Task-Relevant Language-conditioned Segmentation for Robust Generalization in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Humans possess a remarkable ability to filter out irrelevant sensory clutter, extracting only the information needed to anticipate and act within dynamic environments. Prior attempts to mitigate this through augmentation and masking strategies have improved robustness, but remain limited by computational overhead, weak semantic grounding, or instability in actor-critic training. Inspired by how language guides human perception, we introduce Task Relevant Language-conditioned Segmentation (TaLaS), a framework that leverages language-conditioned segmentation to impose semantic structure on visual observations. TaLaS employs a two-phase design where in the first phase, a lightweight masker is pretrained on unaugmented, language-guided masks; in the second phase, a student masker is regularized with strong augmentations to enforce consistency. This yields a task-relevant feature extractor that improves policy stability and removes the need for online segmentation at inference time. To address the actor's deployment distribution shift, we employ asymmetric actor-critic training. TaLaS improves robustness to distractors and achieves particularly strong performance under challenging visual shifts on RL-ViGen, while remaining competitive in easier settings. The benchmark includes challenging variants of the DeepMind Control Suite, Quadruped Locomotion and Dexterous Manipulation tasks. `https://talas-rl.github.io/`

1 INTRODUCTION

Visual Reinforcement Learning (RL) has made substantial progress in learning control policies directly from high-dimensional sensory observations, enabling agents to acquire complex behaviors without access to privileged state information (Levine et al., 2016; Laskin et al., 2020b). Despite these advances, the generalization ability of visual RL agents remains severely limited. In many commonly used benchmarks, observations are generated from simulators in which the visual scene is already strongly aligned with the task, often through carefully designed viewpoints, simplified backgrounds, or environment constructions in which most visible content is implicitly relevant to control (Ha & Schmidhuber, 2018; Hafner et al., 2018; Hansen et al., 2024). Under such conditions, agents can achieve strong in-distribution performance while relying on fragile correlations that do not persist under even modest visual changes. When evaluated under shifts in texture, illumination, background dynamics, camera perturbations, or the presence of distractors, performance often deteriorates sharply (Cobbe et al., 2019; Ma et al., 2025). The fundamental challenge lies in disentangling task-relevant signals from exogenous noise while ensuring that only meaningful information is incorporated in learned representations.

At its core, the problem is of selective representation learning. Pixel-based RL agents typically process the full visual field as a dense input tensor and must implicitly infer, from reward alone, which parts of the scene should be encoded and which should be ignored. This places a substantial burden on the optimization process. Because reward is sparse, delayed, and entangled with exploration, the agent is rarely given direct supervision about what in the image matters for control. As a result, learned representations can absorb noise factors that are predictive in the training distribution but spurious from the standpoint of task structure. Background patterns, distractor motion, lighting statistics, and incidental textures may all become embedded in the policy or critic representation if they correlate with reward during training. Such entanglement reduces robustness under distribution shift and makes it difficult for the agent to transfer

across environments that preserve task semantics but alter appearance. This stands in contrast to human perception, which is inherently structured and selective. Humans do not treat every visible pixel or region as equally relevant for action. Instead, perception is organized around semantically meaningful entities such as objects, object parts, spatial relations, and action-relevant affordances. Irrelevant segments are suppressed to prevent distraction, while attention is dynamically adapted as scene context changes (Nasr et al., 2008; Seidl et al., 2012). The visual system selectively prioritizes structure that is useful for the intended behavior. This mismatch underscores a fundamental gap between how agents and humans parse visual scenes.

Recent work has explored masking-based strategies that mask irrelevant parts asking-based strategies that masks irrelevant parts of the input to focus learning on relevant information, but they exhibit various limitations (Grooten et al., 2024; Zhang et al., 2024; Huang et al., 2023). SGQN (Bertoin et al., 2022) uses Q-function saliency maps to mask observations and suppress distractors. Since supervision is value-derived and shifts with critic training, early saliency is noisy and hyperparameter-sensitive, leaving perception entangled with a partially learned critic and inheriting its biases. Focus-then-Decide (Chen et al. (2024), FTD) performs observation segmentation via the Segment Anything Model (Kirillov et al. (2023), SAM), followed by an attention selector that identifies objects for control. This requires segmentation at every step, leading to high computational overhead and limiting practicality in complex tasks. SAM-G (Wang et al., 2023) uses DINOv2–SAM correspondences with manual point annotations to generate masks that are applied at every step, resulting in high computational overhead and dependence on SAM for entire training and evaluation. Our method avoids per-task annotation and online segmentation by using language identifiers to guide SAM once, then distilling these priors into a lightweight masker trained under augmentations.

A central mechanism through which humans organize perception is language. Verbal descriptions and category labels are not merely communicative abstractions; they actively reshape perceptual processing by biasing attention toward relevant features and suppressing irrelevant alternatives (Lupyan, 2012; Lupyan & Ward, 2013). Language compresses the hypothesis space of perception. For example, when a person is instructed to *"Pick the red apple from the crowded shelf"*, the term *apple* constrains perception to certain shape, size and graspable structure. *Red* sharpens color filters, and *shelf* imposes a spatial prior, allowing irrelevant clutter or background motion to be suppressed. Language narrows the search space of perception into discrete categories, making raw sensory input tractable for decision-making (Yang & Zelinsky, 2009). This observation is particularly compelling for visual RL, where the main challenge is often not lack of information, but excess information without a mechanism for semantic filtering.

Yielding inspiration from these insights, we propose **Ta**sk-Relevant **La**nguage-conditioned **S**egmentation (TaLaS), a novel algorithm that integrates language-guided segmentation with mask distillation to improve generalization in visual reinforcement learning. To ensure computational efficiency, we adopt a two-phase training pipeline. In Phase I, natural language prompts are provided to a segmentation backbone (Cuttano et al. (2025), SAMWISE), which partitions clean observations into semantically grounded masks. These masks serve as supervisory signals for a lightweight convolutional masker, which is optimized to approximate the teacher outputs using trajectories generated under a random exploration policy π_{exp}. This phase serves as an offline semantic distillation stage: the expensive segmentation model and language interface are used only to produce supervisory targets, while the student masker learns to infer task-relevant segmentation directly from raw images. In Phase II, after T_{exp} exploration steps, the segmentation backbone and language interface are removed, and the convolution masker is frozen to serve as a teacher. A second, noisy student masker is then introduced, which receives augmented observations (overlayed with additional image, Appendix A.3) and is trained via consistency regularization to reproduce the teacher's outputs on corresponding clean inputs (Figure 1). This stage performs a form of robustness transfer: semantic priors extracted from clean scenes are propagated into a compact model that remains stable under visual corruption. The resulting role-reversal distillation procedure yields a masker that is both semantically informed and resistant to nuisance variation, without requiring online foundation-model segmentation or language processing during RL optimization. To improve robustness to mask imperfections at deployment, TaLaS uses an asymmetric actor–critic update: during actor updates, the policy is conditioned on clean and augmented masked views, while the critic evaluating the action uses the clean view; during critic updates, the online critic is trained on clean and augmented masked views, while the bootstrap target is computed from the clean view only. The learned masker is not an endpoint by itself, but a compact

perceptual bottleneck that filters irrelevant visual content before downstream control. Under this view, the contribution of TaLaS is not a new segmentation model, but an RL training framework in which semantically grounded masking improves robustness, efficiency, and transfer across visually perturbed domains.

We evaluate TaLaS on **eleven** environments spanning three challenging benchmarks in RL ViGen (Yuan et al., 2023): the DeepMind Control Generalization Benchmark (Hansen & Wang, 2021), Quadruped Locomotion (Hansen & Wang, 2021) and Dexterous Manipulation (Rajeswaran et al., 2018). Across these domains, TaLaS shows its strongest gains over several strong well-established vision-based RL baselines, under challenging distractor-heavy evaluation settings, while remaining competitive in several easier conditions.

2 Related Work

2.1 Generalization with Data Augmentation

RL agents exhibit significant generalization gaps, often suffering sharp performance drops in out-of-distribution settings as a result of overfitting and poor adaptability (Kirk et al., 2023; Jiang et al., 2023). To address this, data augmentation has been widely adopted in visual RL, with methods such as RAD (Laskin et al., 2020b), DrQ (Yarats et al., 2021), and DrQ-v2 (Yarats et al., 2022) using image transformations to improve robustness, whereas approaches like SVEA (Hansen et al., 2021), SODA (Hansen & Wang, 2021) and SADA (Almuzairee et al., 2024) regularize representations by mixing clean and augmented streams and enforcing constraints on encoder features. Information-theoretic methods pursue a similar goal of compression and invariance by penalizing redundant or noisy channels (You et al., 2022; Dave & Rueckert, 2024; Wang et al., 2024), but they too lack explicit semantic grounding, making it difficult to bifurcate which features should be suppressed or retained. Recent work further refines the augmentation toward more selective forms: SRM (Huang et al., 2022) suppresses high-frequency components to mitigate texture bias, but global filtering risks erasing motion-relevant detail. TLDA (Yuan et al., 2022a) perturbs pixels deemed irrelevant through a Lipschitz-based sensitivity analysis of the policy, yet this proxy depends on local smoothness assumptions and may misclassify low-salience but task-critical signals. CG2A (Liu et al., 2023) stabilizes training under multiple transforms by calibrating gradient magnitudes, yet its robustness is bounded by the augmentation set. While these approaches improve generalization, they still enforce invariances through input-level perturbations and lack explicit semantic structure, leaving generalization tied only to augmentations to preserve task-relevant information.

2.2 Masking Distractors in RL

Masking has emerged as a significant strategy to improve generalization in visual RL by suppressing distractors. SGQN (Bertoin et al., 2022) derives saliency masks from Q-function attribution to enforce value consistency, but supervision is critic-dependent and unstable early in training. MLR(Yu et al., 2022) employs random occlusion and latent reconstruction to promote dynamics-relevant encodings, yet remains object-agnostic, offering no guarantee of alignment with task semantics. InfoGating (Tomar et al., 2023) learns differentiable gates via inverse dynamics and Q-learning losses, but inherits biases from auxiliary objectives and risks occluding subtle cues. MaDi (Grooten et al., 2024) uses a reward-driven CNN masker to filter irrelevant pixels, achieving efficiency gains but suffering from sparse, delayed supervision. Recent methods leverage segmentation backbones: SAM-G (Wang et al., 2023) combines DINOv2–SAM correspondences with human prompts to generate task-relevant masks, while FTD (Chen et al., 2024) uses SAM and attention selectors. Both methods, however, use a segmentation model during training and inference, introducing substantial computational overhead and, in some cases, reliance on instance-level labels (Section 5.6.2). Overall, masking improves robustness, but existing methods remain limited by unstable signals (SGQN, InfoGating), lack of semantic grounding (MLR, MaDi), or reliance on expensive foundation-model segmentation (SAM-G, FTD).

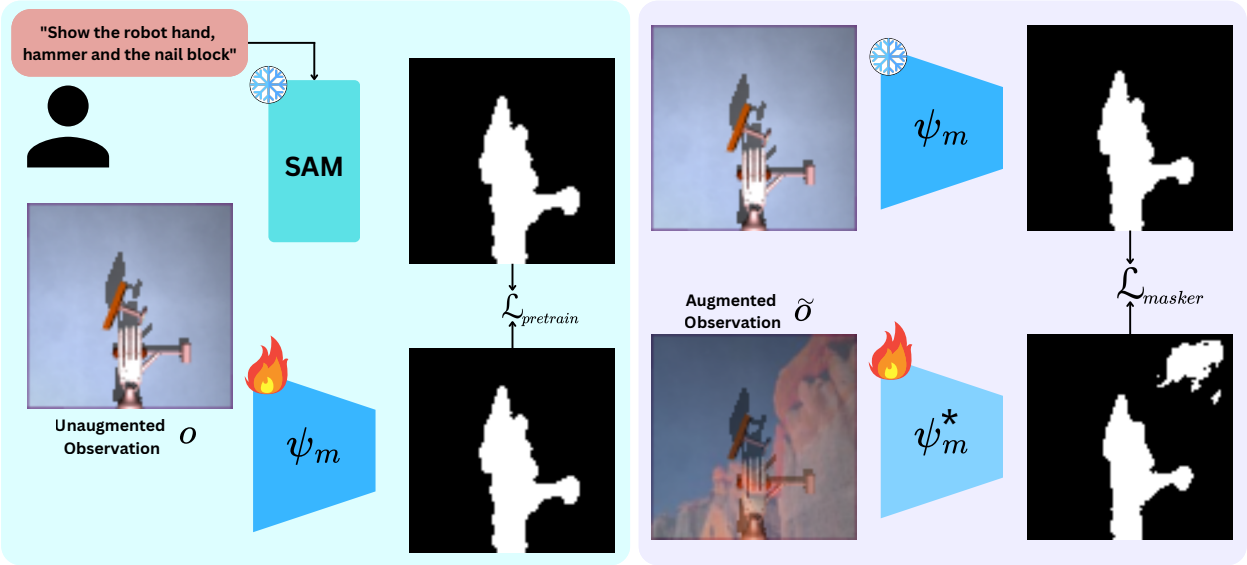


Figure 1: Depiction of the masker training phases of TaLaS. **Phase I (left): Language-conditioned distillation.** A text prompt drives a frozen segmentation model (SAM) on unaugmented frames to produce task masks and a compact masker ψ_m is pretrained by masked reconstruction to imitate these target masks. **Phase II (right): Augmentation-consistent student.** An unaugmented frame yields a target mask from the frozen teacher, while the corresponding augmented frame is fed to the student masker $\psi_m^\star$, which is trained through $\mathcal{L}_{\text{masker}}$ to match the teacher mask. Actor-critic gradients are stopped at the masked observation. Modules marked with 🔥 are trainable, whereas modules marked with ❄ remain frozen.

3 Preliminaries

3.1 Visual Reinforcement Learning

Reinforcement learning (RL) solves sequential decision problems via interaction with an environment (Sutton & Barto, 2018). In visual settings, we model the task as a Partially Observable Markov Decision Process (POMDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, P, \Omega, R, \gamma \rangle$, where $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition kernel, $\Omega : \mathcal{S} \to \mathcal{P}(\mathcal{O})$ the observation kernel, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward, and $\gamma \in [0, 1)$ the discount. To mitigate partial observability (Kaelbling et al., 1998), we use a k-frame stack $s_t = (o_t, o_{t-1}, \ldots, o_{t-k+1})$ as a practical proxy for the history, which enriches temporal context without full belief-state inference. The objective is to learn a policy π_ϕ maximizing discounted return $\mathbb{E}_{\tau \sim (\mathcal{M}, \pi_\phi)}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where trajectories τ is induced by the underlying dynamics of the POMDP.

3.2 Reinforcement Learning Backbone

We build on PIE-G (Yuan et al., 2022b), which couples a frozen pre-trained visual encoder with a lightweight off-policy actor–critic. A fixed image encoder ϕ_{enc} (e.g., ResNet (He et al., 2016)) projects masked observations to features $z_t = \phi_{enc}(o_t)$, where we use early-layer embeddings for better visual generalization and batch normalization to adapt to distribution shift. It is built on DrQ-v2 (Yarats et al., 2022), a DDPG (Lillicrap et al., 2015) style learner with twin critics and target networks. Given two critics Q_{θ_k} ($k \in \{1, 2\}$) and targets $Q_{\bar{\theta}_k}$, the critic minimizes

$$\mathcal{L}_Q(z) = \sum_{k=1}^{2} \mathbb{E}_{(o_t, a_t, r_t, o_{t+1}) \sim \mathcal{D}} \left[\left(Q_{\theta_k}(z_t, a_t) - y_t \right)^2 \right], \tag{1}$$

where the target is $y_t = r_t + \gamma \min_{k=1,2} Q_{\bar{\theta}_k}(z_{t+1}, a_{t+1})$, with $a_{t+1} = \pi_\phi(z_{t+1}) + \epsilon$, and $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma^2), -c, c)$. The actor is trained to maximize critic values via the objective $\mathcal{L}_\pi =$

$-\mathbb{E}\left[\min_{k=1,2} Q_{\theta_k}(z_t, a_t)\right]$. Following standard visual RL practice, we apply strong augmentation like overlay during training.

4 Task-Relevant Language-conditioned Masking

We address robust pixel-based control under distribution shift by enforcing task structure prior to policy learning. Training proceeds in two phases:

(i) A short distillation phase trains a compact masker to reconstruct masks generated by a frozen segmentation model.

(ii) A consistency phase trains a student masker on augmented samples to match the target masks from unaugmented samples, yielding augmentation-stable masks.

Control is learned via off-policy RL on masked inputs. Masking reduces distraction-induced variance, stabilizes value estimates under augmentation, and eliminates segmentation overhead at training and deployment.

4.1 Problem Statement

Our work addresses the challenge of generalization in visual reinforcement learning, where agents are trained in unaugmented or domain randomized environments, then evaluated in visually perturbed environments with unseen distractors. Given a set of POMDPs $\{\mathcal{M}_\nu : \nu \in \mathcal{V}\}$, sharing similar dynamics and reward structure, but varying in observation space $\mathcal{O}_\nu$ (due to distractions), our objective is to learn a policy π that maintains consistent performance across this family of environments $\mathcal{M}$, despite shifts in the observation distribution in a zero-shot manner. Let $\mathcal{M}_{\text{train}}$ be the training environment and $\mathcal{M}_{\text{test}}$ a set of test environments. Following prior work (Kirk et al., 2023; Wang et al., 2024), the generalization gap of a policy π is defined as $\nabla_{\text{gen}} := \eta_{\mathcal{M}_{\text{test}}}(\pi) - \eta_{\mathcal{M}_{\text{train}}}(\pi)$, where $\eta_{\mathcal{M}}(\pi)$ denotes the expected return of π in environment $\mathcal{M}$. We assume that the policy behaves well on the training environment, as this metric remains susceptible to random policy (Jiang et al., 2023). Rather than formulating the problem as one of image segmentation, we view task-relevant masking as a structured perceptual prior for visual reinforcement learning. In TaLaS, language-conditioned segmentation is used only to provide semantic supervision for distillation, while the actual objective remains zero-shot policy generalization under visual distraction and appearance shift.

4.2 Mask Learning

To enforce semantic consistency across visually diverse environments, we construct task-relevant masked observations in two stages as described below. These masked observations serve as inputs to the downstream RL pipeline.

4.2.1 Phase I: Language-Guided Mask Distillation

This first phase, as shown in Figure 1 (left), aims to distill spatial task structure into a compact teacher masker by learning to reconstruct masked images derived from a frozen segmentation model. For each environment, we define a single task-level natural language prompt $\ell \in \mathcal{L}$ at the beginning of training, before the mask-distillation stage and policy learning. The same prompt is then used throughout this entire phase to generate segmentation supervision. For a single environment, it is kept fixed across frames, episodes and random seeds, and is not adapted or tuned using policy performance or test-environment feedback. We have added the full list of task-level prompts in the Appendix A.1.2 (Table 3).

Given a clean input image $o \in \mathcal{O}^{H \times W \times C}$, and a natural language prompt $\ell \in \mathcal{L}$ describing the task semantics, a frozen language-conditioned segmenter $\Phi(o, \ell)$ produces a set of region proposals $\{m_k^+(o)\}_{k=1}^K \subset \{0,1\}^\Omega$. These proposals are then combined into a single task-relevant mask $m^+(o) \in [0,1]^\Omega$ via the logical OR operation $m^+(o) = \bigvee_{k=1}^K m_k^+(o)$. To amortize this computationally expensive segmentation process, we train a lightweight convolutional network masker ψ_m to predict the mask $m(o) = \psi_m(o) \in [0,1]^\Omega$. The training

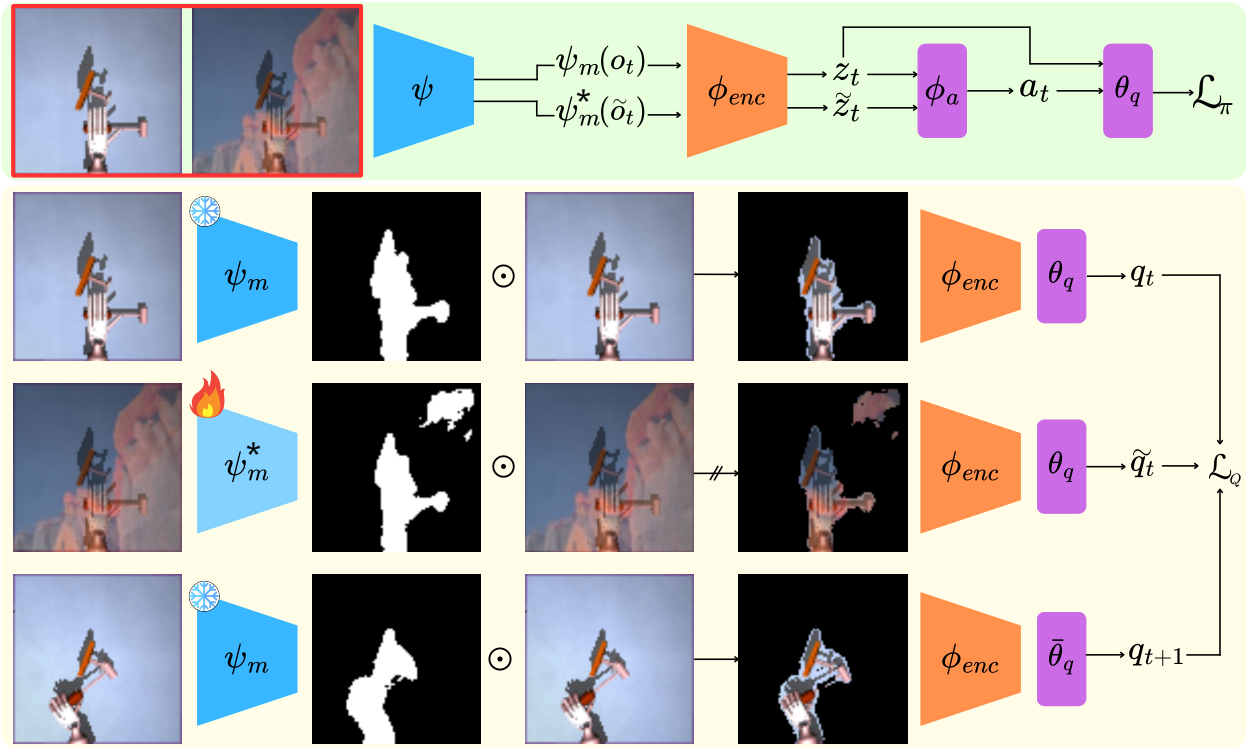


Figure 2: Integrating the masking strategy with the RL backbone for robust policy learning. **Actor.** The frozen masker ψ_m filters clean inputs, while the student masker $\psi^\star_m$ filters augmented inputs. The encoder ϕ_{enc} maps both masked views to z_t and $\tilde{z}_t$. The red rectangle denotes batch-wise concatenation $[z_t, \tilde{z}_t]_N$, where clean and augmented views are stacked as batch elements, not feature-concatenated. **Critic.** Critics θ_q use both clean and augmented masked views, while the target critic bootstraps only from the clean view. The stop-gradient sign blocks actor-critic gradients to $\psi^\star_m$, which is updated only by $\mathcal{L}_{\text{masker}}$. $\bar{\theta}_q$ denotes target critics; 🔥 and ❄️ denote trainable and frozen modules respectively.

objective minimizes the binary cross-entropy between the ground-truth mask $m^+(o)$ produced by segmentation model and the predicted mask $m(o)$ by the convolution masker as $\mathcal{L}_{\text{pretrain}}(\psi) = \text{BCE}\left(m^+(o); m(o)\right)$. This formulation encourages the teacher to learn a spatial filter that preserves task-relevant content while suppressing irrelevant visual regions. These masks are collected during the initial exploration phase of the agent under random policy.

But why use a random policy? A random policy ensures diverse coverage of the observation space, exposing the segmenter to varied agent and object poses, crucial for learning generalizable masks. This avoids reliance on expert policies or co-training with control (online), keeping the segmentation phase modular and self-supervised. Once trained for k-iterations, ψ_m serves as a teacher for the augmentation-consistent student masker trained in Phase II.

4.2.2 Phase II: Augmentation-Consistent Student Masker

For the second phase, as shown in Figure 1 (right), after pretraining the teacher masker ψ_m on unaugmented images, we train a student masker $\psi^\star_m$ to produce task-relevant masks that are consistent under visual augmentations (Figure 1). The goal of this phase is to transfer the clean semantic prior learned by the teacher to visually corrupted observations, rather than to relearn task relevance from reward. The teacher masker ψ_m has only seen unaugmented observations and therefore provides a stable clean-view semantic target. However, because the deployed agent may encounter augmented or visually shifted observations, directly using this clean teacher on corrupted inputs can introduce a distribution mismatch. Let $\tau \sim \mathcal{T}$ be a photometric transform that augments the image to $\tilde{o} = \tau(o)$. The teacher generates a mask $m(o) = \psi_m(o) \in [0,1]^\Omega$ and to align this supervision with the augmented view, the student masker $\psi^\star_m$ then predicts a mask

$m^\star = \psi_m^\star(\tilde{o}) \in [0,1]^\Omega$ from the augmented input. The student is also trained with the teacher target via binary cross-entropy, $\mathcal{L}_{\text{masker}}(\psi_m^\star) = \text{BCE}\,(m^\star, m(o))$.

This teacher-student separation decouples semantic distillation from augmentation robustness. A single masker trained jointly on clean and augmented inputs would have to preserve clean semantic precision and become invariant to strong visual corruption using the same parameters. This simpler alternative entangles the semantic-supervision role and the augmentation-robustness role in a single network, which can weaken the clean semantic anchor under strong overlays. By keeping ψ_m fixed and adapting only $\psi_m^\star$, TaLaS preserves a stationary semantic target while learning a deployment-facing augmentation-robust masker. Importantly, the student masker is not optimized by the actor-critic objective. When the student is updated during the RL phase, it is updated only through this consistency loss against the frozen teacher, and gradients from the RL loss are stopped at the masked observation. For evaluation, the trained student masker $\psi_m^\star$ is deployed as a frozen module generating task-consistent augmentation-invariant masks for downstream policy execution.

4.3 Reinforcement Learning with Masked Observations

To ensure robustness under distribution shift, we integrate the augmentation-consistent student masker $\psi_m^\star$ within our RL framework (Figure 2). During RL, the teacher masker ψ_m remains frozen. The student masker ψ_m^* is used for augmented masked observations and, when updated, is updated only through the augmentation-consistency loss against the frozen teacher. Actor-critic gradients are stopped at the masked observation, so the RL objective updates only the actor and critic. At each timestep t, the agent receives a raw image observation o_t, which is processed by the frozen teacher masker ψ_m to yield a binary mask $\psi_m(o_t) \in [0,1]^\Omega$. The clean masked observation is computed as the Hadamard product $o_t \odot \psi_m(o_t)$, and is passed through the encoder ϕ_{enc}, yielding the clean masked representation $z_t = \phi_{\text{enc}}(o_t \odot \psi_m(o_t)) \in \mathbb{R}^d$. Similarly, given an augmented observation $\tilde{o}_t = \tau(o_t)$, the student masker $\psi_m^\star$ produces an augmented masked observation $\tilde{o}_t \odot \psi_m^\star(\tilde{o}_t)$, which is encoded as $\tilde{z}_t = \phi_{\text{enc}}(\tilde{o}_t \odot \psi_m^\star(\tilde{o}_t)) \in \mathbb{R}^d$. Following SADA (Almuzairee et al., 2024), z_t and $\tilde{z}_t$ are used as clean and augmented training streams, respectively, rather than as two simultaneous inputs to the actor. The actor remains a single-input policy $\pi_\phi(a_t \mid z_t)$. During actor updates, the same actor is applied to both z_t and $\tilde{z}_t$, while the resulting actions are evaluated through the clean representation z_t. This exposes the policy to both clean and corrupted masked observations, while keeping value estimation anchored to the lower-variance clean stream. During critic updates, the critic is trained on both clean and augmented masked representations, but the bootstrap target is computed only from the clean next-state representation. Specifically, the target Q-value is $y_t = r_t + \gamma \min_{k=1,2} Q_{\bar{\theta}_k}(z_{t+1}, \pi_{\bar{\phi}}(z_{t+1}))$, where $z_{t+1} = \phi_{\text{enc}}(o_{t+1} \odot \psi_m(o_{t+1}))$. The critic loss is

$$\mathcal{L}_{\text{critic}} = \frac{1}{2}\sum_{k=1}^{2} \mathbb{E}_{(o_t,a_t,r_t,o_{t+1})\sim\mathcal{D}} \left[\left(Q_{\theta_k}(z_t,a_t) - y_t\right)^2 + \left(Q_{\theta_k}(\tilde{z}_t,a_t) - y_t\right)^2 \right].$$

This objective regresses both the clean masked representation z_t and the augmented masked representation $\tilde{z}_t$ to the same clean-view bootstrap target. The target is computed from z_{t+1}, rather than an augmented next-state representation, to avoid injecting augmentation-induced variance into the temporal-difference target while still training the online critic on augmented masked inputs. The actor is updated using the asymmetric objective

$$\mathcal{L}_\pi(\mathcal{D}) = \mathbb{E}_{o_t\sim\mathcal{D}} \left[-\min_{k=1,2} Q_{\theta_k}\left([z_t, z_t]_N, \pi_\phi([z_t, \tilde{z}_t]_N)\right) \right].$$

Here, $[\cdot,\cdot]_N$ denotes concatenation along the batch dimension, not concatenation along the feature dimension. Thus, $\pi_\phi([z_t, \tilde{z}_t]_N)$ means that the same single-input actor is applied batch-wise to both clean and augmented masked representations. Consequently, the clean counterpart of the augmented view is required only during the SADA-style training update and is not assumed to be available at deployment. At inference time, no clean counterpart is required; the deployed agent receives only the current observation and acts as $a_t = \pi_\phi\left(\phi_{\text{enc}}\left(o_t \odot \psi_m^\star(o_t)\right)\right)$.

Table 1: Performance on DMC Benchmark Environment in Video-Hard (VH) and Video-Easy (VE) settings. T: Tasks, CS: Cartpole Swingup, WW: Walker Walk, WS: Walker Stand, BiC: Ball in Cup Catch, FS: Finger Spin, CR: Cheetah Run.

T(VE)	SAC	DrQ	DrQ-v2	CURL	SVEA	SRM	PIEG	SGQN	MaDi	CNSN	TaLaS
CS	398±60	485±105	267±41	404±67	782±27	724±75	482±51	717±35	**848±6**	353±40	531±65
WW	245±165	682±89	175±117	556±133	819±81	854±42	871±22	860±53	895±24	**923±8**	878±33
WS	389±131	873±83	560±48	852±75	961±8	**963±57**	957±12	955±9	**967±3**	956±9	**961±13**
BiC	192±157	318±157	871±106	316±119	871±106	**924±35**	910±37	761±171	807±144	892±43	856±48
FS	206±169	533±119	456±15	502±19	808±33	**853±76**	837±107	609±61	679±17	683±44	**850±41**
CR	87±21	102±30	64±22	104±24	249±20	257±21	287±20	269±33	294±25	**347±34**	219±31
Avg	253	499	457	456	757	**763**	724	697	748	692	716
T(VH)	**SAC**	**DrQ**	**DrQ-v2**	**CURL**	**SVEA**	**SRM**	**PIEG**	**SGQN**	**MaDi**	**CNSN**	**TaLaS**
CS	158±17	138±9	130±3	114±15	393±45	475±75	323±24	488±18	**619±24**	309±19	395±35
WW	122±47	104±22	34±11	58±18	377±93	535±35	641±63	655±45	504±33	669±42	**787±50**
WS	231±57	289±49	151±13	45±5	834±46	863±57	852±56	851±24	824±287	856±38	**940±22**
BiC	101±37	100±40	97±27	115±33	403±174	566±135	773±74	782±57	758±135	721±7	**864±86**
FS	13±10	91±13	21±4	27±21	335±58	419±32	762±59	554±8	358±25	556±291	**788±38**
CR	10±5	32±13	23±5	21±7	105±37	115±24	154±17	144±34	170±14	162±23	**198±22**
Avg	106	126	76	63	408	496	584	579	539	545	**662**

5 Experiments

In this section, we present our experimental evaluations conducted on generalization benchmarks from RL-ViGen (Yuan et al., 2023). These benchmarks were selected as they encompass a wide range of environments: **(1)** DeepMind Control Generalization Benchmark (Hansen & Wang, 2021) for assessing continuous control under dynamic and reward variation; **(2)** Quadruped Locomotion (Zakka et al., 2022), which includes a Unitree quadruped robot, with complex balance and control demands; **(3)** Dexterous Manipulation (Rajeswaran et al., 2018), featuring multi-object interactions requiring precise strategies with sparse rewards.

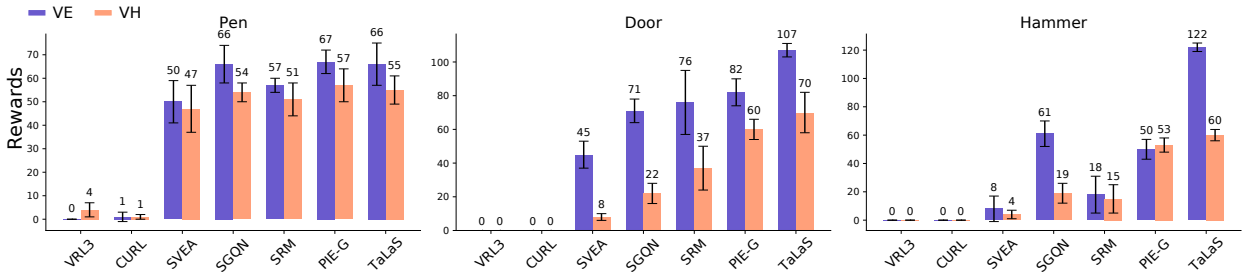


Figure 3: Performance on Adroit tasks (Pen, Door, Hammer) under *video-easy* (VE) and *video-hard* (VH) settings, averaged over 5 seeds. Error bars indicate standard deviation.

5.1 Baseline Methods

We compare TaLaS against a comprehensive set of visual RL baselines designed to improve generalization: DrQ (Yarats et al., 2021), DrQ-v2 (Yarats et al., 2022), CURL (Laskin et al., 2020a), SVEA (Hansen et al., 2021), SRM (Huang et al., 2022), PIE-G (Yuan et al., 2022b), SGQN (Bertoin et al., 2022), MaDi (Grooten et al., 2024), CNSN(+PIEG) (Li et al., 2024), and VRL3 (Wang et al., 2022). These methods were chosen because they represent the most established approaches to generalization in visual RL, covering the main strategies explored in the literature, including data augmentation, pretrained encoders, saliency or masking, normalization, and learning from demonstrations.

5.2 Zero-shot Evaluation.

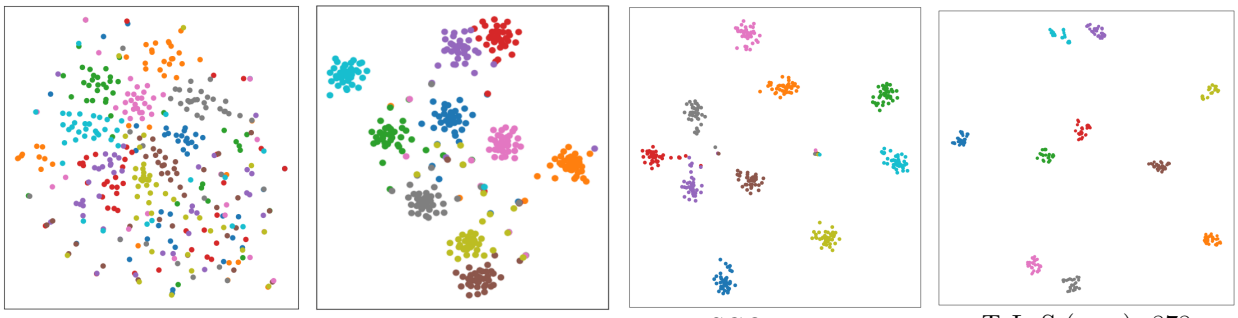
We measure generalization without any fine-tuning on unseen environments spanning multiple distraction intensities. Specifically, we evaluate performance on the *video-easy* (10 background videos) and the more challenging *video-hard* (100 background videos without surface) configurations across all environments. Each seed undergoes evaluation over 100 episodes, corresponding to the designated noise levels.

Table 2: Performance on Unitree Walk and Unitree Stand task.

Task (VE)	**DrQ**	**DrQ-v2**	**CURL**	**SVEA**	**SRM**	**PIEG**	**SGQN**	**TaLaS (ours)**
Walk	67±9	98±16	75±14	98±28	98±9	140±64	152±87	**189±32**
Stand	341±20	375±65	431±38	**587±40**	553±28	380±66	447±50	325±51
Average	204	236	253	**343**	326	260	332	257
Task (VH)	**DrQ**	**DrQ-v2**	**CURL**	**SVEA**	**SRM**	**PIEG**	**SGQN**	**TaLaS (ours)**
Walk	40±22	83±24	61±26	74±52	72±29	204±76	123±68	**206±23**
Stand	66±26	96±37	99±25	279±11	**300±34**	202±43	140±47	**289±36**
Average	53	89	80	177	186	203	131	**247**

5.3 Deepmind Control Suite

We evaluate our algorithm on the DMC-GB (Hansen & Wang, 2021) benchmark, spanning six tasks as shown in Table 1.



SRM: 535 PIEG: 641 SGQN: 655 TaLaS (ours): 878

Figure 4: t-SNE visualization of clustering results for TaLaS and three selected baseline methods.

5.3.1 Generalization Performance

We assess generalization across five seeds per task, reporting mean and standard deviation of episode returns. As shown in Table 1, TaLaS performs best in 7 of 12 environments. Its advantage is most pronounced in the challenging *video-hard* setting, outperforming all baselines with a **13.4%** gain over next-best method, PIEG, and **17.8%** over the average of the next-best four methods (PIEG, SGQN, CNSN, MaDi). In the *video-easy* setting, TaLaS remains competitive, trailing SRM by a modest **6%**. These results highlight the robustness of our masking strategy under challenging visual shifts, where TaLaS shows its clearest advantage, while remaining competitive in easier conditions.

5.3.2 Representations under Distractors

We evaluate domain invariance by visualizing encoder features via t-SNE (van der Maaten & Hinton, 2008). Ten states are augmented with 40 unseen backgrounds, and their embeddings are plotted with same-state points sharing a color. As shown in Figure 4, TaLaS yields the tightest clusters across background variations, indicating a strong ability to learn domain-invariant representations.

5.3.3 Robustness to Distraction Levels

We assess robustness by quantifying the performance drop from *video-easy* to *video-hard* settings using average returns. As shown in Table 1, most methods suffer significant degradation under elevated distraction. Even strong baselines like SRM and PIEG drop by 35% and 19% respectively. In contrast, TaLaS maintains consistent, with only a **7.5**% drop and yielding high rewards.

5.4 Locomotion

For locomotion, we evaluate on Unitree Stand and Walk tasks from the Unitree Series (Zakka et al., 2022), using the same training and evaluation setup as in the DMC benchmark. Generalization is measured over three seeds per task by computing mean and standard deviation of returns. As shown in Table 2, TaLaS achieves the highest score on Unitree Walk in both video-easy and video-hard settings, outperforming all baselines. However, on Unitree Stand under video-easy, TaLaS yields a lower average score than most of the baselines. Nevertheless, under the more challenging video-hard configuration, TaLaS demonstrates robust performance across both tasks, achieving a **22**% gain in average return over the next-best method PIEG, highlighting its strong generalization to distractor-heavy environments.

5.5 Manipulation

In Adroit (Rajeswaran et al., 2018), we evaluate three tasks: Door, Hammer, and Pen, from a single view in RL-ViGen. The presence of numerous distractor objects requiring masking further increases the difficulty of this environment.

We evaluate generalization on three random seeds per task, reporting mean and standard deviation of returns. As shown in Figure 3, in the *video-easy* setting, TaLaS achieves an average improvement of **48**% over PIE-G, the strongest baseline, and **62**% over the average of the next three best methods (PIE-G, SRM, SGQN). In the more challenging *video-hard* setting, TaLaS still delivers substantial gains, surpassing PIE-G by **9**% and outperforming the average of PIE-G, SRM, and SGQN by **51**%. These results highlight the robustness of our masking strategy, which consistently yields better generalization under both easy and hard visual shifts.

5.6 Ablation Study

5.6.1 Actor Training

To quantify the effect of asymmetric training, we compare TaLaS with and without the proposed asymmetric actor–critic update. In the "without" setting, the actor is trained on only masked unaugmented observations. The critic training remains the same. Figure 5 shows the average reward in all the six DMC-GB environments, averaged over 5 seeds in both *video-easy* and *video-hard* settings. Asymmetric training yields substantial gains over both, easy and hard settings.

5.6.2 Wall Time

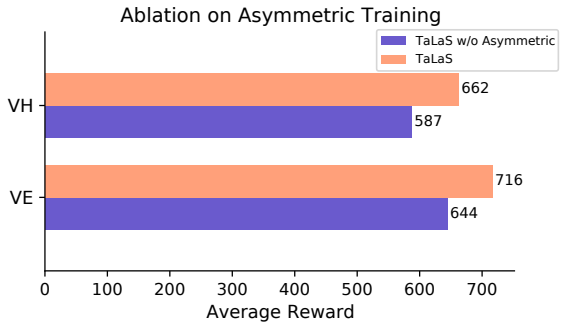


Figure 5: Ablation on asymmetric training in DMC-GB under *video-easy* (VE) and *video-hard* (VH), averaged over 5 seeds.

To assess the computational efficiency of our method, we compare the wall-clock times of TaLaS and a baseline that directly uses SAMWISE during both training and evaluation for Walker Walk task. For TaLaS, we report the actual wall time: ≈ 7.2 seconds per 100 environment steps during training (≈ 10 hours), and ≈ 5 seconds per evaluation episode. For the SAMWISE-based baseline, we report estimated latency, which amounts to ≈ 52 seconds per 100 environment steps during training (≈ 3 days) and ≈ 150 seconds per episode during evaluation, due to the cost of online segmentation. Because end-to-end runs with online SAMWISE segmentation were prohibitively expensive, we estimate its total wall time by measuring the

per-step segmentation latency and extrapolating it to the full training and evaluation horizon under the same hardware and environment configuration used for TaLaS. This demonstrates that TaLaS substantially reduces compute overhead while maintaining generalization, making it more practical for real-world deployments.

5.6.3 Prompt sensitivity

To assess the effect of prompt specificity, we evaluate mask agreement under our default task-level prompts and a weaker coarse-prompt variant. For each task, we sample 10 random frames from the unaugmented setting and manually annotate task-relevant regions as human ground-truth masks. We then compute the IoU between the human annotation and the mask generated by either the default task-level prompt or the coarse prompt. The coarse prompts remove color cues and reduce semantic specificity. Averaged across all tasks, the default prompt set achieves a higher IoU of 0.6917, while the coarse prompt obtains 0.3890. For example, in Hammer, the default prompt specifies the yellow robot hand, hammer, and red nail block, whereas the coarse prompt only refers to the robot hand and object, leading to reduced mask agreement. Although a few tasks show task-specific deviations, the overall trend indicates that TaLaS is sensitive to prompt specificity, and that prompts naming the controllable body, task-relevant objects and their colors provide more accurate mask supervision. The full prompt variants, per-task IoU scores, human-annotation examples, and qualitative mask visualizations are provided in Appendix A.4.1.

5.6.4 Effect of Two-Stage Masker Distillation

To isolate the effect of the two-stage masker design, we compare full TaLaS against a single-stage masker variant on Walker Walk over five random seeds. In the single-stage variant, the masker is trained directly to predict SAMWISE masks without the teacher-student consistency stage. This variant achieves 842 ± 61 in the video-easy setting and 664 ± 97 in the video-hard setting, whereas full TaLaS achieves 878 ± 33 and 787 ± 50, respectively. This corresponds to a 4.3% improvement in video-easy and a 18.6% improvement in video-hard. These results indicate that the two-stage design provides a measurable benefit over directly training a single masker, with a larger gain under stronger visual distraction. This supports our claim that separating semantic distillation from augmentation-consistent student training improves the robustness of the agent in distraction setting.

6 Conclusion

We presented TaLaS, a two-stage framework for improving generalization in visual reinforcement learning by introducing semantic structure into the observation pipeline. Rather than requiring an RL agent to infer task relevance solely from reward, TaLaS uses language-conditioned segmentation to obtain task-aligned masks and distills these masks into a compact convolutional masker. In the first stage, a frozen segmenter provides semantically grounded supervision on clean observations; in the second stage, the distilled masker is regularized under strong augmentations to produce stable and augmentation-consistent masks. This design yields a lightweight perceptual bottleneck that suppresses nuisance variation while preserving action-relevant visual content, and removes the need for online segmentation during policy learning and deployment. To address the distribution shift induced by imperfect masks at test time, we combine this perceptual pipeline with an asymmetric actor–critic update. The actor is trained to operate under both clean and augmented masked views, while the critic maintains stable value targets through clean-view bootstrapping. Empirically, this combination leads to the clearest gains under challenging distractor-heavy settings, where TaLaS consistently improves robustness and maintains strong performance without incurring the computational overhead of running a foundation segmentation model online. Taken together, these results support the broader view that robust visual control is not only a matter of stronger augmentation or larger encoders, but also of imposing the right semantic structure on what the agent is allowed to perceive.

At the same time, several limitations remain. First, the quality of the learned perceptual bottleneck depends on the quality of the language-conditioned supervision. If the task description is vague, incomplete, or semantically misaligned with the control objective, the resulting masks may omit relevant regions or preserve irrelevant ones. Second, the method depends on sufficient coverage during the initial exploration phase: if

important objects, viewpoints, contact configurations, or interaction states are rarely observed early in training, the distilled masker may not learn to preserve them reliably once frozen. Finally, TaLaS is most naturally suited to settings in which the semantic identity of relevant objects remains stable across domains. When the distribution shift directly changes the appearance, structure, articulation, or visibility of the object of interest itself, segmentation-based filtering may become less reliable. These limitations point to several promising directions for future work. A first direction is to make the semantic interface more adaptive, for example through prompt refinement or prompt selection conditioned on the evolving task and state distribution. A second direction is to move beyond purely offline distillation toward iterative or online refinement of the masker during policy learning, so that the perceptual prior can be corrected when the agent encounters previously unseen but task-relevant structures. More broadly, it would be valuable to evaluate TaLaS in settings involving delayed object discovery, long-horizon interaction, stronger object-level distribution shifts, and real-world robotic data with dynamic backgrounds and sensor noise. Extending the framework to multimodal observations, where language-guided masking is combined with proprioceptive or tactile feedback, is another natural step toward more robust and deployment-ready embodied learning systems.

References

Abdulaziz Almuzairee, Nicklas Hansen, and Henrik I Christensen. A recipe for unbounded data augmentation in visual reinforcement learning. *RLJ*, 2024.

David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35, 2022.

Chao Chen, Jiacheng Xu, Weijian Liao, Hao Ding, Zongzhang Zhang, Yang Yu, and Rui Zhao. Focus-then-decide: segmentation-assisted reinforcement learning. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 2024.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Int. Conf. on Machine Learning*. PMLR, 2019.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2025.

Vedant Dave and Elmar Rueckert. Denoised predictive imagination: An information-theoretic approach for learning world models. In *17th European Workshop on Reinforcement Learning*, 2024.

Bram Grooten, Tristan Tomilin, Gautham Vasan, Matthew E Taylor, A Rupam Mahmood, Meng Fang, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Madi: Learning to mask distractions for generalization in visual deep reinforcement learning. In *Proc. of the 23rd Int. Conf. on Autonomous Agents and Multiagent Systems*, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *Int. Conf. on Learning Representations*, 2018.

Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE Int. Conf. on Robotics and Automation*. IEEE, 2021.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *The 12th Int. Conf. on Learning Representations*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2016.

Sili Huang, Yanchao Sun, Jifeng Hu, Siyuan Guo, Hechang Chen, Yi Chang, Lichao Sun, and Bo Yang. Learning generalizable agents via saliency-guided features decorrelation. *Advances in Neural Information Processing Systems*, 36, 2023.

Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. Spectrum random masking for generalization in image-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Yiding Jiang, J Zico Kolter, and Roberta Raileanu. On the importance of exploration for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 1998. ISSN 0004-3702.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, 2023.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76, 2023.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *Proc. of the 37th Int. Conf. on Machine Learning*, 2020a.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems*, 33, 2020b.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17, 2016.

Lu Li, Jiafei Lyu, Guozheng Ma, Zilin Wang, Zhenjie Yang, Xiu Li, and Zhiheng Li. Normalization enhances generalization in visual reinforcement learning. In *Proc. of the 23rd Int. Conf. on Autonomous Agents and Multiagent Systems*, 2024.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, 2023.

Gary Lupyan. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in psychology*, 3, 2012.

Gary Lupyan and Emily J. Ward. Language can boost otherwise unseen objects into visual awareness. *Proc. of the National Academy of Sciences*, 110, 2013.

Guozheng Ma, Zhen Wang, Zhecheng Yuan, Xueqian Wang, Bo Yuan, and Dacheng Tao. A comprehensive survey of data augmentation in visual reinforcement learning. *Int. Journal of Computer Vision*, 2025.

Shahin Nasr, Ali Moeeny, and Hossein Esteky. Neural correlate of filtering of irrelevant information from visual working memory. *PLoS One*, 3, 2008.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Proc. of Robotics: Science and Systems*, 2018.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The 13th Int. Conf. on Learning Representations*, 2025.

Katharina N Seidl, Marius V Peelen, and Sabine Kastner. Neural evidence for distracter suppression during visual search in real-world scenes. *Journal of Neuroscience*, 32, 2012.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Manan Tomar, Riashat Islam, Matthew Taylor, Sergey Levine, and Philip Bachman. Ignorance is bliss: Robust control via information gating. *Advances in Neural Information Processing Systems*, 36, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

Che Wang, Xufang Luo, Keith Ross, and Dongsheng Li. Vrl3: A data-driven framework for visual deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Shuo Wang, Zhihao Wu, Jinwen Wang, Xiaobo Hu, Youfang Lin, and Kai Lv. How to learn domain-invariant representations for visual reinforcement learning: an information-theoretical perspective. In *Proc. of the 33rd Int. Joint Conf. on Artificial Intelligence*, 2024.

Ziyu Wang, Yanjie Ze, Yifei Sun, Zhecheng Yuan, and Huazhe Xu. Generalizable visual reinforcement learning with segment anything model. *arXiv preprint arXiv:2312.17116*, 2023.

Hyejin Yang and Gregory J Zelinsky. Visual search is guided to categorically-defined targets. *Vision research*, 2009.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *Int. Conf. on Learning Representations*, 2021.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *Int. Conf. on Learning Representations*, 2022.

Bang You, Jingming Xie, Youping Chen, Jan Peters, and Oleg Arenz. Self-supervised sequential information bottleneck for robust exploration in deep reinforcement learning. *arXiv preprint arXiv:2209.05333*, 2022.

Tao Yu, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Mask-based latent reconstruction for reinforcement learning. *Advances in Neural Information Processing Systems*, 35, 2022.

Zhecheng Yuan, Guozheng Ma, Yao Mu, Bo Xia, Bo Yuan, Xueqian Wang, Ping Luo, and Huazhe Xu. Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning. In *Proc. of the 31st Int. Joint Conf. on Artificial Intelligence*, 2022a.

Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35, 2022b.

Zhecheng Yuan, Sizhe Yang, Pu Hua, Can Chang, Kaizhe Hu, and Huazhe Xu. Rl-vigen: A reinforcement learning benchmark for visual generalization. *Advances in Neural Information Processing Systems*, 36, 2023.

Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022.

Di Zhang, Bowen Lv, Hai Zhang, Feifan Yang, Junqiao Zhao, Hang Yu, Chang Huang, Hongtu Zhou, Chen Ye, et al. Focus on what matters: Separated models for visual-based rl generalization. *Advances in Neural Information Processing Systems*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2018.

A Appendix

A.1 Masker Implementation Details

A.1.1 SAMWISE

We use SAMWISE, built atop the SAM2 (Ravi et al., 2025) backbone, to generate language-conditioned video masks. It integrates temporal and text cues via a Cross-Modal Temporal (CMT) adapter at intermediate layers and a Conditional Memory module to reduce tracking bias, all without fine-tuning SAM2 or relying on external VLMs. The added overhead is minimal (∼4.2–4.9M parameters), and the model operates in a streaming setup. A short task-specific prompt is used to obtain binary mask proposals.

A.1.2 Natural language Prompts

For each environment, we define a fixed set of task-level prompts at the beginning of training, before the mask-distillation stage and before policy learning. The prompt is manually specified from the task description and names the controllable agent component and the task-relevant object or objects. The prompt is used only to query the frozen language-conditioned segmenter during Phase I. After the masks are distilled into the lightweight masker, the prompt interface and the segmentation model are removed from the reinforcement learning pipeline. Table 3 lists the prompts used for all environments in our experiments.

Table 3: Task-level prompts used for language-conditioned mask generation. Prompts are specified once at the beginning of training and kept fixed across frames, episodes, seeds, and evaluation domains.

Benchmark	Task	Prompt(s)
DMC-GB	Cartpole Swingup	"Show the orange cartpole"
DMC-GB	Walker Walk	"Show the orange walker"
DMC-GB	Walker Stand	"Show the orange walker"
DMC-GB	Ball in Cup Catch	"Show the orange cup", "Show the orange ball"
DMC-GB	Finger Spin	"Show the orange finger", "Show the orange rectangle"
DMC-GB	Cheetah Run	"Show the orange cheetah"
Unitree	Unitree Walk	"Show the black robot dog"
Unitree	Unitree Stand	"Show the black robot dog"
Adroit	Pen	"Show the yellow robot hand", "Show the pen"
Adroit	Door	"Show the yellow robot hand", "Show the orange door", "Show the door handle"
Adroit	Hammer	"Show the yellow robot hand", "Show the hammer", "Show the red nail block"

A.2 Learning Algorithm Implementation Details

A.2.1 RL Hyperparameters

We follow RL-ViGen (Yuan et al., 2023) defaults, using ResNet encoders (He et al., 2016) for PIE-G. For TaLaS, we set a masker learning rate of 10^{-4}, 2 k-step exploration phase (T_{exp}), 2 k masker pretraining steps, and batch sizes of 128 and 32 for RL and masker training, respectively. For SAMWISE (Cuttano et al., 2025), we retain its default setup, adjusting the input resolution to 90×90 for Hammer and Door in Adroit.

A.2.2 Masker Sigmoid

The masker is a compact CNN with three convolutional layers of channel dimensions $3 \to 64 \to 32 \to 1$, using kernel sizes 7×7, 5×5, and 3×3, respectively, with padding to maintain spatial resolution. Each layer is followed by Group Normalization and ReLU, except the final, which uses a temperature-controlled

Table 4: Prompt variants used for prompt-sensitivity analysis. The default prompt corresponds to the prompt used in the main experiments. Coarse prompts remove color cues and reduce semantic specificity. IoU is computed against human-annotated task-relevant masks over 10 randomly sampled unaugmented frames per task.

Task	Default prompt(s)	Coarse prompt(s)
Cartpole Swingup	"Show the orange cartpole"	"Show the cartpole"
Walker Walk	"Show the orange walker"	"Show the walker"
Walker Stand	"Show the orange walker"	"Show the walker"
Ball in Cup Catch	"Show the orange cup", "Show the orange ball"	"Show the cup and ball"
Finger Spin	"Show the orange finger", "Show the orange rectangle"	"Show the finger and object"
Cheetah Run	"Show the orange cheetah"	"Show the cheetah"
Unitree Walk	"Show the black robot dog"	"Show the dog"
Unitree Stand	"Show the black robot dog"	"Show the dog"
Pen	"Show the yellow robot hand", "Show the pen"	"Show the hand and object"
Door	"Show the yellow robot hand", "Show the orange door", "Show the door handle"	"Show the hand and door"
Hammer	"Show the yellow robot hand", "Show the hammer", "Show the red nail block"	"Show the hand and object"

sigmoid $\max(0, \min(1, 1/(1+e^{-sx})))$ with slope $s = 5$ ($\tau = 0.2$), producing near-binary, differentiable masks that stabilize gradients.

A.3 Augmentations

All baselines except SAC employ data augmentations. TaLaS uses SVEA-style overlay augmentation (Hansen et al., 2021) with random images from Places (Zhou et al., 2018) dataset: $\tilde{o}_t = \delta \cdot o_t + (1 - \delta)n$, where n is a randomly selected background and $\delta = 0.5$. We use the same augmentation setting across all methods to ensure a controlled and fair comparison. The coefficient δ controls the strength of the visual corruption: larger values preserve more of the original observation, whereas smaller values increase the contribution of the distractor image. TaLaS is not architecturally tied to $\delta = 0.5$, since the student masker is trained to recover the clean teacher mask from corrupted observations. However, we do not claim invariance to arbitrary overlay strengths. Evaluating performance across different values of δ is a useful sensitivity analysis and is left for future work.

A.4 Ablation Study

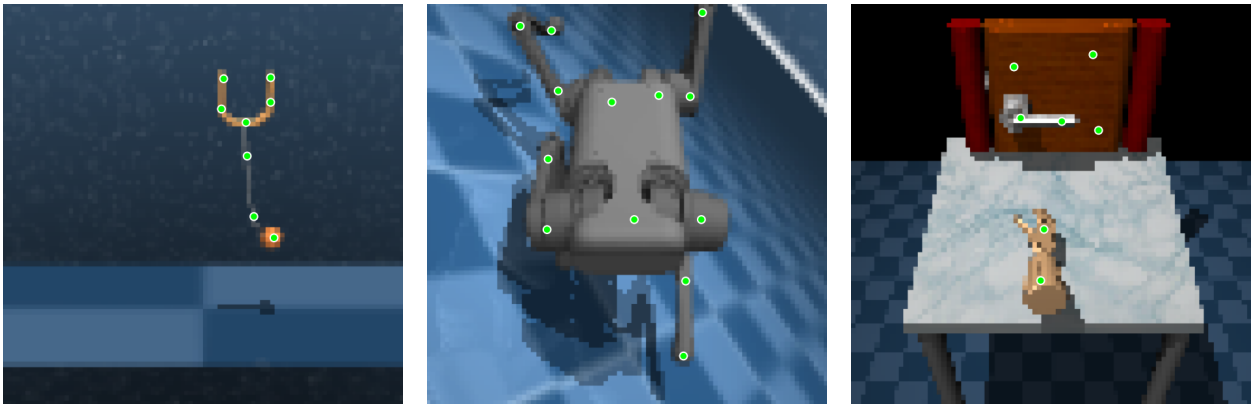
A.4.1 Prompt Ablation

To quantify prompt sensitivity, we sample 10 random frames per task from the unaugmented (original) setting and manually annotate task-relevant regions as human ground-truth masks. We then compute the intersection-over-union (IoU) between each prompt-induced mask and the corresponding human annotation. All masks are binarized before IoU computation.

Figure 6 illustrates the human-annotation protocol used to obtain the ground-truth masks. The prompt variants are listed in Table 4, the per-task IoU scores are reported in Table 5, and qualitative prompt-sensitivity examples are shown in Figure 7.

Table 5: Per-task prompt sensitivity under unaugmented observations. IoU is computed against human-annotated task-relevant masks using 10 randomly sampled frames per task. The default prompt corresponds to the prompt used in the main experiments.

Task	Default prompt IoU	Coarse prompt IoU
Cartpole Swingup	0.8394	0.0209
Walker Walk	0.6399	0.1799
Walker Stand	0.8064	0.6694
Ball in Cup Catch	0.6059	0.6173
Finger Spin	0.7808	0.6856
Cheetah Run	0.8840	0.6853
Unitree Walk	0.7270	0.4092
Unitree Stand	0.4039	0.4315
Pen	0.5008	0.1949
Door	0.5897	0.0016
Hammer	0.8307	0.3835
Average	0.6917	0.3890



DMC: Ball-in-Cup Catch Unitree Stand Adroit: Door

Figure 6: Human-annotation for prompt-sensitivity evaluation. We manually annotate task-relevant regions on 10 randomly sampled unaugmented frames per task and use these annotations as ground-truth masks for IoU computation. The examples show annotations from Ball-in-Cup Catch, Unitree Stand, and Hammer.

A.4.2 Qualitative Mask Analysis under Video-Hard Distractors

We provide additional qualitative mask examples to analyze the behavior of TaLaS under strong visual distractors. Figure 8 shows video-hard observations and the corresponding masked observations across representative tasks. The masked views preserve the controllable agent and task-relevant objects while suppressing background distractors.

Figure 7: Qualitative prompt-sensitivity examples on unaugmented observations. Each row corresponds to one representative task. Columns show the original observation, human-annotated ground-truth mask, default-prompt mask, and coarse-prompt mask. The rows correspond to: first row, Walker Walk; second row, Finger Spin; third row, Hammer; fourth row, Door; and fifth row, Unitree Walk.

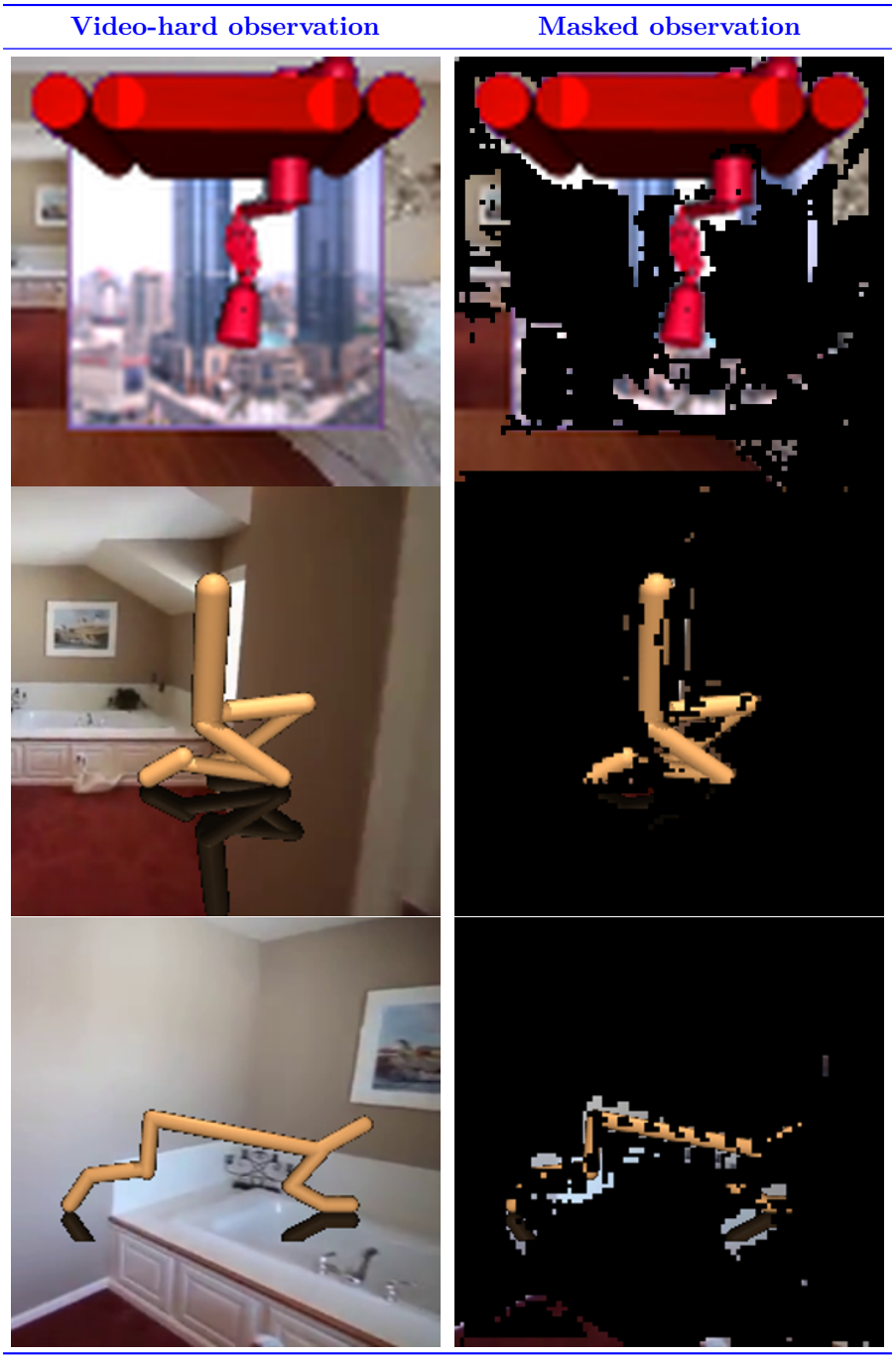


Figure 8: Qualitative mask examples under video-hard distractors. The rows correspond to: first row, *Door Opening*; second row, *Walker Walk*; third row, *Cheetah Run*. Columns show the original video-hard observation and the corresponding masked observation produced by TaLaS.