotes independence of factor reresentations in the source domain. In this section, we provide experimental evidence supporting our claim. To this end, we compare cross-prediction accuracies with and without the CI constraint, for each factor among $z_1, z_2, \ldots, z_K$. Results are visualized in Figure 7. We can see that, while direct-prediction accuracies are comparable with and without the CI constraint, the cross-prediction performance decreases significantly when the CI constraint is introduced. This supports our hypothesis that the CI constraint induces a higher degree of independence among factor representations in the source domain.

A.4 Importance of Association Matrix Assignment

We hypothesize that the shape is a generic robust factor that can be used to predict object types. So, we manually associate the shape factor from the source domain to the object type of target domains in the association matrix A in all fully-correlated target domain scenarios. To validate this hypothesis, we perform an ablation study to evaluate network performance when different configurations of source factors are chosen in association matrix A. Performance of all configurations can be visualized as two-dimensional heatmaps for different datasets as in Figure 8. The value in each cell C_{ij} of a heatmap represents the average HM accuracy when target attribute and target object associate to source factor i



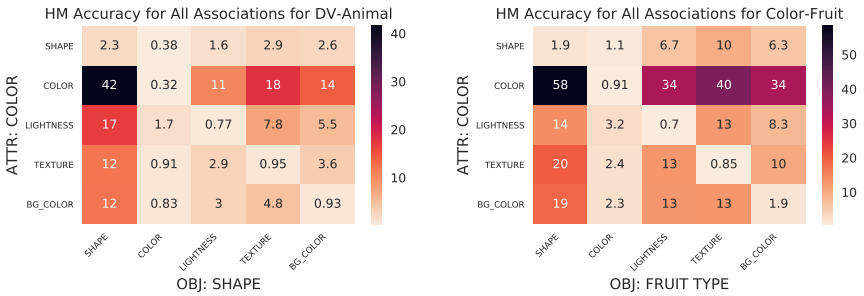
(a) FactorSRC Heatmap (b) FactorSRC-CI Heatmap

Fig. 7: Heatmaps displaying direct and cross-factor prediction accuracies using DiagVib-Caltech and DiagVib-Animal as source and target domain respectively. Each cell indicates the accuracy attained when a single factor representation (z_k in each row) is used to predict labels ($\hat{y}_s^l$) with a linear model. A higher degree of independence among factor representations is expected to yield similar diagonal values but lower off-diagonal ones (cross-prediction). Factor indices from 1 to 5 correspond to shape, color, lightness, texture and background respectively.

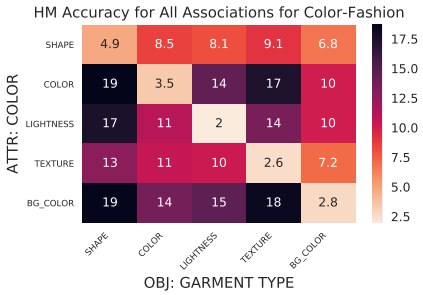
(row of heatmap) and j (column of heatmap) in the source domain respectively. In this regard, cell C_{21} in each heatmap represents HM accuracy of a configuration setting when target attribute/object associate with source color/shape factors.

First of all, considering the results from DiagVib-Animal and Color-Fruit target datasets in Figure 8a and 8b, the highest values of C_{21} (42% and 58%) for both datasets empirically support our hypothesis that the shape factor is a robust factor for predicting object types in the target domain. In the Color-Fruit dataset, an interesting observation can be made as texture factor is also a predictive of fruit type in addition to the shape factor (C_{13} of 40% in Figure 8b). This result shows capability of the texture to predict fruit type which aligns to the estimated association matrix in Figure 5e. From these results, we can empirically validate our hypothesis and show that a proper configuration of the association matrix is important to alleviate model vulnerability to shortcuts.

In a more challenge dataset Color-Fashion, even though our configuration C_{21} is among the best, there are other configuration settings that reach similar result (Figure 8c). This behavior can be explained intuitively: considering the target object type (garment type), models have high performance when associating the object type to either shape or texture factors (can be seen as cells of high values on the first and the fourth columns). This behaviour is similar to the case of the fruit type in Color-Fruit dataset emphasizing the fact that both shape or texture



(a) DiagVib-Animal (b) Color-Fruit



(c) Color-Fashion

Fig. 8: HM Accuracies with different configurations of association matrices in different target datasets. The cell C_{ij} in each heatmap corresponds to the accuracy when the association matrix associating target attribute and object type to the i-th and j-th factor respectively. The order of factors is shape, color, lightness, texture and background color.

can be a predictive factor for object types. For the target attribute type (garment color), its associations to color or background color produce high accuracies (can be seen as cells of high values on the second and the fifth rows). This implies that information of the garment color is contained in factor representations of both color and background color. The underlying reason can be due to the design of our source domain. In the DiagVib-Caltech source domain, boundaries between foregrounds and backgrounds are simple as backgrounds are only plain colors. However, in the case of the Color-Fashion target domain, its backgrounds are more complex representing realistic scenes. This suggests redesigning of the source domain. One possibility is to use more realistic backgrounds such as place images similar to [1].

A.5 Implementation Details

In this section, we provide details of our network design and training hyperparameters.

Table 8: Accuracies on DiagVib-Animal, Color-Fruit, AO-CLEVR and Color-Fashion target domains with the similar experiment setup as in Table 1. However, calibrated bias terms are incorporated before computing seen, unseen and HM accuracies.

Approach	Use Source?	DiagVib-Animal			Color-Fruit			Color-Fashion		
		Seen	Unseen	HM	Seen	Unseen	HM	Seen	Unseen	HM
LabelEmbed+	✗	96.3	10.3	13.2	100	19.7	12.5	90.0	13.4	16.9
TMN	✗	95.7	7.0	12.2	100	17.9	29.2	89.4	7.0	8.9
CGE	✗	92.8	11.8	15.0	100	24.9	32.9	88.1	20.8	21.0

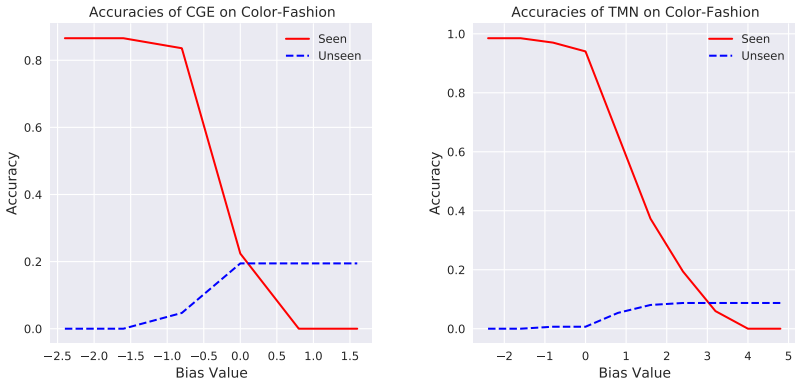
For all variants of the factorized architecture illustrated in Figure 2b (Factor-0, FactorSRC, FactorSRC-CI and FactorSRC-IL), the encoder G is a fully-connected network with 2 hidden layers, which outputs multiple factor representations, each one of length 64. All branches of H_s and H_t (i.e., all prediction heads in $\{h_s^k\}_{k=1}^K \cup \{h_s^o, h_s^a\}$) consist of a fully-connected network with 1 hidden layer. We set the hyperparameter λ equal to 10 when we include the source dataset for all experiments for fair comparison. In section 3.4, we introduce strategy to learn the factor association matrix with additional regularization constraints. Hyperparameters α, β and τ used for the regularization constraints are 5, 20 and 0.33 respectively.

For training we use Adam as an optimizer, a learning rate of 0.01 and weight decay equal to $5e^{-5}$. The optimal network is selected based on the loss on a validation split over 100 epochs.

A.6 Bias Terms for Adjusting Likelihood of Unseen Combinations

As mentioned earlier in section 4, unlike some prior works [23, 21], we evaluate compositional generalization without bias terms to adjust the likelihood of unseen combinations (using higher bias makes the model more likely to predict unseen combinations). The reason is that tuning of the bias terms requires availability of samples from unseen combinations. This violates the zero-shot assumption. Additionally, bias terms are designed to be applicable only with certain baselines based on compatability scores (such as LabelEmbed+, TMN and CGE) but not the others leading to unfair comparison.

For completeness, we will also provide results when the calibrated bias terms are incorporated during evaluation for LabelEmbed+, TMN and CGE. The seen, unseen and HM accuracies reported here correspond to their maximum values when their optimal bias terms are used (maximum seen and maximum unseen accuracies usually employ different optimal values of bias terms). Adopting the same experiment setup similar to Table 1, baseline performance with calibrated bias terms is presented in Table 8. According to the results, the accuracies are higher when the calibrated biases are incorporated. However, the overall HM accuracies are still lower than results from our approaches. This still highlights vulnerability of these baselines to shortcuts.



(a) Seen/Unseen Accuracies of CGE (b) Seen/Unseen Accuracies of TMN

Fig. 9: Seen/Unseen Accuracies of TMN and CGE baselines evaluated with different bias terms.

Here, we also investigate why seen accuracies of certain baselines are low in Table 1 (e.g., CGE on Color-Fashion). We can understand this behavior by observing seen/unseen accuracies using different bias terms. According to Figure 9a, the seen accuracy of CGE on Color-Fashion can be as high as 88.1 (similar to Table 8) when low bias term is used. However, in our experiment, we choose not to use bias terms for evaluation as per the reasons described above. Therefore, the reported seen accuracies on Table 1 are computed with bias terms of zero values. From Figure 9a, the seen accuracy of CGE on Color-Fashion is reduced to 21.6 (similar to Table 1). In contrast to CGE, the seen accuracy of TMN with zero bias term is already high (see Figure 9b). Therefore, we do not see low seen accuracy of TMN on Table 1.

A.7 Sweeping Weight of Loss for the Source Domain

The hyperparameter λ is used to weight the importance of $\mathcal{L}_{source}$ during training. Here we investigate its impact on the generalization performance attained in the target domain. Results of our analysis are shown in Figure 10. We note that, for FactorSRC and FactorSRC-CI, the harmonic mean of seen and unseen accuracies increases with higher λ values. This suggests that these two models are less sensitive to biases in the target dataset when λ is increased. High values of λ encourage FactorSRC and FactorSRC-CI to be more similar to FactorSRC-IL as $\mathcal{L}_{target}$ becomes less important to update G relative to $\mathcal{L}_{source}$. FactorSRC-IL, on the other hand, performs consistently when $\lambda > 0$. This result is reasonable since, when the IL constraint is introduced, $\mathcal{L}_{source}$ and $\mathcal{L}_{target}$ are independently used to update different parts of the network (they update $\{G, H_s\}$ and $\{H_t\}$ respectively). We note that, even though the higher λ leads to better performance, we reserve to use λ at 10 in our experiment so that we can study

effects from other loss terms. It should be noted that changing the value of λ here does not play a major role in our analysis since the key trends would be the same.

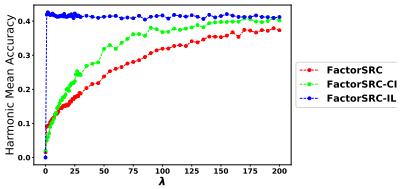


Fig. 10: HM Accuracies using the DiagVib-Animal target domain and the DiagVib-Caltech source domain with different λ values to weight the importance of $\mathcal{L}_{source}$. Notice that, higher λ values encourage models to behave closer to FactorSRC-IL. We sample λ on the low values with higher frequency to better highlight the faster increasing trends.

A.8 Cross-Factor Independence Constraint Algorithm

The Cross-Factor Independence constraint is implemented as a two-step optimization approach. In the first step, we update H' by minimizing the sum of cross entropy loss terms for all cross-factor predictions, i.e.,

$$\mathcal{L}_{H'} = \sum_{\forall k_1, k_2; k_1 \neq k_2} CE(H'_{k_1 k_2}(z_{k_1}), y_s^{k_2}) \; . \tag{3}$$

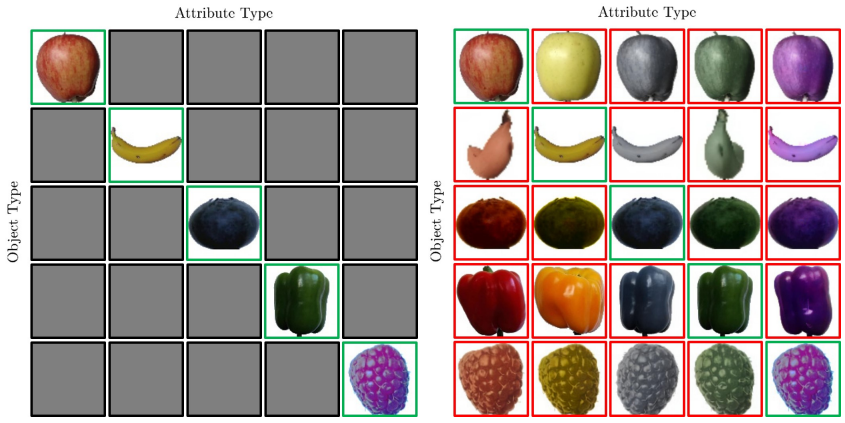
Subsequently, we update the whole network by minimizing the combination of $\mathcal{L}_{target}$, $\mathcal{L}_{source}$, and an additional independence loss $\mathcal{L}_{CI}$. In principle, $\mathcal{L}_{CI}$ could be formulated as $-\mathcal{L}_{H'}$ but we found that this leads to training instabilities due to the fact that such a loss is unbound. Instead, we minimize the cross entropy between the predictions of H' and a uniform label distribution. This encourages each factor representation to be uninformative with respect to all other factors. Mathematically, $\mathcal{L}_{CI}$ can be written as follows:

$$\mathcal{L}_{CI} = \gamma \sum_{\forall k_1, k_2; k_1 \neq k_2} CE\left(H'_{k_1 k_2}(z_{k_1}), \frac{\mathbf{1}^{N_{\mathcal{F}_s^{k_2}}}}{N_{\mathcal{F}_s^{k_2}}}\right) \tag{4}$$

, where $\mathbf{1}^N$ indicates a vector of ones with length N, $N_{\mathcal{F}_s^k}$ is the number of factor values of the k-th factor and $\gamma \geq 0$ is a hyperparameter (we use $\gamma = 5$).

A.9 Seen Accuracy form FactorSRC-IL on DiagVib-Animal

According to the result from Table 1, we notice that, on DiagVib-Animal, even though the HM accuracy of FactorSRC-IL is significantly higher than all other



(a) Training samples (b) Test samples

Fig. 11: Examples of Color-Fruit dataset images. (a) Training samples containing images of fully-correlated attribute-object combinations denoted with green-bordered boxes (One object type always has one color and vice versa). (b) Test samples are, on the other hand, uncorrelated i.e., consisting of images with any attribute-object combinations (i.e., each object (fruit type) can appear with any attributes (color)). Fruit images whose colors are not available in the original Fruits 360 dataset are obtained by using the recolorization technique in [29]

approaches, the seen accuracy is dropped significantly (to 56.3%). The drop of the seen accuracy only presents in the case of DiagVib-Animal but not other target domains. We suspect that this behavior could stem from the lower random chance accuracy (1% on DiagVib-Animal compared to 4% and 4.7% on other target domains) or just the complexity of the DiagVib-Animal (with high intra-class variations and various backgrounds). In this regard, we conduct an experiment with reduced number of attribute/object labels from 10 to 5 so that it has the random chance accuracy of 4% which is the same as the one of Color-Fruit. In this regard, seen, unseen and HM accuracies on this reduced version of the DiagVib-Animal target domain are 74.2%, 52.6% and 61.4% respectively. Notice that, the seen accuracy is higher than the one on the original version but it is still relatively lower compared to the seen accuracies on other target domains. We can, therefore, conclude that the lower of the seen accuracy on DiagVib-Animal stems not only from its lower random chance accuracy but also from the complexity of the target domain itself.

A.10 Color-Fruit Dataset Generation

In order to generate the *Color-Fruit* dataset, used in our experiments, we use fruit images from the Fruits 360 dataset [20]. Five fruits (Apple, Banana, Blueberry, Pepper and Raspberry) are selected as they have distinct colors (red, yel-

low, blue, green and magenta), which facilitate the evaluation of compositional generalization in the case of fully-correlated seen combinations.

During evaluation, however, fruits with different colors are required. Thus, we perform recolorization of images in the test split using the approach described in [29]. Basically, an original test image is recolorized into median colors of all other fruits (e.g. a banana image is transformed such that it has a color similar to that of an apple, a blueberry, a pepper and a raspberry). More detailed visualization of the dataset is presented in Figure 11.