A. The "What If" Challenges for Deep Neural Networks (DNNs)

A causal inferable DNNs for reasoning chaotic real-world patterns [56, 58, 92] would be necessary for many practice scenarios. Recent regulatory concerns (e.g., GDPR [3]) on Artificial Intelligence (AI) safety and self-driving automobile accidents also highlight the importance and the emergence of understanding: (1) "What" does a DNNs model "learn for accurate label-prediction and (2) utilizing the "Why" relationship between labels generated by human knowledge and their conceptional pattern representations in the real world.

We propose COCO$_{CP}$ dataset based on MS-COCO [38]. COCO$_{CP}$ dataset includes classes having similar object(concept) but are different in context that could possibly result in confusing interpretation and recognition by human psychology studies (e.g., giraffe and elephant, etc.). In total, three different common causal pairs are formatted, namely giraffe-elephant (g-e) with 3316 images, stop sign-fire hydrant (s-f) with 2419 images, and bike-motorcycle (b-m) with 4729 images. A pair of stop sign and fire hydrant has been selected to study public awareness of on-road visual detection.

A.1. Correlation and Causation in Vision Task

Correlation [56, 58, 59] is a statistical measure that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the difference in the values of the other variable. Causation [58, 59] indicates that one event is the result of the occurrence of the other event; i.e., there is a causal relationship between the two events. This is also referred to as cause and effect.

Theoretically, the difference between the two types of relationships are easy to identify — an action or occurrence can cause another (e.g., having rain droplet causes an increase in the risk of developing rain day), or it can correlate with another (e.g., a visual rain droplet is correlated with a red umbrella, but it does not lead to having a representation of a red umbrella in vision directly). In practice, however, it remains difficult to establish cause and effect, compared with establishing correlation clearly. Yet, most of the current deep learning method focused on directly visualizing the patterns after the visual model is trained without verifying (e.g., intervention methods [59, 92]) the causation between each representing patterns.

Neural Causation Coefficient (NCC) [43] is an existing observational causal discovery validation method for the joint distribution of a pair of related proxy variables. However, they mainly focus on discussing the object-context causal hypothesis in the image setting, but barely covers

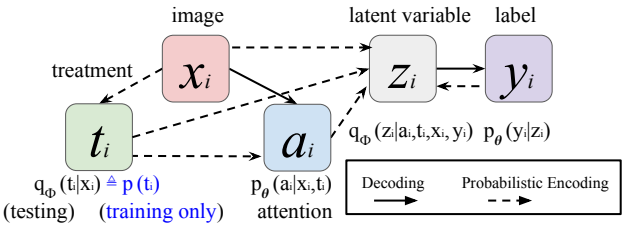


Figure 6: Generative process of Treatment Learning Transformer model. **Solid lines**: decoding phase. **Dashed lines**: probabilistic encoding phase. The treatment information could be only observed during training.

the measurement of perturbation effect. NCC leverages such embedded joint distribution as a regularization term to encourage the learning of causal or anticausal patterns in neural networks. Extended from Eq. 12, **object-feature ratio (OFR)** [43] could be extended to a **treatment-feature ratio (TFR)** score s_l^t by do-operator [59] as:

$$s_l^t = \frac{\sum_{j=1}^{m} \left| (f_{jl}^c | do(t)) - (f_{jl} | do(t)) \right|}{\sum_{j=1}^{m} \left| (f_{jl)} | do(t) \right|}, \qquad (13)$$

where (l=512) is the input feature length, and f_j^t corresponding vectors of feature scores is equivalent to $F_{DNN}(x_j^t)$.

A.2. Basic Metrics for Causal Inference

By defining the CE of the individual treatment effect (ITE) as the difference between two potential outcomes for the individual [44, 76], the average treatment effect (ATE) is defined as the expected value of the potential outcomes over the subjects.

B. Dataset

B.1. Dataset Statistics: CPS General

In order to accurately measure visual causality, we select six categories from MS-COCO [38] and match them into three causal pairs, which are giraffe-elephant, bicycle-motorcycle, and stop sign-fire hydrant. Each pair is chosen by its matching characteristic and background and split into train and validation set by MS-COCO default setting. With similar characteristics, we are able to study on the causality of how network classifies different categories. To avoid network relying on texture to predict, every pair has a similar background, which makes our classifier more robust to evaluate the causal effect. Next, we analyze the properties of each pair and compare them. The number of images and instances per category for both train and validation set are shown in Fig.7 and Fig.8. In addition, the segmentation size of each pair is shown in Table.6. The segmentation size of each pair influence the performance of our classifier for

excessive noise from background increasing the difficulty to find the correct hidden features as shown in Table.7.

Table 6: Percentage of Segmentation in Images

Pair 1	**Giraffe**	**Elephant**
	14.60%	24.13%
Pair 2	**Bicycle**	**Motorcycle**
	5.74%	15.59%
Pair 3	**Stop Sign**	**Fire Hydrant**
	7.58%	7.64%

B.2. Dataset Statistics: CPS Medical

In total, 2,633 three-dimensional images (with 658 test images) were collected across multiple anatomies of interest, multiple modalities, and multiple sources (or institutions) representative of real-world clinical applications followed by COCO-CP processing. All images were identified using processes consistent with institutional review board polices at each contributing site. We reformatted the images to reduce the need for specialized software packages for reading to encourage use by specialists in medical imaging for high-level feature reasoning.

Human Evaluation

Chest MRI can provide important features to diagnose lung problems such as a tumor or pleural disorder, blood vessel problems, or abnormal lymph nodes. We collaborate with three **board-certified thoracic surgeons** to review the activate region generated by guided grad-CAM [75] on the test images. The surgeons individually retrospectively reviewed and labeled each study from the generated 100 image results as a DICOM file as consistent or inconsistent saliency compared with their diagnosis using the PACS system. The radiologists have averaged 6.43 years of experience on average, ranging from 5 to 16 years.

The TIT-generated saliency results also attain the highest consistency (61.2%) from thoracic surgeons compared with the results from VAE_{Res} (48.1%) and CEVAE^{*}_{Res} (58.8%). The consistency from a randomly generated saliency map is only (3.2%).

B.3. Ablation Study

Starting from a ResNet, we modified the architecture towards proposed TIT and compare the accuracy performance of various architecture. Table. 9 shows the impact of each change of the architecture on COCO_{CP} classification. Among all variation, attention mechanism is the most important feature, while having bilinear fusion (BF) is also more effective than concatenation.

C. Parameter and Architecture

C.1. Adversarial Perturbation

With recent security concerns of adversarial example over visual recognition, we also made a broad study on the accuracy and causal effect under adversarial examples. Fast Gradient Sign Method (FGSM) [19] is a classical gradient-based adversarial noise to generate adversarial examples by one step gradient update along the direction of the sign of gradient at each pixel by:

$$X_{\text{Adversarial}} = X + \varepsilon \cdot \text{sign}\left(\nabla_X J(X, Y)\right), \quad (14)$$

where J is the training loss (e.g. cross entropy) and Y is the groundtruth label for X. We adopt FGSM as an visual modification with $\varepsilon = \mathbf{0.3}$ ℓ_∞ perturbation constraint. This treatment could be further extended on other adversarial examples combined with causal analysis [92]. Instead of FGSM, we also study the accuracy performance under Carlini-Wagner attack (C&W) [9] and projected gradient descent (PGD) [47] as treatment. As shown in Table 8, our proposed TIT attains higher accuracy and less accuracy degradation in FGSM, C&W and PGD settings compared to CVAE' and CEVAE' for CPS classification.

C.2. Overparameterization

For a fair comparison, we study the performance of architectures with similar number of parameters. To align with the number of parameter in TIT, We modify the number of Resblocks in CVAE' as 4 and add attention mechanism to CEVAE'. As shown in Table 10, with similar number of parameters, our proposed TIT acquire the highest accuracy and better utilize the power of more parameters to compete with the state-of-art CVAE' architecture.

C.3. Different Mask Size

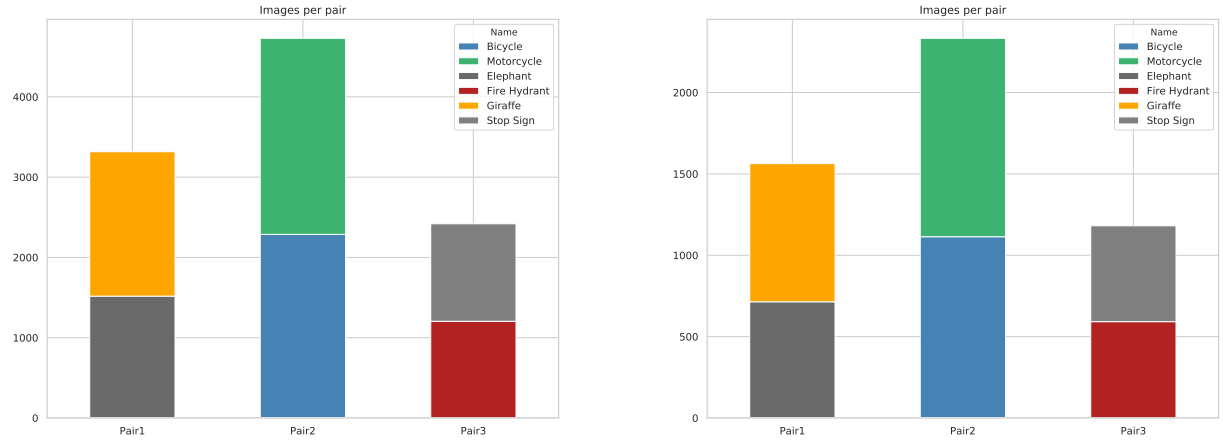
We study the effect of different mask sizes with same intervention flipping rate. The object-masking and background-refilling are used as visual perturbation in the experiments. To observe the effect, we gradually increased the mask ratio among the target object. The results in Table 12 and Table 14 show the impact of changing the ratio to the accuracy of CPS general and medical dataset classification. We find the accuracy drops as the ratio increasing, while our proposed TIT is relatively resilient to the high noise ratio scenario and perform better classification.

C.4. NICO Dataset Settings

NICO dataset provides several settings to simulate the Non-I.I.D dataset on different levels. 4 typical settings to generate Non-I.I.D training and testing subset.

Minimum Bias

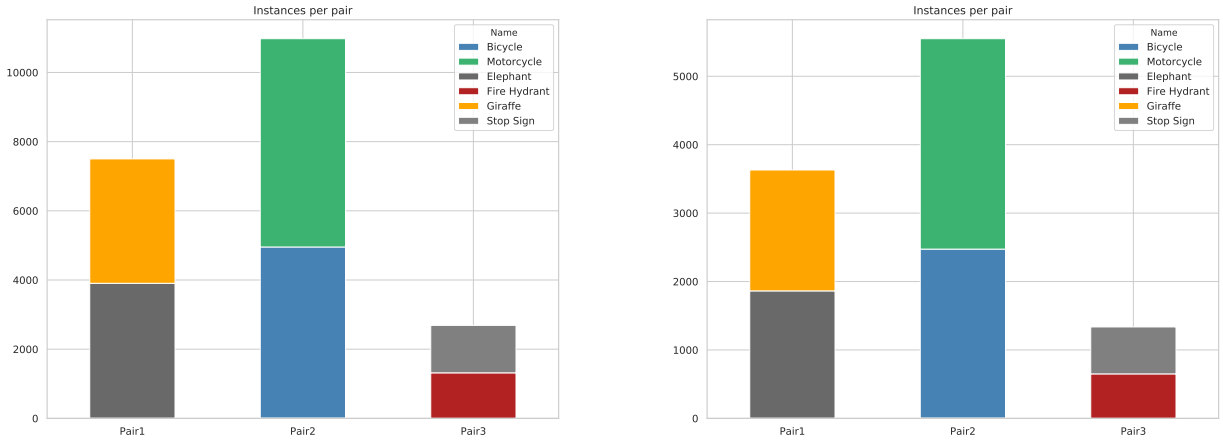
The setting choose the images in target class as positive samples and images in other classes as negative samples ignoring



(a) Train Set

(b) Validation Set

Figure 7: **Image** details per pairs in CPS: CPS_1 – Giraffe-Elephant, CPS_2 – Bicycle-Motorcycle, and CPS_3 – Stop sign-Fire hydrant.



(a) Train Set

(b) Validation Set

Figure 8: **Instance** details per pairs in CPS: CPS_1 – Giraffe-Elephant, CPS_2 – Bicycle-Motorcycle, and CPS_3 – Stop sign-Fire hydrant.

the context, which could lead to a minimum distribution shift in training and testing subset.

Proportional Bias

The setting takes all context into consideration but the ratio of each context are different in training and testing subset. In this setting, the level of distribution shift can be adjusted based on the difference of context ratio.

Compositional Bias

In this setting, the contexts exist in test subset are not guaranteed to exist in training subset. The distribution shift is higher between training and testing set. The shift could be enhanced by adding proportional bias.

We carefully observe the effect of different settings imposing on different model. The result in Table. 11 shows our proposed TLT performs better in all settings compared to other models including the CNBB model proposed by [23].

Table 7: **CPS General:** performance for common object [38] causal pairs with different visual treatments.

	Treatment	CVAE'	CEVAE$_{att}$	TLT
Bicycle & Motorcycle	Object Masking 0.0	78.31 ±0.17	80.79 ±0.06	**83.05** ±0.11
	Object Masking 0.5	74.98 ±0.09	79.46 ±0.12	**80.51** ±0.16
	Object Masking 1.0	71.65 ±0.23	72.85 ±0.13	**73.29** ±0.08
	Background Refilling 0.5	75.28 ±0.15	77.5 ±0.27	**78.68** ±0.20
	Background Refilling 1.0	71.11 ±0.42	**74.49** ±0.38	73.95 ±0.41
Stop Sign & Fire Hydrant	Object Masking 0.0	74.59 ±0.29	75.79 ±0.26	**77.41** ±0.19
	Object Masking 0.5	72.28 ±0.10	73.91±0.05	**74.08** ±0.08
	Object Masking 1.0	68.67 ±0.34	**71.22** ±0.28	71.06 ±0.24
	Background Refilling 0.5	69.13 ±0.16	73.79 ±0.21	**75.45** ±0.14
	Background Refilling 1.0	65.62 ±0.47	66.65 ±0.37	**68.24** ±0.44
Elephant & Giraffe	Object Masking 0.0	93.72 ±0.25	93.53 ±0.28	**94.67** ±0.20
	Object Masking 0.5	90.14 ±0.11	93.01 ±0.19	**93.15** ±0.09
	Object Masking 1.0	80.12 ±0.19	82.73 ±0.21	**83.06** ±0.11
	Background Refilling 0.5	90.44 ±0.08	91.71 ±0.11	**91.73** ±0.10
	Background Refilling 1.0	81.32 ±0.28	82.59 ±0.29	**83.91** ±0.17

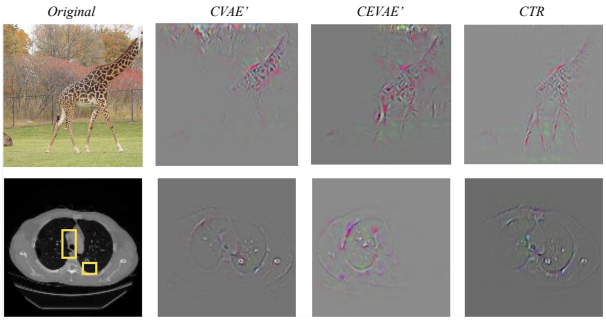


Figure 9: We use class activation mapping methods [75, 96] to explain our medical classification model. The yellow bounding box is ground truth label from the Decathlon [78] dataset. The guided-grad CAM method shows a highest false-negative scores on the region of interest.

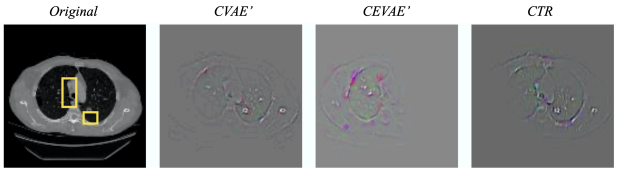


Figure 10: We show Guided-Grad-CAM results on different classification model. The results generated from CAM is much matching yellow bounding box from the Decathlon [78] dataset in the test dataset.

Table 8: Accuracy performance under adversarial attack as the treatment. TIT attains higher accuracy in both FGSM, C&W, and PGD settings.

Method	CVAE'	CEVAE'	TLT
FGSM [19]	91.92±0.11	92.02±0.11	**92.86**±0.12
C&W [9]	82.32 ±2.34	74.23±4.18	**88.12**±1.26
PGD [47]	74.32 ±1.38	86.34±1.70	**89.43**±1.08

Table 9: Model architecture ablation study in COCO$_{CP}$

Architecture	**Val. Acc. (%)**
ResNet	81.23±0.12
ResNet + CVAE = CVAE'	82.31±0.13
ResNet + CEVAE = CEVAE"	82.17±0.24
CEVAE + BF - bernoulli = CEVAE'	82.68±0.15
Treatment Learning Transformer (TLT)	**84.32**±0.07

Table 10: Overparameterization ablation study in COCO$_{CP}$

Model	Para.	Val. Acc. (%)
CVAE'	4.03M	82.31±0.13
CVAE' + 2 Resblocks	5.92M	82.38 ±0.18
CVAE' + 4 Resblocks	**7.83M**	81.96 ±0.19
CEVAE' + Attention$_C$	7.81M	83.62 ±0.21
TLT (ours)	7.39M	**84.92**±0.07

D. Reproducibility

D.1. Hyper-Parameters and Experiment Setup

Causal **E**ffect **A**utoencoder[44] (**CEVAE**$^*_{Res}$) baseline: To empower CEVAE for the visual data, our input images use dim(C)=3, dim(X)=128, dim(Y)=128. The encoder part of VAE model utilized in paper takes the ResNet34 as feature extractor. Then we sample the $q(t|x)$ by Bernoulli distribution, and $q(y|x,t)$ and $q(z|x,y,t)$ are sampled by densely connected hidden layer of 512 neurons. Sequentially, the Z is generated by reparameterization from $q|t$. The decoder starts from 3 ResBlocks with 512 width for the

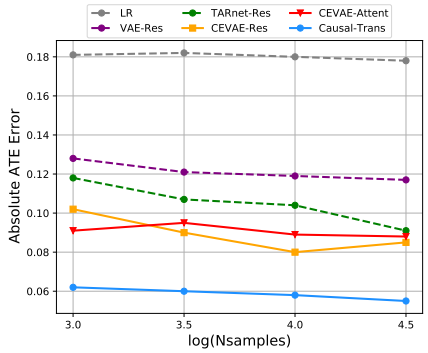


Figure 11: Error bar of average treatment effect

Table 11: Classification accuracy (%) on NICO dataset with different model and setting. Setting 1,2 and 3 refers to minimum bis, proportional bias and compositional bias mentioned in C separately.

Setting	CNBB [23]	CVAE'	CEVAE'	TLT
Setting 1	42.96 ±1.54	48.72 ±2.01	53.94 ±1.74	**57.02** ±1.42
Setting 2	44.15 ±1.48	50.10 ±1.98	54.33 ±1.64	**58.75** ±1.44
Setting 3	45.16 ±1.52	50.23 ±2.12	56.17 ±1.82	**60.98** ±1.52

dim(Z)=512 to reconstruct the $p(x|z)$, we further used 5 upsample blocks with 2 times scaling up and convolution layers with [512,256,128,64,32] width. For the last convolution layer, we use reflection padding with width set as 3. Also, we sample the $p(t|z)$ and $p(y|t,z)$ by projecting t and $\mu_y(t)$ through adaptive pooling and the densely connected hidden layer with 512 width.

CEVAE with Attention baseline (CEVAE$_{Att}$): For a much fair comparison with proposed Treatment Learning Transformer (TLT), we apply the dual attention module in the encoder part of CEVAE to approximate the $q(z|x,t,y)$. The dual attention module consists of position and channel attention module.

- Global feature: After 1x1 convolution, the input is scaled by 4 times larger with bilinear interpolation.
- Position attention: The module outputs the position attention combined by 2 convolution layers with 64 width to calculate the attention.
- Channel attention: The module outputs the channel attention by fusing the channel into spatial information and pass the feature to 2 convolution layers with 64 width.
- Combination: The output has the channel of 512, the same width as the input.

The weights of the network are initialized with weights from a model pre-trained on ImageNet. The Adam algorithm with standard parameters and learning rate **0.001** are utilized for optimization. We use mini-batches of size 128 and pick the models with the highest accuracy. All experiments of our model are implemented in PyTorch using an NVIDIA GeForce GTX 2080 Ti GPU with 12GB memory. The training time for each MS-COCO [38] causal pair with different visual treatment takes one hour to two hours on average. The reproducible code of CAN networks and a causal graphical model have been provided in the supplementary and will be open source[2].

D.2. Cognitive Response to Attention Mechanism

Cognitive psychology and neuroimaging [12] studies have found a distinct neural response to the different visual scene, as the visual attention mechanism [45]. Attention exercises, Luck et al., [45] have been proved to be enhanced learning capacities by executive control and transferring to cognitive abilities. Since images from different categories vary systematically in their visual properties as well as their semantic category, variation in visual property may influence our cognitive process of visual stimuli. The human brain has the ability to distinguish the visual scenes from different categories when categorical perception is impaired. For example, scrambling and masking are used widely when experimenting with visual pattern sensitivity. Although these perturbation preserved many of their visual characteristics, perception of scene categories was severely impaired, which makes scrambling and masking suitable metrics to compare the visual perception process between neural network and the human brain. These experiments [35, 45] have been validate on adding attention training for a improved learning performance in human education.

D.3. Using Saliency Map to Associate Learned Causal Patterns

To better understand the learned causal patterns from **TLT**, we use class activation mapping [96] (CAM) to study [56, 58, 59, 92] the causal patterns. CAM removes all fully-connected layers at the end, and including a tensor product (followed by softmax), which takes as input the global-average-pooled convolutional feature maps, and outputs the probability for each class. To obtain the class-discriminative localization map, Grad-CAM computes the gradient of $\mathbf{yc}$ (score for class $\mathbf{c}$) with respect to feature maps $\mathbf{A}$ and importance weights $\alpha_{\mathbf{k}}^{\mathbf{c}}$ of a convolutional layer. Similar to CAM, Grad-CAM [75] heat-map is a weighted combination of feature maps, and followed by a ReLU:

$$L^{c}_{\text{GradCAM}} = \text{Re}\, LU\left(\sum_{k} \alpha_k^c A^k\right) \tag{15}$$

[2]Please follow the readme in the supplementary open-source code for more information

Table 12: Accuracy (%) of varying mask sizes when intervention flipping rate $\mathbf{n} = 0.05$. $\mathbf{I_{OM}}$/$\mathbf{I_{BR}}$ denotes Object-Masking/Background-Refilling. The number $X\%$ means masking $X\%$ of a target object. Our proposed method (TLT) maintains relatively high accuracy as X increases.

$\mathbf{I_{OM}}$	CVAE'	CEVAE$_{att}$	TLT	$\mathbf{I_{BR}}$	CVAE	CEVAE$_{att}$	TLT
10%	93.31 ±0.17	92.58 ±0.21	**94.32**±0.08	10%	93.04 ±0.16	94.25 ±0.18	**94.65** ±0.09
30%	91.19 ±0.15	93.37 ±0.19	**94.13**±0.07	30%	91.27 ±0.21	93.53 ±0.23	**94.25** ±0.11
50%	90.14 ±0.11	93.01 ±0.19	**93.15** ±0.09	50%	90.44 ±0.08	91.71 ±0.11	**91.73** ±0.10
70%	86.53 ±0.13	**91.90** ±0.23	91.26 ±0.18	70%	86.62 ±0.21	88.85 ±0.18	**90.46** ±0.12
100%	80.12 ±0.19	82.73 ±0.21	**83.06** ±0.11	100%	81.32 ±0.28	82.59 ±0.29	**83.91** ±0.17

In our DNN visualization experiment, we use the state-of-the-are CAM method, guide-GradCAM [75] for comparing CVAE', CEVAE', and our TLT. guided-GradCAM fuse guided backpropagation and the Grad-CAM visualizations via a point-wise multiplication.

Interestingly, according to the intervened image after class-activation mapping techniques in Fig. 4 in the main context and Fig. 12 in the supplementary, we could find out when the area of interest are much central on the texture and edge effect. To reduce the texture dependent variable, we utilize a neural style transfer on the image set before the intervention.

Figure 12: We conduct activation saliency experiment on the COCO-CP trained by bike patterns under partial masking intervention. The guided GradCAM results show current DNNs methods often overfit on the texture and background patterns instead of desired training label(s), which echoes to previous studies [28, 63, 92].

E. Identification of Visual Causal Effect

E.1. Causality

Rubin's Causal Model (Sekhon, 2008) is a framework developed for the statistical analysis of cause of effect based on the idea of potential outcomes. Consider:

• t_i, a binary treatment $\mathbf{t}$ for individual i with 1 referring to assigning the treatment and 0 to no treatment;

• y_i is the outcome on individual i given a treatment value.

Each individual can have two potential outcomes or (counterfactuals) available as $\{y_i(1), y_i(0)\}$ $\mathbf{t}$ corresponding to receiving the treatment or not.

Identifying conceptional treatment effect Individual Treatment Effect (ITE) can be defined as the difference between the two potential outcomes for the individual; Average Treatment Effect (ATE) as the expected value of the potential outcomes over the subjects. For a binary outcome, it is defined as: given by:

$$y_i = y_i(0)\,(1-t_i) + y_i(1)t_i;\;\; ATE = \mathbb{E}\,[y_i(1)] - \mathbb{E}\,[y_i(0)]\,; \tag{16}$$

$$\begin{aligned} y_i &= y_i(0)\,(1-t_i) + y_i(1)t_i; \\ ATE &= \mathbb{E}\,[y_i(1)] - \mathbb{E}\,[y_i(0)]\,; \end{aligned} \tag{17}$$

The above mentioned metric cannot be properly estimated if there are confounding variables in the system, which will introduce bias (Greenland et al., 1999). The causal effect by a treatment variable t on an outcome y is represented by $E[y|do(t = 1)]$, where do represents the fact that the treatment has been kept at a specific value by external interventions on the system which do not affect other variables and their causal relationships in the system. Pearl defines the causal effect for a given treatment t and an outcome y and other confounding variables $\mathbf{Z}$ as:

Given a proxy X, outcome y, binary treatment t and confounder Z, we use the back-door criteria to get:

$$\begin{aligned} &P(y|X, do(t = 1)) = \\ &\int_Z P(y|X, do(t = 1), Z)P(Z|X, do(t = 1))dZ \end{aligned} \tag{18}$$

Using the intervention manipulation rules, we obtain:

$$P(y|X, do(t = 1)) = \int_Z P(y|X, t = 1, Z)P(Z|X)dZ. \tag{19}$$

Table 13: Refuting tests of the causal estimate [59, 59] with causal effect (CE) over different treatment. The validation tests show our method is confident since the common random selection (T_c) and Subset test (T_s) are closed to the original CE, and all the CE results after replacing treatment with a random (placebo) variable (T_p) are close to zero.

Treatment	**Original ATE**	**Test-Common (T_c) ↑**	**Test-Placebo (T_p) ↓**	**Test-Subset (T_s) ↑**
IS: TLT	0.288	0.288	**0.00479**	0.288
CEVAE$_{att}$	0.2948	0.2941637	0.0427	0.276
CVAE'	0.057	0.05673101	0.0385	0.0583
AT: TLT	0.036	0.035	0.012	0.035
CEVAE$_{att}$	0.027	0.0274	**0.0062**	0.024
CVAE'	0.0247	0.0242	0.01347	0.0156
SB: TLT	0.2334	0.23385	0.0253	0.238
CEVAE$_{att}$	0.2417	0.2431	0.0364	0.2353
CVAE	0.1853368	0.1853	**0.01157**	0.1834
IM: TLT	0.1855	0.1854	0.037	0.191
CEVAE$_{att}$	0.22	0.22	**0.0038**	0.1736
CVAE'	0.222	0.22285	0.0200707	0.1609
ST: TLT	0.31763	0.317641	0.0278	0.3351
CEVAE$_{att}$	0.3431	0.342221	0.0225	0.3252
CVAE'	0.354412	0.354334	**0.01127**	0.3257

Table 14: Classification accuracy (%) with error bars (with 10-fold cross-validation) comparison between different visual treatments under intervention in the **CPS** dataset.

Treatment	**CVAE'**	**CEVAE$_{att}$**	**TLT**
Object Masking 0.0	93.61 ±0.15	93.31±0.11	**94.91**±0.15
Object Masking 0.1	93.31 ±0.17	93.58±0.21	**94.32** ±0.08
Object Masking 0.3	91.19 ±0.15	93.37±0.19	**94.13** ±0.07
Object Masking 0.5	90.14 ±0.11	93.01 ±0.19	**93.15** ±0.09
Object Masking 0.7	86.53 ±0.13	**91.90**±0.23	91.26±0.18
Object Masking 1.0	80.12 ±0.19	82.73 ±0.21	**83.06** ±0.11
Background Refilling 0.1	93.04 ±0.16	94.25 ±0.18	**94.65** ±0.09
Background Refilling 0.3	91.27 ±0.21	93.53 ±0.23	**94.25** ±0.11
Background Refilling 0.5	90.44 ±0.08	91.71 ±0.11	**91.75**±0.10
Background Refilling 0.7	86.62 ±0.21	88.85 ±0.18	**90.46** ±0.12
Background Refilling 1.0	81.32 ±0.28	82.59 ±0.29	**83.91** ±0.17
Image Scrambling	59.42 ±2.19	77.3 ±1.17	**78.8** ±0.62
Style Transfer	67.73 ±2.19	68.12 ±1.21	**68.29** ±0.42
Adversarial Example	91.92 ±0.11	92.02 ±0.11	**92.86** ±0.12

The refuting test for all conditional visual model show sustainable performance to the original ATE by random common cause variable test (T_c) and random subset test (T_s) and an ideally nearby zero ATE results on replacing treatment (T_r) with a random variable test. Above validation show our CGM and its associated neural are robust and validated for causal modeling and measurement.

F. Evidence Lower Bound of VAE

To validate an Evidence Lower Bound (ELBO) of our CAN, we assume $\mathbf{p}(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ is the observed data and $\mathbf{Z}$ is the latent representation. $\mathbf{p}(\mathbf{X}, \mathbf{Z})$ can be decomposed into the likelihood and the prior as: $\mathbf{p}(\mathbf{X}, \mathbf{Z}) = \mathbf{p}(\mathbf{X}|\mathbf{Z})\mathbf{p}(\mathbf{Z})$. Using Baye's inference to calculate the posterior gives:

$$p(Z|X) = \frac{p(X|Z)p(Z)}{\int_z p(X|z)p(z)} \tag{20}$$

VAE approximates it with the family of distributions $q_\lambda(Z|X)$, where λ is the variational of parameters for the given family. We minimize the KL divergence to ensure that the approximate distribution used is close to the true

Table 15: Validation of causal effect by three causal refuting tests. The causal effect estimate is tested by random common cause variable test (T_c), replacing treatment with a random (placebo) variable (T_r – lower is better), and removing a random subset of data (T_s). TLT outperforms in most tests.

Noise : do(t)	**Measurement of ATE**			
Method	Original	w/ T_c	w/ T_p	w/ T_s
TLT	**0.2432**	**0.2431**	**0.0114**	**0.2481**
CEVAE'	0.2414	0.2414	0.0248	0.2329
CVAE'	0.1792	0.1763	0.0120	0.1751

posterior:

$$KL\left(q_\lambda(Z|X)\|p(X|Z)\right) = \mathbb{E}_q\left[\log\left(q_\lambda(Z|X)\right)\right] - \mathbb{E}_q[\log p(X,Z)] + \log p(X). \tag{21}$$

The posterior for inference network will be :

$$q^*_\lambda(Z|X) = \arg\min_\lambda KL\left(q_\lambda(Z|X)\|p(X|Z)\right). \tag{22}$$

However, due to the occurrence of $\mathbf{p}(\mathbf{X})$, the KL is still intractable. We can manipulate the above equation by defining the ELBO:

$$\begin{aligned} ELBO(\lambda) &= \log(p(X)) - KL\left(q_\lambda(Z|X)\|p(X|Z)\right) \\ &= \mathbb{E}_q[\log p(X|Z)] - KL\left(\log q_\lambda(Z|X)\|p(Z)\right). \end{aligned} \tag{23}$$

Then, the negative of the ELBO is the loss function used for the neural networks:

$$\begin{aligned} l(\theta,\phi) = &- \mathbb{E}_{q_\theta(z|x)}\left[\log p_\phi(X|Z)\right] \\ &+ KL\left(\log q_\theta(Z|X,\lambda)||p(Z)\right). \end{aligned} \tag{24}$$

θ and ϕ, are the weights and biases of the DNN which are chosen to maximize the ELBO using gradient descent algorithm.

Training Objective of Treatment Inference Transformer. In the TLT setting, where the architecture is adapted from TARnet [76]'s inference network, i.e., split input for each treatment group in t after a shared representation, the objective function $\mathcal{L}$ is given by:

$$\begin{aligned} \mathcal{L} = &\sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i,t_i,y_i)}\left[\log p\left(\mathbf{x}_i,t_i|\mathbf{z}_i\right) + \log p\left(y_i|t_i,\mathbf{z}_i\right)\right] \\ &+ \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i,t_i,y_i)}[\log p\left(\mathbf{z}_i\right) - \log q\left(\mathbf{z}_i|\mathbf{x}_i,t_i,y_i\right)] \end{aligned} \tag{25}$$

For predicting new subject predictions, the treatment assignment t along with outcome y are required. We have introduced Bernoulli distributions which help predict y and t (a binary index of treatment) for new samples with the theoretical foundation from CEVAE [44]. We then leverage bilinear fusion for $q(z|x,y,t,a)$ instead of concatenation [44] and remove Bernoulli sampling for classification label inference. The attention decoding $p(a|x,q(y)) = q(a)$ is incorporating with the known treatment for training.

G. Future Work

Treatment Learning for Video Clip Classification.

The binary requirement [20, 71, 76] of treatment variable aims to model a **single independent causal condition** (e.g., either the noise is "the cause" or "not the cause" affecting visual patterns) in our causal graphical model [56] (CGM). Breaking down the binary hypothesis conflicts with treatment effect estimation and the CGM. Interestingly, it is possible to model treatment variables as a time-dependent "noise ratio" for video classification (e.g., a Gaussian noise over time frames).

As a proof of concept, we tested using TLT on a simple UCF-101 dataset [80], where the naive results show TLT remains a high accuracy of 99.6±0.2% against various mean values, as shown in Table 16. In this preliminary study, TLT also outperformed conditional transformer-based backbones [42], sharing a similar takeaway of our current focus, which may connect to our cognitive discussion in the future.

Table 16: Classification accuracy (%) on UCF-101.

Model	**Video-Trans** [42]	**CVAE'**	**TLT**
Acc.	99.1 ±0.3	98.5 ±0.4	**99.6** ±0.2

Discovering Visual Causality beyond Vision Classification Tasks.
In conclusion, we find out causal effect do exist in different DNN-based visual modification methods, and this effect could be visualized to see its effectiveness on understanding targeted DNN layer. By introducing a new extended dataset, COCO-CPs, our CAN networks show competitive visualization results and potential combined with existing saliency-based methods. For future work, we plan to extend our proposed CAN framework to discover visual causality over more visual tasks, such a video detection, cross-model adaption, and obvious question answering (VQA).