Learning Concept-Based Causal Transition and Symbolic Reasoning for Visual Planning

Yilue Qian[1,2,3*], Peiyu Yu[4], Ying Nian Wu[4], Yao Su[1], Wei Wang[1†], Lifeng Fan[1†]

***Abstract*— Visual planning simulates how humans make decisions to achieve desired goals in the form of searching for visual causal transitions between an initial visual state and a final visual goal state. It has become increasingly important in egocentric vision with its advantages in guiding agents to perform daily tasks in complex environments. In this paper, we propose an interpretable and generalizable visual planning framework consisting of i) a novel Substitution-based Concept Learner (SCL) that abstracts visual inputs into disentangled concept representations, ii) symbol abstraction and reasoning that performs task planning via the learned symbols, and iii) a Visual Causal Transition model (ViCT) that grounds visual causal transitions to semantically similar real-world actions. Given an initial state, we perform goal-conditioned visual planning with a symbolic reasoning method fueled by the learned representations and causal transitions to reach the goal state. To verify the effectiveness of the proposed model, we collect a large-scale visual planning dataset based on AI2-THOR, dubbed as *CCTP*. Extensive experiments on this challenging dataset demonstrate the superior performance of our method in visual planning. Empirically, we show that our framework can generalize to unseen task trajectories, unseen object categories, and real-world data. Further details of this work are provided at `https://fqyqc.github.io/ConTranPlan/`.**

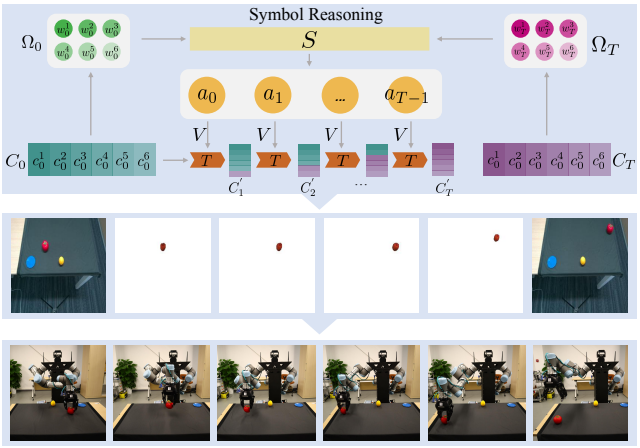

Fig. 1: **Our visual planning framework.** Given an initial state and a goal state, we aim to predict the intermediate states (in the second row) that will guide a robot to manipulate the target objects (in the third row). The disentangled concept-based representation C and abstracted symbol representation Ω, as well as their corresponding causal transition $\mathcal{T}$ and symbol reasoning S, are effectively combined into a bi-level planning framework for better generalization (in the first row).

I. Introduction

As one of the fundamental abilities of human intelligence, planning is the process of insightfully proposing a sequence of actions to achieve desired goals, which requires the capacity to think ahead, to employ knowledge of causality, and the capacity of imagination [1, 2], so as to reason and foresee the proper actions and their consequences on the states for all the intermediate transition steps before finally reaching the goal state. Visual planning simulates this thinking process of sequential causal imagination in the form of searching for visual transitions between an initial visual state and a final visual goal state. With its advantages in guiding agents to perform daily tasks in the first-person view, visual planning has become more and more important in egocentric vision [3] and embodied AI. In robotics, visual planning could also avoid manually designing the required specific goal conditions, action preconditions, and effects for robots.

Previous works for visual planning can be roughly categorized into three tracks, *i.e.*, neural-network-based models [4, 5], reinforcement-learning-based models [6, 7] and classic search-based models [8, 9]. Neural-network-based models can be trained in an end-to-end manner, which tends to fall short in terms of its interpretability [10]. Reinforcement-learning-based models can perform goal-conditioned decisions, but could suffer from sparse reward, low data efficiency [11], and low environment and task generalization ability [12]. Considering these limitations and inspired by human cognition, our method falls into the third "search-based" category and further proposes three key components for visual planning, namely **representation learning, symbolic reasoning, and causal transition modeling**. Representation learning focuses on extracting objects' dynamic and goal-oriented attributes. Symbolic reasoning performs action planning at the abstract higher level via learned symbols. Causal transition models the visual preconditions and action effects on attribute changes.

At the **perception** level, we propose to learn concept-based disentangled representation and believe such human-like perception ability to abstract concepts from observations is vital for visual causal transition modeling [13]. The reason is that such representation learning could encode images at a higher semantic level beyond pixels, identifying distinct attribute concepts and isolating "essential" factors of variation, thus serving causal learning [14]. This also enhances both robustness and interpretability [14–16], and facilitates compositional generalization to unseen scenarios in zero-shot inference [14, 17–19], thereby supporting a wide range of real-world downstream tasks. At the **reason-**

† Corresponding authors: {wangwei, lifengfan}@bigai.ai
* This work was done during Yilue Qian's internship at BIGAI. [1]State Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI). [2]Institute for Artificial Intelligence, Peking University. [3]Yuanpei College, Peking University. [4]Department of Statistics, University of California, Los Angeles (UCLA).

ing and planning level, we argue that understanding the atomic causal mechanisms is inevitable for task planning. Leveraging learned disentangled representations of concepts, we gain insight into the core of atomic causal transitions, which involves identifying key relevant variable concepts and predicting the outcomes of actions executed upon them. The understanding and reasoning of the abstract higher-level task planning composed of the lower-level atomic causal transition also have the potential to be more generalizable and interpretable [20, 21]. Thus, we propose a visual causal transition model (ViCT) and its abstracted symbolic transition model, which corresponds to the discrete higher-level task planning and avoids the problem of "error accumulation" [22]. Guided by symbolic transition, the visual transition reconstructs intermediate and final goal images.

Technically, there are **three critical modules** in our visual planning framework. First, a novel Substitution-based Concept Learner (SCL) (Sec. III-A) is learned by switching the latent concept representations of a pair of images with different attribute concepts. Second, a set of state symbols is abstracted from clustering low-level concept token representations (Sec. III-B). The most efficient symbolic transition path can be found via a Markov Decision Process (MDP). Third, a visual transition model (Sec. III-C) is proposed to learn the action-induced transition of the changeable attributes given the concept representations of the precondition image and thus generate the resulting effect image. To verify the effectiveness of our framework, we collect a large-scale visual planning dataset, which contains a concept learning dataset and a causal planning dataset. Extensive comparison experiments and ablation studies on this dataset demonstrate that our model achieves superior performance in the visual planning task and various forms of generalization tests.

To summarize our **main contributions**: (i) We propose a novel concept-based visual planning framework, which models both discrete symbolic transition and continuous visual transition for efficient path search and intermediate image generations. Comprehensive experiments show that our method achieves superior performances in visual task planning and generalization tests. (ii) Apart from generalizability, our method has better interpretability by generating a causal chain (the action sequences and the intermediate state images) to explicitly demonstrate the transition process to the goal. (iii) We collect a large-scale visual planning dataset, which can foster future research in the community.

A. Related Work

Visual planning is feasible with the learned representation and atomic causal effects. [23] proposed a method for long-horizon deformable object manipulation tasks from sensory observations, which relies heavily on differentiable physics simulators. [8] performed a tree-search-based planning algorithm on the learned world representation after applying high-level actions for visual robot task planning, but they ignored learning disentangled representations. [4] learned how to plan a goal-directed decision-making procedure from real-world videos, leveraging the structured and plannable latent state and action spaces learned from human instructional videos, but their transformer-based end-to-end model is hard to generalize to unseen planning tasks. [5] proposed a model based on deep neural networks consisting of encoding, action-conditional transformation, and decoding for video prediction in Atari Games, but they do not abstract symbols for efficient reasoning. [24] is the most similar to ours, which learned symbolic operators for task and motion planning, but cannot generate intermediate images.

Concept-based disentangled representation learning has emerged as a popular way of extracting human-interpretable representations [25]. Discrete and semantically-grounded representation is argued to be helpful for human understanding and abstract visual reasoning, enables few-shot or zero-shot learning and leads to better down-stream task performance [26, 27]. Previous studies tried to learn disentangled concept representation either in a completely unsupervised manner [28–31], or via weak supervision and implicit prototype representations [32], or by employing supervision from the linguistic space [33, 34]. There have been diverse learning techniques, such as Transformer [31], (sequential) variational autoencoder [29, 30], and information maximizing generative adversarial nets [28], etc. Existing techniques have proved successful on objects mostly with limited variation, such as digits, simple geometric objects [32], and faces [28]. In this work, we propose a variant of [31] by imposing more reconstruction constraints, which works very well on more complex household objects with diverse variations (Sec. II) and better benefits the downstream planning task compared to prior works.

Causal reasoning for task understanding is one of the essential capabilities of human intelligence, and a big challenge for AI with the difficulty of generating a detailed understanding of situated actions, their dependencies, and causal effects on object states [35]. Various evaluated state-of-the-art models only thrive on the perception-based descriptive task, but perform poorly on the causal tasks (*i.e.*, explanatory, predictive, and counterfactual tasks), suggesting that a principled approach for causal reasoning should incorporate not only disentangled and semantically grounded visual perception, but also the underlying hierarchical causal relations and dynamics [36]. [37] built a sequential Causal And-Or Graph (C-AOG) to represent actions and their effects on objects over time, but suffers from ambiguity in real-life images due to their not-well-disentangled representation. Our work benefits from our disentangled concept representation by finding a latent space where important factors could be isolated from other confounding factors [17], and we ground actions to their causal effects on relevant object attributes. Our bi-level causal planning framework with discrete symbolic transition and continuous visual transition also helps to resist the real-world data noises and ambiguity.

II. Environment & Dataset

To facilitate the learning and evaluation of the concept-based visual planning task, we collect a large-scale RGB-D image sequence dataset named *CCTP* (Concept-based Causal Transition Planning) based on AI2-THOR simulator [38]. We exclude scene transitions in each task by design to focus more on concept and causal transition learning, *i.e.*, each task is performed on a fixed workbench, although

the workbenches and scenes vary from task to task. The frame resolution is 384×256, converted into 256×256 at the beginning of our method. The whole dataset consists of a concept learning dataset and a visual causal planning dataset, which we will illustrate in detail below.

A. Concept Learning Dataset

We learn six different kinds of concepts: `TYPE`, `POSITION_X`, `POSITION_Y`, `ROTATION`, `COLOR`, and `SIZE`. `TYPE` refers to the object category. The dataset has eight different types of objects in total, including *Bread*, *Cup*, *Egg*, *Lettuce*, *Plate*, *Tomato*, *Pot*, and *Dyer*, all of which can be manipulated on the workbench. We manually add the `COLOR` concept to the target object by editing the color of the object in its HSV space. This leads to 6 different colors for each object, and 20 samples are provided for each color to avoid sample bias. For `SIZE` concept, we rescale each target object to 4 different sizes as its concept set. As for the position, we use `POSITION_X` and `POSITION_Y` to refer to the coordinates along the horizontal X-axis and the vertical Y-axis w.r.t. the workbench surface. We discretize `POSITION_X` with 3 values and `POSITION_Y` with 5. Notably, changes in `POSITION_X` and `POSITION_Y` also cause variant perspectives of an object. For `ROTATION`, we set 0, 90, 180 and 270 degrees for all types of objects. We exhaustively generate all possible target objects with different value combinations of the six concepts, resulting in 234,400 images. Leveraging the masks provided by AI2-THOR, we isolate the foreground images, containing only the target object with a black background. We randomly choose 40% of the concept combinations for training. For each image $X_{0,f}$ in the training set and each concept index i, we search for image $X_{1,f}$ within the training set such that $X_{0,f}$ and $X_{1,f}$ differ only in the i-th concept. We use such paired images and the corresponding label i for concept learning.

B. Causal Planning Dataset

A causal planning task consists of several steps of state transitions, each caused by an atomic action. We define seven different atomic actions in our dataset, including `move_front`, `move_back`, `move_left`, `move_right`, `rotate_left`, `rotate_right`, and `change_color`. The magnitude of each action is fixed. The target object states (*e.g.*, its color) are randomly initialized in each task from our dataset. The task lengths (*i.e.*, the number of steps for each task) are not fixed. We collect four subsets of tasks, each representing a difficulty level. In the first level, the workbench has no obstacles, and the ground truth actions involve only movements. In the second level, several fixed obstacles appear on the workbench. In the third level, a dyer additionally appears on the workbench, and the target object must be moved adjacent to the dyer to change its color if necessary before moving to the target position. In the fourth level, rotation actions are involved additionally. The action sequence in each task is paired with the corresponding visual observations. The maximum task lengths for each subset are 6, 9, 15, and 16, and the average are 2.67, 2.81, 4.66, and 5.64, respectively. Each subset contains 10,000 tasks: 8,000 for training, 1,000 for validation, and 1,000 for testing.

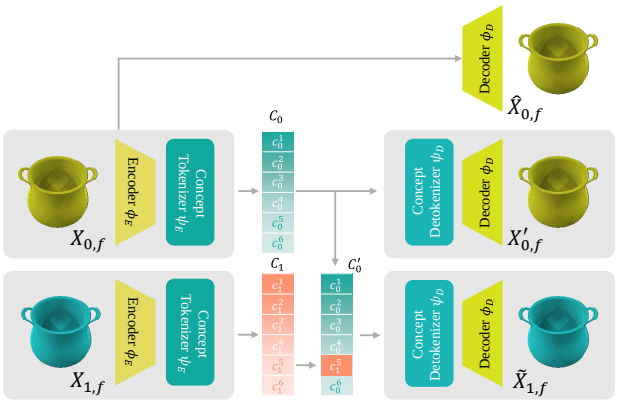

Fig. 2: **Architecture of SCL**. Foreground images $X_{0,f}$ and $X_{1,f}$ differ only in the `COLOR` concept. After extracting their concept tokens and assuming the token c_i^5 to represent the color concept, the `COLOR` concept c_0^5 of $X_{0,f}$ is substituted by c_1^5 from $X_{1,f}$, which are then fed into the detokenizer and decoder to reconstruct images.

We construct additional generalization test benchmarks based on our collection. We provide four levels of **Unseen Object** generalization tests for object-level generalization. For each level, we generate 1000 tasks with the target object types unseen in the training dataset, including object types of *Cellphone*, *Dish Sponge*, *Saltshaker*, and *Potato*. Additionally, we provide datasets for generalization tests for unseen tasks. The training and testing tasks in the **Unseen Task** dataset have different combinations of action types. For example, the training dataset may include tasks that consist of only `move_left` and `move_front` actions, as well as tasks that consist of only `move_right` and `move_back` actions, while the testing dataset contains tasks from the held-out data with different combinations. The **Unseen Task** dataset is limited to level-1 and level-2 because limited combinations of actions are insufficient to accomplish more difficult tasks.

III. Method

Given an initial RGB-D state image X_0 and a final RGB-D state image X_T, our task is to find a valid and efficient state transition path with an inferred sequence of actions $\mathbf{\Gamma} = \{a_t\}_{t=1,\dots,T}$, as well as generating intermediate and final state images $\tilde{\boldsymbol{X}} = \{\tilde{X}_t\}_{t=1,\dots,T}$. To fulfill this task, we use a concept learner to extract disentangled concept representations for state images, abstract concept symbols for reasoning, and train a ViCT to generate intermediate state images.

A. Substitution-based Concept Learner

The architecture of our SCL is illustrated in Fig. 2. A pair of foreground images $X_{0,f}$ and $X_{1,f}$ are given as input, where these two images contain two objects differing only in one concept, *e.g.*, a yellow pot and a green pot. Then a shared encoder ϕ_E is applied to the foreground images to obtain the latent embeddings $Z_{i,f} = \phi_E(X_{i,f})$. The embedding $Z_{i,f}$ is further fed into a concept tokenizer ψ_T to generate the concept tokens $C_i = \{c_i^k\}_{k=1,\dots,6} = \psi_T(Z_{i,f})$. Here k is the concept index, and we assume there exist six

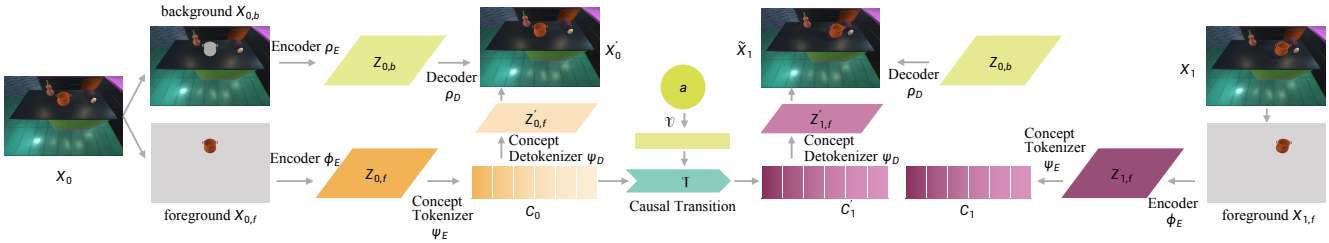

Fig. 3: **Architecture of ViCT**. The concept tokenizer extracts object concept tokens for causal transition. The causal transition model transforms concept tokens from C_0 to $C_1^{'}$ with the action embedding $\mathcal{V}(a)$. The background encoder converts the background image into latent vectors, which are then combined with predicted concept tokens $C_1^{'}$ to generate the effect image $\tilde{X}_1$.

visual concepts, *i.e.*, `TYPE`, `POSITION_X`, `POSITION_Y`, `ROTATION`, `COLOR`, and `SIZE`, representing the visual attributes of the target objects (refer to Sec. II-A for details).

The concept token c_0^i is substituted with c_1^i to get a new concept token vector $C_0^{'}$, where i indexes the different concept between the paired images $X_{0,f}$ and $X_{1,f}$. For example, the token c_i^5 assumes to represent the `color` concept in Fig. 2, so replacing c_0^5 with c_1^5 will change the original yellow pot to a green pot. The token vector $C_0^{'}$ is fed into a concept detokenizer ψ_D to reconstruct the latent embedding $Z_{1,f}^{'} = \psi_D(C_0^{'})$, which is further decoded into image $\tilde{X}_{1,f} = \phi_D(Z_{1,f}^{'})$. After the concept detokenizer and decoder, we obtain a combined reconstruction loss as follows:

$$\mathcal{L}_1 = \mathcal{L}_{MSE}(X_{0,f}^{'}, X_{0,f}) + \mathcal{L}_{MSE}(\tilde{X}_{1,f}, X_{1,f}), \quad (1)$$

where $\mathcal{L}_{MSE}$ is the mean squared error. In addition, we add another branch that directly connected the encoder to the decoder. This branch aims to distinguish the role of the encoder from that of the concept tokenizer; it enforces the encoder to learn hidden representations by reconstructing $X_{0,f}$. The reconstructed image and reconstruction loss of this branch are $\hat{X}_{0,f}$ and $\mathcal{L}_{MSE}(\hat{X}_{0,f}, X_{0,f})$, respectively. Similar to [31], a Concept Disentangling Loss (CDL) is employed to reduce interference between the concept tokens. The CDL can be formulated as follows:

$$\mathcal{L}_{CDL} = \mathcal{L}_{CE}(\|C_0 - C_1\|_2, i), \quad (2)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss. $\|C_0 - C_1\|_2$ calculates the l_2 norm of the variation of each concept token. i is the ground-truth token index and indicates that the i-th concept token is replaced. The total loss $\mathcal{L}_C$ of concept learner is as follows:

$$\mathcal{L}_C = \mathcal{L}_1 + \mathcal{L}_{MSE}(\hat{X}_{0,f}, X_{0,f}) + \mathcal{L}_{CDL}, \quad (3)$$

where the equal weights for each loss work well in our experimental settings.

B. Symbol Abstraction and Reasoning

Symbol abstraction aims to convert concept tokens into discrete symbols for later symbolic reasoning. Our empirical results in Fig. 6 show that the concept tokens learned in Sec. III-A are well-disentangled and can be easily clustered into several categories. Therefore, a clustering algorithm could be applied to the concept tokens to generate symbols. Specifically, we collect all the concept tokens extracted from the training data using the SCL and create the concept token spaces: $\mathbf{C} = \{c_n\}$. Then, we employ the K-means algorithm to cluster data points within the concept spaces, resulting in the concept centers $\{\bar{c}\}$ and a symbol assignment $\omega = \sigma(c, \{\bar{c}\})$ for each concept token c. Here, σ is the nearest neighbor function, which assigns the symbol of the nearest concept center to c. This procedure is independently applied to six defined concepts, with each concept assigned a specific number of clustering centers that correspond to their predefined value spaces, abstracting a set of concept symbols $\Omega = \{\omega^k\}_{k=1,\ldots,6}$ for each image.

The symbolic reasoning aims to find the most plausible transition path from the initial state to the goal state at the symbol level, which can be formulated as an MDP. Given the initial concept symbols $\Omega_0 = \{\omega_0^k\}_{k=1,\ldots,6}$ and the action a_0, the symbol reasoner computes the distribution of concept symbols at the next timestep $\Pr\left[\Omega_1^{'} \mid a_0, \Omega_0\right]$. The concept symbol distribution at the timestep t can be obtained as follows:

$$\begin{aligned} &\Pr\left[\Omega_t^{'} \mid a_{0:t-1}, \Omega_0\right] \\ &= \sum_{o \in \mathbf{\Omega}} \Pr\left[\Omega_t^{'} \mid a_{t-1}, \Omega_{t-1}^{'} = o\right] \cdot \Pr\left[\Omega_{t-1}^{'} = o \mid a_{0:t-2}, \Omega_0\right], \end{aligned} \quad (4)$$

where $\mathbf{\Omega}$ denotes the the entire concept symbol space. Additionally, two legality checks are implemented during the reasoning process to ensure the validity of the action sequence, involving action legality and state legality checks. The action legality is defined as $\mathbf{1}_{\Pr[a|\Omega] > \text{thresh}}$. This check aims to prevent the use of noise-inducing transformations caused by the SCL, thereby modifying Eq. (4) to:

$$\begin{aligned} &\Pr\left[\Omega_t^{'} \mid a_{0:t-1}, \Omega_0\right] \\ &= \sum_{o \in \mathbf{\Omega}} \mathbf{1}_{\Pr[a_{t-1}|o] > \text{thresh}} \Pr\left[\Omega_t^{'} \mid a_{t-1}, \Omega_{t-1}^{'} = o\right] \Pr\left[\Omega_{t-1}^{'} = o \mid a_{0:t-2}, \Omega_0\right]. \end{aligned} \quad (5)$$

The state legality check is designed to eliminate contributions to the distribution originating from invalid states (e.g., collisions with obstacles on the workbench). It can be written as follows:

$$\begin{aligned} &\Pr\left[\Omega_t^{'} = o_0 \mid a_{0:t-1}, \Omega_0; \{\Omega_{\text{env}}\}\right] \\ &= \frac{\mathbf{1}_{o_0 \in \mathbf{\Omega}_{\text{valid}}} \cdot \Pr\left[\Omega_t^{'} = o_0 \mid a_{0:t-1}, \Omega_0\right]}{\sum_{o \in \mathbf{\Omega}_{\text{valid}}} \Pr\left[\Omega_t^{'} = o \mid a_{0:t-1}, \Omega_0\right]} \end{aligned} \quad (6)$$

where $\mathbf{\Omega}_{\text{valid}} \subseteq \mathbf{\Omega}$ represents the set of valid concept symbols given the concept symbols of other objects in the environment, and o_0 is an arbitrary element within $\mathbf{\Omega}$. To

reduce computational complexity, the reasoning process is individually applied to each concept. This approach is effective due to the well-designed disentangled concepts, which ensure that the changes in each concept are independent given a particular action. The MDP estimates symbol-level transition probability distribution by recording the (input, action, output) triplets in the training data. The objective is to discover the action sequence $a_{0:T-1}$ that is most likely to result in a distribution of concept symbols $\Omega_T^{'}$ closely approximating the goal concept symbols Ω_T. This action sequence is then passed into the ViCT (See Sec. III-C) to generate predicted intermediate images (Fig. 1).

C. Visual Causal Transition Learning

The aim of ViCT is to generate visual effect images based on precondition images and human actions. For example, Fig. 3 shows an action that moves the pot one step to the right. ViCT predicts image $\tilde{X}_1$ by transforming the pot in image X_0 with a `move_right` action.

As seen in Fig. 3, three parts exist in the framework of ViCT. Firstly, the causal transition is the key part of ViCT. This process transforms object concept tokens from C_0 to $C_1^{'}$ with the help of an action embedding $\mathcal{V}(a)$. The action a is encoded into a one-hot vector and further embedded via an embedding function $\mathcal{V}$ to achieve this. The transition process is as follows:

$$C_1^{'} = \mathcal{T}(C_0, \mathcal{V}(a)), \tag{7}$$

where $C_1^{'}$ represents the resulting concept tokens. $\mathcal{T}$ denotes the causal transition function involved in this process. In addition to the causal transition component, two other crucial parts in ViCT are dedicated to managing visual extraction and reconstruction. The second part contains a concept tokenizer to extract foreground object concept tokens C_0 for later transitions. This concept tokenizer has been trained as described in Sec. III-A and fixed here. This part also involves a background encoder ρ_E, which processes the background image to produce latent vectors represented as $Z_{0,b}$. The vectors $Z_{0,b}$ store background-related information and will be used to generate the resultant image $\tilde{X}_1$. The third part combines foreground object concept tokens and background latent vectors to predict effect image $\tilde{X}_1$ with the background decoder ρ_D. Instead of directly using concept tokens, we convert them back to latent embeddings, *i.e.*, from $C_1^{'}$ to $Z_{1,f}^{'}$, and then concatenate $Z_{1,f}^{'}$ with latent vectors $Z_{0,b}$ as the input to the decoder. Similarly, we can also combine $Z_{0,f}^{'}$ and $Z_{0,b}$ to obtain a reconstruction image $X_0^{'}$.

Now two losses can be computed during training: a reconstruction loss $\mathcal{L}_{MSE}(X_0^{'}, X_0)$ and a prediction loss $\mathcal{L}_{MSE}(\tilde{X}_1, X_1)$. In addition to measuring image-level prediction errors, we can also evaluate token-level prediction errors. Given a ground-truth effect image X_1, we extract its concept tokens C_1, and introduce a token prediction loss $\mathcal{L}_{MSE}(C_1^{'}, C_1)$. The total loss of ViCT is summarized as follows:

$$\mathcal{L}_T = \mathcal{L}_{MSE}(C_1^{'}, C_1) + \mathcal{L}_{MSE}(\tilde{X}_1, X_1) + \mathcal{L}_{MSE}(X_0^{'}, X_0). \tag{8}$$

The ViCT is trained on our causal planning dataset (see Sec. II-B).

IV. Experiments

Our experiments aim to answer the following questions: (1) Is our model design effective and applicable to visual planning tasks? (2) How do the proposed key components contribute to the model performance? (3) Are the learned concepts and causal transitions interpretable? (4) Does the proposed method exhibit generalization on novel tasks? To answer these questions, we perform extensive experiments, showing the proposed methods are interpretable, generalizable, and capable of producing significantly better results than baseline methods.

A. Evaluating Visual Planning on Dataset CCTP

To validate the effectiveness of our model design, we employ PlaTe [4], the state-of-the-art method for visually-grounded planning, as our baseline. To probe the contribution of our proposed components, we replace each component with alternative baselines to compare with. We replace the proposed concept learner with strong baselines such as beta-VAE [39] and VCT [31] model to verify the effectiveness of our concept learning module. Additionally, we compare our model to a goal-conditioned Double DQN agent [40] trained with prioritized experience replay [41], noted as "w/ RL". Furthermore, to verify the necessity of our symbolization process, we apply the reasoning process directly to the concept tokens, employing our causal transition model to search for states closest to the goal state within the concept token spaces. We also conduct experiments where we further remove the concept learning process. Instead, we use an autoencoder to extract latent embedding for causal transition. The corresponding results are denoted as "w/o. symbol" and "w/o. concept", respectively. The "w/o. concept" experiments are limited to the level-1 dataset, as this method is unable to avoid object collisions in the higher-level datasets. Finally, we replace the explicit planning module with a transformer architecture. It takes the initial and goal state concept symbols, provided by our concept learner and symbolizer, as inputs to generate the action sequence. We refer to this variant as "w/o causal". We also substitute the planning module with random action predictions for each step as an additional baseline for reference.

a) Evaluation metrics: To thoroughly inspect the performance of visual planning, we design metrics including Action Sequence Prediction Accuracy (ASAcc), Action Sequence Efficiency (ASE), and Final State Distance (FSD). ASAcc is measured as the success rate of sequence prediction. In level-1 and level-2 tasks, a successful prediction entails moving the target object accurately to the position of goal states without encountering any collisions with obstacles (if present). In level-3 tasks, when the target object's color changes, success requires moving the object adjacent to the dyer, applying the `change_color` action, and then moving it to the goal position. In level-4 tasks, the target object must also be correctly rotated for success. During testing, MDP and search-based methods, including "Ours", "Ours w/ β-VAE", "Ours w/o symbol", and "Ours w/o concept" generate the 5 most possible paths, and randomized algorithms "Chance" and "Ours w/ RL" make 5 attempts for each task. The top-1 accuracy evaluates the success of the most

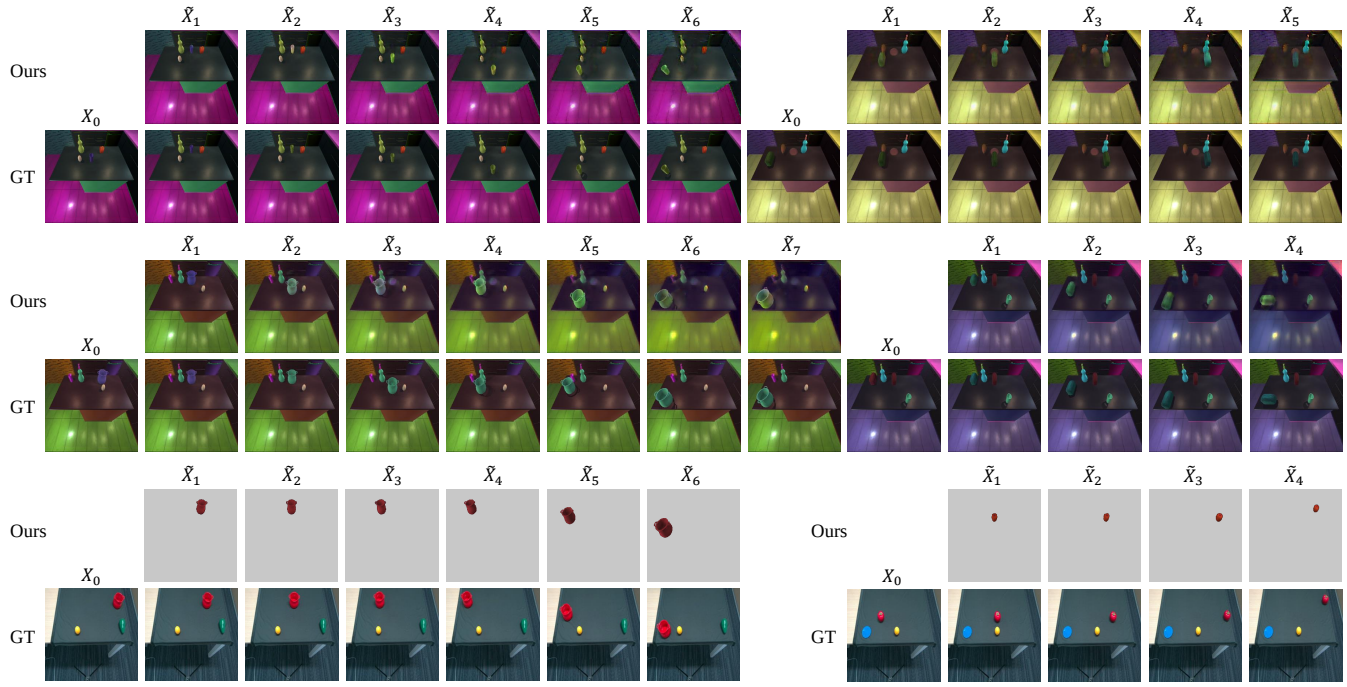

Fig. 4: **Qualitative results of our visual planning model**. The top two samples are obtained from the level-3 dataset, and the middle two are from the level-4 dataset. Our model demonstrates its ability to manage tasks of varying lengths, effectively plan action sequences, and generate intermediate and goal state images. Notably, the first sample from the level-4 dataset generates a different path than the ground truth but still achieves success and maintains high efficiency. The bottom two samples are from our real-world data experiments, corresponding to the level-3 dataset. To simplify implementation, we focus on visually planning the target objects in the real-world images, ignoring the background.

likely path or the initial attempt, while the top-5 accuracy checks if any of the 5 paths are successful. ASE measures the efficiency of the planning by comparing the length of the ground truth sequence to that of the predicted sequence. It only considers the successfully predicted sequences. The ASE is defined as follows:

$$ASE = \frac{\sum_{i=1}^{N} \mathbb{I}(\mathbf{\Gamma}_i^{pred}) \ell(\mathbf{\Gamma}_i^{gt}) / \ell(\mathbf{\Gamma}_i^{pred})}{\sum_{i=1}^{N} \mathbb{I}(\mathbf{\Gamma}_i^{pred})}, \tag{9}$$

where $\mathbb{I}$ is a indicator function for a successful prediction, ℓ represents the length of an action sequence. Of note, the ground truth action sequences in *CCTP* are the most efficient, so the efficiency of a predicted sequence will be no more than 1. FSD calculates the distance between the positions of the foreground object in the final predicted state and in the goal state. The distance is defined based on the object's coordinates w.r.t. the workbench.

b) Results: We can see from Tab. I that the proposed method achieves significantly higher performance compared with baselines. Specifically, we compare our method with different ablative variants on *CCTP* dataset. Our method outperforms baselines in terms of ASAcc by a large margin and achieves the smallest FSD, which demonstrates our method can obtain an accurate planning path to reach the goal state. Our method achieves very competitive ASE. Notably, certain baselines (*e.g.*, Ours w/ RL) attain high levels of ASE, but with a disproportionately lower ASAcc. Moreover, our model maintains strong performance when encountering hard tasks, while competitive baselines' performances significantly decrease as task difficulty increases.

TABLE I: **Quantitative results for visual task planning.** The best scores are marked in **bold**.

Model ID	ASAcc.(%)(↑)		ASE(↑)	FSD(↓)	ASAcc.(%)(↑)		ASE(↑)	FSD(↓)
	Top-1	Top-5			Top-1	Top-5		
	Dataset level-1				Dataset level-2			
Chance	1.3	7.3	-	3.139	0.4	2.2	-	3.499
PlaTe [4]	38.9	-	-	-	15.3	-	-	-
Ours w/ β-VAE [39]	0.5	3.0	0.970	3.220	0.0	3.5	-	3.670
Ours w/ VCT [31]	54.1	60.6	0.972	1.483	1.6	4.9	0.988	1.294
Ours w/o symbol	65.8	76.9	0.983	1.197	41.0	52.6	0.962	1.627
Ours w/o concept	56.9	77.6	0.986	1.644	-	-	-	-
Ours w/o causal	1.4	-	-	3.326	0.3	-	-	3.419
Ours w/ RL	29.7	35.1	**0.991**	2.418	2.5	6.0	**1.000**	3.150
Ours	**97.9**	**99.2**	0.971	**0.025**	**99.4**	**99.6**	0.981	**0.013**
	Dataset level-3				Dataset level-4			
Chance	0.0	0.4	-	3.513	0.1	0.4	-	3.147
PlaTe [4]	0.7	-	-	-	0.4	-	-	-
Ours w/ β-VAE [39]	0.0	0.5	-	3.596	0.0	0.0	-	3.107
Ours w/ VCT [31]	0.7	1.2	0.968	3.442	0.2	0.3	1.000	3.193
Ours w/o symbol	15.4	24.1	0.970	2.278	9.8	14.0	0.981	2.149
Ours w/o causal	0.0	-	-	3.691	0.0	-	-	3.201
Ours w/ RL	3.0	3.9	**1.000**	3.030	2.8	3.5	**1.000**	2.498
Ours	**86.5**	**87.0**	0.966	**0.037**	**55.1**	**76.7**	0.978	**0.003**

These results demonstrate the effectiveness of our model design. Our full model achieves the best overall performance in all four levels of tests, and each component of our model contributes remarkably to the performance improvements. The qualitative results are shown in Fig. 4.

B. Interpretable Concepts and Causal Transitions

We qualitatively show the interpretability of the concept learned by our model. We randomly choose 2 images $X_{0,f}$ and $X_{1,f}$, substituting the concept token c_0^i with c_1^i for $i = 1, 2, 3, 4, 5, 6$, which are then fed into the concept detokenizer and the decoder to generate new images. As Fig. 5 shows,

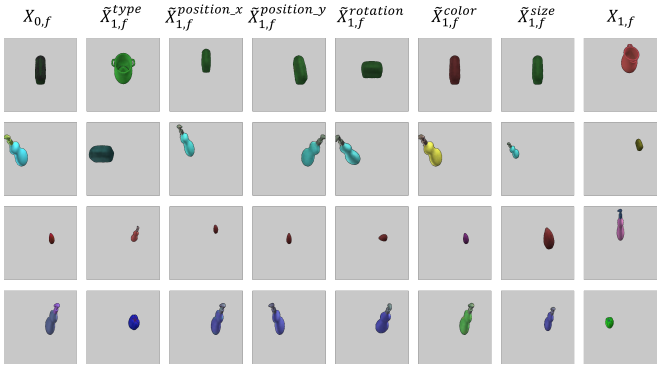

Fig. 5: **Fine-grained attribute-level concept manipulation.** The concept learner generates new images by substituting each concept token c^i_0 from $X_{0,f}$ with c^i_1 from $X_{1,f}$.

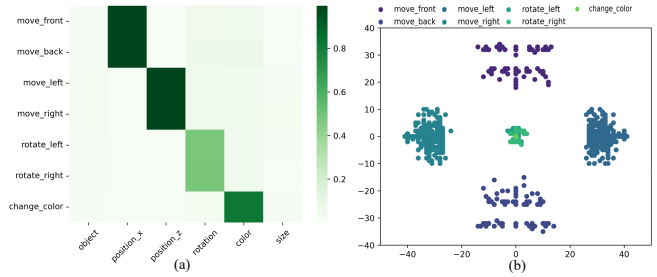

Fig. 6: **Action effects on the learned disentangled concept representations.** (a) l_2 norm between the concept vectors before and after each action. (b) Distributions of position change induced by each action.

with the properly learned concept representations, we could perform fine-grained attribute-level concept manipulation. This indicates that our concept learner is capable of disentangling concept factors and demonstrates the interpretability of our method.

We quantitatively demonstrate the interpretability of our learned causal transitions with statistics of the corresponding causal effects. To be specific, we aim to answer the question: do the learned causal transitions have semantic meaning consistent with the corresponding action? Fig. 6 (a) shows the correlation between concepts and actions, measured with l_2 norm between the concept vectors before and after each action. A larger l_2 norm means a higher correlation. We can see that the learned rotation actions only affect the rotation status in the concept vector. Similarly, the horizontal and vertical movements only affect the x and y coordinates. Fig. 6 (b) shows the distribution of position change induced by 7 displacement actions. For example, the position changes of `move_front` distribute along the positive y-axis, while those of `move_back` distribute along the negative y-axis. This evidence indicates that 1) our learned concept is successfully disentangled, which makes it possible for our model to learn causal transitions, and 2) the learned causal transition is consistently grounded to real-world actions with similar semantics.

C. Generalization Tests

We design three experiments to test the generalizabiliy of our model.

TABLE II: **Quantitative results for generalization tests.** The best scores are marked in **bold**.

	Model ID	ASAcc.(%)(↑) Top-1	ASAcc.(%)(↑) Top-5	ASE(↑)	FSD(↓)	ASAcc.(%)(↑) Top-1	ASAcc.(%)(↑) Top-5	ASE(↑)	FSD(↓)
		Dataset level-1				Dataset level-2			
Unseen Object	Chance	0.6	4.7	-	3.203	1.1	3.2	-	3.591
	PlaTe [4]	18.5	-	-	-	9.7	-	-	-
	Ours w/o symbol	44.0	59.9	0.968	1.507	29.0	43.8	0.986	1.880
	Ours w/o concept	37.1	60.5	0.950	1.319	-	-	-	-
	Ours w/o causal	1.7	-	-	3.233	0.2	-	-	3.563
	Ours w/ RL	30.2	35.9	**0.989**	1.887	2.2	6.1	**1.000**	3.549
	Ours	**72.4**	**97.2**	0.987	**0.470**	**73.2**	**93.6**	0.978	**0.491**
		Dataset level-3				Dataset level-4			
	Chance	0.0	0.0	-	3.544	0	0.1	-	3.518
	PlaTe [4]	0.6	-	-	-	0.8	-	-	-
	Ours w/o symbol	12.6	22.5	0.990	2.710	6.9	11.7	0.972	2.917
	Ours w/o causal	0.0	-	-	3.467	0.0	-	-	3.183
	Ours w/ RL	1.9	5.3	**1.000**	3.484	1.4	4.9	**1.000**	3.370
	Ours	**61.8**	**66.9**	0.960	**0.307**	**29.1**	**43.9**	0.954	**0.424**
Unseen Task		Dataset level-1				Dataset level-2			
	Chance	0.4	2.1	-	3.550	0.1	0.3	-	3.513
	PlaTe [4]	1.4	-	-	-	0.5	-	-	-
	Ours w/o symbol	63.1	78.0	0.974	1.022	40.0	51.9	0.980	1.407
	Ours w/o concept	42.7	70.7	0.971	1.485	-	-	-	-
	Ours w/o causal	0.0	-	-	3.536	0.0	-	-	3.525
	Ours w/ RL	26.3	30.1	**0.994**	2.159	2.8	7.0	**1.000**	3.417
	Ours	**98.7**	**99.3**	0.985	**0.015**	**98.2**	**99.4**	0.991	**0.019**
Real-world Data		Dataset level-1				Dataset level-2			
	Chance	2.0	5.0	-	3.261	1.0	2.0	-	3.370
	PlaTe [4]	12.0	-	-	-	5.0	-	-	-
	Ours	**52.0**	**71.0**	**0.980**	**1.341**	**36.0**	**47.0**	**0.987**	**1.765**
		Dataset level-3				Dataset level-4			
	Chance	0.0	1.0	-	3.498	0.0	0.0	-	3.552
	PlaTe [4]	1.0	-	-	-	1.0	-	-	-
	Ours	**21.0**	**27.0**	**0.993**	**1.436**	**11.0**	**15.0**	**1.000**	**1.735**

a) Unseen Objects: Through this experiment, we aim to investigate if our model can perform visual planning tasks on objects unseen during training. We test our model on the *Unseen Object* testing dataset (see Sec. II-B for details) and compare the results with several baselines to demonstrate the generalizability of our concept-based object representation module. We expect our concept learner to recognize the color, position, and size attributes of unseen object types during testing. If this is the case, the transition model could consequently apply transitions on these visual attributes for successful manipulation tasks. As shown in Tab. II, our model is significantly more robust than PlaTe and RL-based methods against novel objects.

b) Unseen Tasks: Moreover, we aim to verify that our model is flexible in processing atomic actions. We train our model on tasks with only limited types of action combinations, *i.e.*, the *Unseen Task* dataset. In this experiment, PlaTe only performs at the same level as a random guess, while our model performs as well as it does when being trained on the whole dataset (see Tab. II), which demonstrates the generalizability of our method on unseen tasks.

c) Real-world Data: Finally, we assess our model's potential to generalize to real-world data. We collect a dataset of real images using Intel RealSense D455, which consists of four subsets, each representing a different difficulty level and comprising 100 tasks (equivalent to one-tenth of our validation sets' size). The real-world tasks are the same as the test tasks in dataset CCTP. The quantitative results are demonstrated in Tab. II. Since the real-world data and the *CCTP* dataset have inherent discrepancies, our model, which was not finetuned with real-world data, exhibited a reduction in ASAcc and FSD. However, our model can successfully identify objects' color, position, and size within the real-

world images, and outperform all the comparison models. The qualitative results are shown at the bottom of Fig. 4. To simplify implementation, we focus on visually planning the target objects in the real-world images and ignore encoding and decoding the background.

V. Conclusion

In this paper, we propose a novel visual planning model based on concept-based disentangled representation learning, symbolic reasoning, and visual causal transition modeling. In the future, we plan to extend our model to more complex planning tasks with diverse concepts and actions, and assist robots in real down-stream application tasks.

Acknowledgement: This work is supported by the National Science and Technology Major Project (2022ZD0114900). The authors thank Mr. Zhitian Li and Dr. Meng Wang at BIGAI for help with experiment setup.

References

[1] C. M. Walker and A. Gopnik, "22 causality and imagination," *The Oxford handbook of the development of imagination*, p. 342, 2013.

[2] S. Wang, M. Han, Z. Jiao, Z. Zhang, Y. N. Wu, S.-C. Zhu, and H. Liu, "Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning," in *IEEE/RAS International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[3] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] J. Sun, D.-A. Huang, B. Lu, Y.-H. Liu, B. Zhou, and A. Garg, "Plate: Visually-grounded planning with transformers in procedural tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4924–4930, 2022.

[5] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," *Advances in neural information processing systems*, vol. 28, 2015.

[6] O. Rybkin, C. Zhu, A. Nagabandi, K. Daniilidis, I. Mordatch, and S. Levine, "Model-based reinforcement learning via latent-space collocation," in *International Conference on Machine Learning*, 2021.

[7] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

[8] C. Paxton, Y. Barnoy, K. Katyal, R. Arora, and G. D. Hager, "Visual robot task planning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[9] K. Liu, T. Kurutach, C. Tung, P. Abbeel, and A. Tamar, "Hallucinative topological memory for zero-shot visual planning," in *International Conference on Machine Learning*, 2020.

[10] L. Gao and L. Guan, "Interpretability of machine learning: Recent advances and future prospects," *IEEE MultiMedia*, 2023.

[11] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: A survey," *Information Fusion*, vol. 85, pp. 1–22, 2022.

[12] C. Packer, K. Gao, J. Kos, P. Krähenbühl, V. Koltun, and D. Song, "Assessing generalization in deep reinforcement learning," *arXiv preprint arXiv:1810.12282*, 2018.

[13] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu *et al.*, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[14] F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer, "On disentangled representations learned from correlated data," in *International Conference on Machine Learning*, 2021.

[15] R. Suter, D. Miladinovic, S. Bauer, and B. Schölkopf, "Interventional robustness of deep latent variable models," *arXiv*, pp. 1811–00 007, 2018.

[16] T. Adel, Z. Ghahramani, and A. Weller, "Discovering interpretable representations for both deep generative and discriminative models," in *International Conference on Machine Learning*, 2018.

[17] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1462–1473, 2020.

[18] I. Higgins, A. Pal, A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner, "Darla: Improving zero-shot transfer in reinforcement learning," in *International Conference on Machine Learning*, 2017.

[19] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *International Conference on Machine Learning*, 2020.

[20] M. Edmonds, *Learning How and Why: Causal Learning and Explanation from Physical, Interactive, and Communicative Environments*. University of California, Los Angeles, 2021.

[21] B. Schölkopf, "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 765–804.

[22] A. d. Garcez, S. Bader, H. Bowman, L. C. Lamb, L. de Penning, B. Illuminoo, H. Poon, and C. G. Zaverucha, "Neural-symbolic learning and reasoning: A survey and interpretation," *Neuro-Symbolic Artificial Intelligence: The State of the Art*, vol. 342, no. 1, p. 327, 2022.

[23] X. Lin, Z. Huang, Y. Li, J. B. Tenenbaum, D. Held, and C. Gan, "Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools," *arXiv preprint arXiv:2203.17275*, 2022.

[24] T. Silver, R. Chitnis, J. Tenenbaum, L. Kaelbling, and T. Lozano-Perez, "Learning symbolic operators for task and motion planning," in *IEEE/RAS International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[25] D. Kazhdan, B. Dimanov, H. A. Terre, M. Jamnik, P. Liò, and A. Weller, "Is disentanglement all you need? comparing concept-based & disentanglement approaches," *arXiv preprint arXiv:2104.06917*, 2021.

[26] S. Van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, "Are disentangled representations helpful for abstract visual reasoning?" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] P. Yu, S. Xie, X. Ma, B. Jia, B. Pang, R. Gao, Y. Zhu, S.-C. Zhu, and Y. N. Wu, "Latent diffusion energy-based model for interpretable text modelling," in *International Conference on Machine Learning*, 2022.

[28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[29] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3vae: Self-supervised sequential vae for representation disentanglement and data generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner, "Early visual concept learning with unsupervised deep learning," *arXiv preprint arXiv:1606.05579*, 2016.

[31] T. Yang, Y. Wang, Y. Lu, and N. Zheng, "Visual concepts tokenization," *arXiv preprint arXiv:2205.10093*, 2022.

[32] W. Stammer, M. Memmel, P. Schramowski, and K. Kersting, "Interactive disentanglement: Learning concepts by interacting with their prototype representations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] N. Saini, K. Pham, and A. Shrivastava, "Disentangling visual embeddings for attributes and objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," *arXiv preprint arXiv:1904.12584*, 2019.

[35] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, "Egotaskqa: Understanding human tasks in egocentric videos," *arXiv preprint arXiv:2210.03929*, 2022.

[36] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.

[37] A. Fire and S.-C. Zhu, "Inferring hidden statuses and actions in video by causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[38] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "AI2-THOR: An Interactive 3D Environment for Visual AI," *arXiv*, 2017.

[39] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2016.

[40] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[41] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.