SALVO: Training World Models in VLM Perceptual Space for Semantically-Grounded Control

1 Observed Phenomenon

Generative world models learn by compressing high-dimensional observations into a low-dimensional latent space, typically using a pixel-level reconstruction loss as a primary learning signal. However, this objective forces the model to expend capacity on high-frequency visual details that are often irrelevant for semantic understanding. This "irreversible semantic compression" means that subtle but task-critical visual cues (e.g., the difference between a "limp" and a "stroll") that are easily distinguishable by a Vision-Language Model (VLM) can be lost in the world model's latent representation.

2 Problem Statement

Current world models, trained to be good at reconstructing pixels, are not explicitly optimized to preserve the semantic information that modern VLMs use to interpret the world. This creates a semantic gap between the world model's understanding of the environment and the VLM's interpretation of a task prompt.

Condition: We don't yet have world models whose representation learning is fundamentally guided by a rich, pre-trained perceptual metric, rather than a pixel-based one. The world model's learning objective is misaligned with the semantic nature of VLM-provided goals.

Consequence: This misalignment leads to an impoverished imaginative faculty. The agent's imagined trajectories may lack the necessary semantic detail to accurately plan for complex, language-defined tasks. This results in policies that fail to capture the nuances of the instructions, leading to poor generalization and suboptimal performance on tasks requiring fine-grained visual understanding.

3 Motivation

State-of-the-art methods like "Dream to Control" (Dreamer) build powerful world models based on reconstruction and dynamics prediction, but their representations are semantically agnostic. "GenRL" brilliantly addresses task specification by connecting a pre-trained VLM to a world model post-hoc via a "latent connector". However, this connector must bridge the gap between the VLM's rich semantic space and the world model's potentially semantically-sparse latent space.

SALVO is motivated by a simple but profound question: What if the world model's learning objective was not to fool the eye, but to fool a VLM? We propose to replace the conventional pixel-level reconstruction loss entirely with a perceptual reconstruction loss defined within the embedding space of a frozen, pre-trained VLM. Instead of forcing the model's latents s_t to reconstruct the raw observation x_t, we train them to reconstruct the VLM's **perception** of x_t.

This approach is hypothesized to be superior because it directly aligns the world model's optimization with the semantic concepts embedded in the VLM. It encourages the latent space to be inherently structured around VLM-salient features, thereby learning a more semantically meaningful model of the world's

dynamics. This should not only lead to more accurate and nuanced imagined rollouts for VLM-specified tasks but also simplify the subsequent learning of the GenRL-style latent connector.

4 Hypothesis

Training a generative world model by replacing the pixel-reconstruction loss with a perceptual loss in a VLM's embedding space will create a world model that:

1. Produces reconstructions that are semantically richer, as measured by a held-out Visual Question Answering (VQA) model.
2. Maintains sufficient physical fidelity for accurate dynamics prediction in imagination.
3. Enables an agent to achieve higher task performance and demonstrate superior zero-shot generalization to novel, language-specified stylistic variations compared to models trained with pixel-reconstruction or with semantics as only an auxiliary objective.

5 Proposed Method

SALVO re-purposes the generative world model architecture (from Dreamer/GenRL) for semantic, rather than visual, fidelity.

1. **Base World Model Architecture:** We use a standard architecture with an encoder ($q_\phi(s_t \mid x_t)$), a sequence model (e.g., GRU), a dynamics predictor ($p_\phi(s_t \mid h_t)$), and a decoder ($p_\phi(\hat{x}_t \mid s_t)$).
2. **Frozen Foundation VLM:** A pre-trained VLM (e.g., InternVideo2) f_{VLM} provides a fixed, rich perceptual mapping from images to embeddings.
3. **Perceptual Reconstruction Loss:** The core of SALVO. We discard the traditional pixel-space L_{recon}. The new objective for the encoder/decoder (ϕ) is to minimize the distance between the VLM embedding of the original image and the VLM embedding of the reconstructed image:

   $$L_{\text{perceptual}} = \left\| \text{sg}\big(f_{\text{VLM}}(x_t)\big) - f_{\text{VLM}}\big(p_\phi(\hat{x}_t \mid s_t)\big) \right\|^2$$

   The gradient from this loss flows back through f_{VLM}, the decoder p_ϕ, and the encoder q_ϕ, forcing the entire system to operate in a way that respects the VM's perceptual space.
4. **Combined World Model Training Objective:** The world model is trained to jointly predict dynamics in latent space and reconstruct in perceptual space.

   $$L_{\text{SALVO}} = \lambda_{\text{dyn}} L_{\text{dyn}} + \lambda_{\text{perc}} L_{\text{perceptual}}$$

   where $L_{\text{dyn}} = D_{\text{KL}}\big[q_\phi(s_t \mid x_t) \,\|\, p_\phi(s_t \mid h_t)\big]$ is the standard dynamics consistency loss from Dreamer/GenRL, and λ are balancing weights.
5. **Curriculum Learning:** To prevent the perceptual loss from overwhelming the learning of stable dynamics initially, we propose a curriculum schedule. Training starts with λ_{dyn} high and λ_{perc} low, gradually shifting the emphasis to perceptual reconstruction as the dynamics model stabilizes.
6. **Task Specification and Policy Learning:** Following the training of the SALVO world model, we adopt the policy learning framework from GenRL (Section 3.3). A latent Connector and Aligner are trained to map language/visual prompts to target latent state sequences $s_{\text{task\_seq}}$ in SALVO's now-semantic latent space. An actor-critic policy is then trained in imagination to match these target trajectories, using the cosine-distance reward from GenRL (Eq. 3).

6 Experimental Outline

1. **Datasets & Environments:** DMControl Suite (Walker, Cheetah, Quadruped, Stickman) and Kitchen, using vision-only, reward-free datasets as in GenRL.

2. **Models & Baselines:**

   - SALVO (Proposed): World model trained with L_{SALVO} (dynamics + perceptual loss).
   - Baseline 1 (GenRL): The method from Paper 1, using $L_{\text{recon}} + L_{\text{dyn}}$.
   - Baseline 2 (Parameter-Matched GenRL): GenRL with an oversized decoder, matching the total parameter count of SALVO's decoder + VLM's encoder (for the forward pass on $\hat{x}_t$). This controls for performance gains due to model size.
   - Baseline 3 (SALVO-Aux): Our original idea. A world model trained with $L_{\text{recon}} + L_{\text{dyn}} + \beta\, L_{\text{semantic\_aux}}$, where $L_{\text{semantic\_aux}}$ is an auxiliary loss from a small MLP predicting $f_{\text{VLM}}(x_t)$ from s_t. This tests if **augmenting** is as good as **replacing**.

3. **Experimental Phases:**

   - Phase 1 (WM Training): Train all world models (SALVO and baselines) on the offline dataset. Pre-compute and store VLM embeddings $f_{\text{VLM}}(x_t)$ for all training data to manage computational cost.
   - Phase 2 (Connector/Policy Training): For each trained world model, train a task-prompt connector and an actor-critic policy in imagination for a suite of in-distribution tasks (e.g., "run fast," "stand on one foot").

4. **Primary & Secondary Metrics:**

   - Primary Metric: Average episodic return on a set of held-out in-distribution and out-of-distribution language/visual prompts.
   - Secondary Metrics:
     - Zero-Shot Generalization: Evaluate all models on a curated set of novel prompts describing stylistic variations not present in the training data (e.g., "walk with a limp," "prance like a pony," "perform a hesitant pirouette").
     - Semantic Richness (VQA Analysis): Generate reconstructions $\hat{x}_t$ from all models. Feed these reconstructions into a frozen, third-party VQA model (e.g., LLaVA) and ask questions about the scene (e.g., "Is the agent balanced?", "Are the agent's arms raised?"). Report VQA accuracy.
     - Dynamics Fidelity Analysis: For SALVO, plot the open-loop dynamics prediction error (MSE in latent space over 50 steps) against the perceptual loss weight λ_{perc} to visualize the trade-off with physical accuracy.

5. **Ablation Studies:**

   - Loss Function Form: Compare the MSE-based $L_{\text{perceptual}}$ with a cosine distance alternative.
   - Contrastive Alignment: As a more advanced ablation, implement a contrastive version of SALVO-Aux. An MLP projects s_t into the VLM space, and a contrastive loss (InfoNCE) is used to pull this projection toward the true $f_{\text{VLM}}(x_t)$ (positive) and push it away from embeddings of other states in the batch (negatives).
   - Curriculum Learning: Compare performance with and without the curriculum for λ weights.

7 Concrete Example

Consider the language prompt "robot performing a graceful pirouette."

- **Baseline (GenRL):** The world model, optimized for pixel reconstruction, learns to represent "turning." Its gradient is driven by minimizing pixel differences. It has no intrinsic concept of "grace." The latent connector maps the VLM's embedding for "graceful pirouette" to the closest available "turning" trajectory in this semantically-crude latent space. The resulting policy might turn, but likely in a mechanically efficient, not graceful, way.
- **SALVO (Proposed Method):** SALVO's world model is optimized via $L_{\text{perceptual}}$. Its learning gradient is directly shaped by the VLM's assessment. If the VLM distinguishes between clumsy and graceful turns in its embedding space, SALVO's decoder is explicitly forced to generate reconstructions that capture this distinction. The latent states s_t therefore must encode the necessary physical information (limb extension, smooth velocity curves) to produce a "graceful-looking" reconstruction. When the policy is trained in this world, it learns to control these semantically-meaningful latents, resulting in a motion that is far more likely to match the V-L's nuanced understanding of "graceful."

8 Potential Pitfalls & Mitigations

- **Risk 1 – Semantic Dominance vs. Physical Accuracy:** The perceptual loss might be prioritized at the expense of physically plausible dynamics, leading to "pretty but impossible" imagined rollouts.
  - **Fallback:** The L_{dyn} term, inherited from Dreamer (Paper 2), acts as a strong regularizer, anchoring the model to real-world transitions. The curriculum learning approach for λ weights is designed to establish stable dynamics first. The "Dynamics Fidelity Analysis" will quantify this trade-off.
- **Risk 2 – VLM Inductive Biases:** The world model will inherit any biases, blind spots, or artifacts present in the frozen VLM.
  - **Fallback:** This is an inherent risk of using foundation models. The L_{dyn} term again provides a crucial grounding in reality. While out-of-scope for this project, future work could mitigate this by using an ensemble of diverse VLMs for the perceptual loss.
- **Risk 3 – Computational Cost:** The forward pass through a large VLM $f_{\text{VLM}}\big(p_\phi(\hat{x}_t \mid s_t)\big)$ inside the training loop adds overhead.
  - **Fallback:** We will pre-compute and store the target VLM embeddings $f_{\text{VLM}}(x_t)$ for the entire dataset offline. The per-step overhead is then one VLM forward pass on the reconstructed image, which is significant but manageable on modern hardware. We will quantify this overhead in our experiments.
- **Risk 4 – Training Instability:** Replacing a well-understood loss like pixel-MSE with a complex, high-dimensional perceptual loss could lead to unstable training.
  - **Fallback:** We will carefully monitor gradients. The curriculum learning approach is the primary mitigation. We will also experiment with simpler loss formulations (cosine distance vs. MSE) and potentially use gradient clipping on $L_{\text{perceptual}}$.

References

[1] P. Mazzaglia, T. Verbelen, B. Dhoedt, A. Courville, and S. Rajeswar. *GenRL: Multimodal-foundation world models for generalization in embodied agents.* https://arxiv.org/abs/2406.18043, arXiv:2406.18043, 2024.

[2] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. *Dream to Control: Learning Behaviors by Latent Imagination.* https://arxiv.org/abs/1912.01603, arXiv:1912.01603, 2019.