Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

VISION-SLS: Safe Perception-Based Control from Learned Visual Representations via System Level Synthesis

Antoine P. Leeman*[1] Shuyu Zhan*[2] Melanie N. Zeilinger[1] Glen Chou[3]

Project Website: (Link) | **Code**: (Github) | **Video**: (YouTube)

Abstract—**We propose VISION-SLS, a method for nonlinear output-feedback control from high-resolution RGB images which provides robust constraint satisfaction guarantees under calibrated uncertainty bounds despite partial observability, sensor noise, and nonlinear dynamics. To enable scalability while retaining guarantees, we propose: (i) a learned low-dimensional observation map from pretrained visual features with state-dependent error bounds, and (ii) a causal affine time-varying output-feedback policy optimized via System Level Synthesis (SLS). We develop a scalable, novel solver for the resulting nonconvex program that leverages sequential convex programming coupled with efficient Riccati recursions. On two simulated visuomotor tasks (a 4D car and a 10D quadrotor) with $\geq 512 \times 512$ pixels and a 59D humanoid task with partial observability, our method enables safe, information-gathering behavior that reduces uncertainty while guaranteeing constraint satisfaction with empirically-calibrated error bounds. We also validate our method on hardware, safely controlling a ground vehicle from onboard images, outperforming baselines in safety rate and solve times. Together, these results show that learned visual abstractions coupled with an efficient solver make SLS-based safe visuomotor output-feedback practical at scale.**

* Equal contribution. [1]Institute for Dynamic Systems and Control, ETH Zürich {aleeman, mzeilinger}@ethz.ch. [23]Georgia Institute of Technology, Schools of [2]Interactive Computing, [3]Aerospace Engineering and Cybersecurity & Privacy {szhan45, chou}@gatech.edu. This work was supported in part by the European Space Agency OSIP 4000133352, an ETH Zurich Doc.Mobility Fellowship, and a Mathworks Microgrant.

I. Introduction

Reliable robot deployment demands safety and performance guarantees. Yet, perception-in-the-loop control, i.e., *output-feedback* control, makes both control synthesis and safety verification difficult: robots must act under partial observability and noisy sensor feedback while efficiently stabilizing their nonlinear dynamics. These challenges are magnified when feedback comes from high-dimensional, non-smooth visual sensor outputs. Due to these difficulties, existing methods typically trade off information-seeking (i.e., active reduction of state uncertainty by moving to regions with more informative sensor readings) under partial observability against safety certification. For instance, reinforcement learning (RL) [26] and imitation learning [9] scales to high-dimensional visuomotor tasks but yields policies that are difficult to certify. In contrast, model-based methods like model predictive control (MPC) [27, 7], Lyapunov-based control via sum-of-squares (SOS) [33, 10, 12] or belief-space planning [19] offer guarantees but are mainly limited to linear, low-dimensional settings or rely on the classical separation principle between estimation and control, thereby precluding information-seeking behaviors that are often necessary in robotics for robustness to visual occlusions.

To close this gap, we present VISual Output-feedback Nonlinear System Level Synthesis (VISION-SLS), a scalable

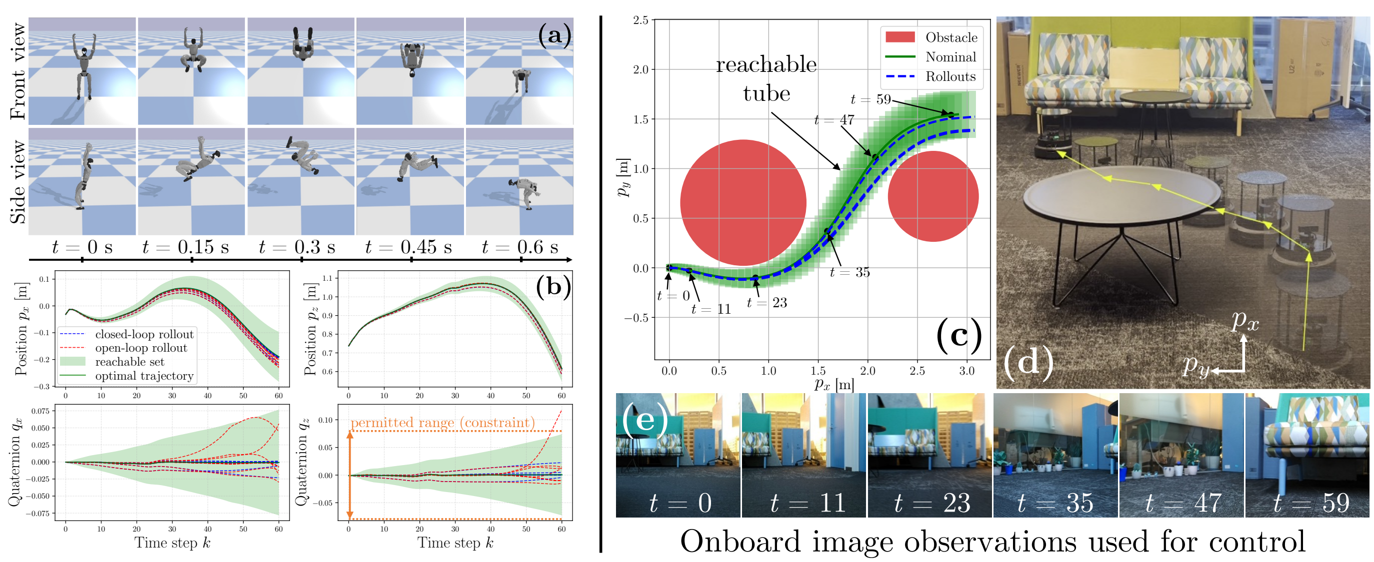

Fig. 1. We stabilize a Unitree G1 humanoid (59 states) around a backflip trajectory that is jointly optimized, enabling robust constraint satisfaction. (a) Two views of a timelapse of the backflip trajectory that our method stabilizes. (b) A subset of the reachable tubes (green), closed-loop rollouts (blue), and open-loop rollouts (red). The optimized closed-loop controller via Alg. 1 keeps the system within the reachable tubes, whereas executing the nominal trajectory open-loop (i.e., without feedback) causes the system to exit the tubes and result in constraint violations (e.g., on the quaternion q_z). See App. D6 for reachable tubes for all states. (c) We evaluate our method on a physical TurtleBot robot, controlled via onboard images. Closed-loop rollouts obtained by using our controller (blue) stay within the tubes (green), avoiding obstacles (red). (d) Rollout time-lapse. (e) Onboard images collected to implement our controller.

method for uncertainty-aware output-feedback control from high-dimensional visual inputs, with robust constraint satisfaction guarantees under calibrated uncertainty bounds. VISION-SLS does not rely on the separation principle and performs active information-gathering with safety guarantees. To achieve this, VISION-SLS integrates learned visual representations into system level synthesis (SLS)-based robust control design. Our key idea is to model perception error as a bounded disturbance on a reduced observation, allowing partial observability, information gathering, and robust constraint satisfaction to be addressed directly and jointly within the closed-loop response parameterization optimized by SLS. This enables scalable output-feedback synthesis for nonlinear systems with high-dimensional sensor data. Our main contributions are:

- We develop a novel controller parameterization for partially observed *nonlinear* systems by coupling nonlinear trajectory optimization with output-feedback SLS. In contrast to prior nonlinear SLS results, which focus on state feedback [25, 24], and prior output-feedback SLS results [48, 2], which focus on linear or linear time-varying (LTV) systems, our method enables robust perception-based control with safety guarantees under nonlinear dynamics (Prop. 3).
- We construct a low-dimensional reduced observation from foundation-model features using a dynamics-aware observability loss, and calibrate a state-dependent overbound on the resulting reduction error (Sec. VI-C1–VI-C2). This makes high-dimensional visual observations compatible with our robust output-feedback synthesis.
- We show unconstrained output-feedback linear-quadratic-Gaussian (LQG) in the SLS parameterization has a double-Riccati solution. This structure enables an efficient solver for the output-feedback subproblems in our method (Prop. 4) that scales synthesis to large systems.
- We validate the approach on a 4D car and a 10D quadrotor with RGB camera observations, a 59D humanoid with partial state measurements, and a hardware TurtleBot with an onboard camera, showing safe information-gathering behavior, scalability to high-dimensional dynamics, and successful vision-based deployment on hardware.

II. Related Work

RL has achieved impressive performance in visuomotor control, scaling to high-dimensional tasks and enabling end-to-end pixel-to-torques policies [26, 41, 31, 6, 46]. However, RL methods are notoriously data-hungry and difficult to certify, as the resulting policies are complex and lack interpretability. While visual foundation models (VFMs) [28, 1, 47] have recently improved generalization by providing rich feature representations, coupling them with RL still leaves the overall perception–control pipeline without calibrated guarantees.

SLS offers a principled, scalable model-based control framework for large LTV systems for output feedback [2, 14, 29], with recent extensions to nonlinear state-feedback under robust constraints [25]. Belief-space planning (BSP) instead casts output-feedback as a partially-observable Markov decision process (POMDP) to model sensor noise and information gathering, but exact solutions are intractable and practical Gaussian/sampling-based approximations (e.g., based on the extended Kalman filter (EKF) or iterative LQG) typically do not consider constraints under uncertain dynamics [19, 15]. Finally, a line of work on safe control with learned models and visual inputs has emerged [37, 10], but only heuristically encourages safety or is limited to low-dimensional systems.

In summary, prior visuomotor control work typically lacks end-to-end guarantees, assumes full state reconstruction from observations, or does not handle robust constraint satisfaction under nonlinear dynamics. These limitations motivate our approach, which combines visual abstractions from VFMs with dynamic output-feedback control using SLS, enabling scalable, information-gathering, and safe visuomotor control.

a) Notation: For vectors or matrices a and b with the same number of rows, we denote their horizontal concatenation by $[a,\ b]$, and their vertical stacking as $(a, b) = [a^\top,\ b^\top]^\top$. Let $0_{n\times m}$ and I_n denote the $n\times m$-zeros and $n\times n$ identity matrix, respectively, with dimensions inferred from context when omitted. We use n_x, n_u, and n_y for the dimensions of state, input, and observation. The matrix $\mathbf{Z}$ denotes the block downshift operator, with I_{n_x} on the first subdiagonal and zeros elsewhere. For sets, $A \oplus B$ denote the Minkowski sum. For a vector $a \in \mathbb{R}^n$, we denote as a_i its ith entry. We use the unit ℓ_∞ ball $\mathcal{B}^n := \{z \in \mathbb{R}^n : \|z\|_\infty \leq 1\}$. We use $\mathbb{N}_{0:T} := \{0, \ldots, T\}$, and use the Frobenius norm as $\|\cdot\|_\mathrm{F}$.

III. Problem Statement

We consider discrete-time, partially-observed, nonlinear systems, with states $x \in \mathbb{R}^{n_\mathrm{x}}$, controls $u \in \mathbb{R}^{n_\mathrm{u}}$, and observations $y \in \mathbb{R}^{n_\mathrm{y}}$,

$$x_{k+1} = f(x_k, u_k) + E(x_k)w_k \tag{1a}$$
$$y_{k+1} = h(x_k) + F(x_k)e_k, \tag{1b}$$

where $f : \mathbb{R}^{n_\mathrm{x}} \times \mathbb{R}^{n_\mathrm{u}} \to \mathbb{R}^{n_\mathrm{x}}$ are the dynamics and $h : \mathbb{R}^{n_\mathrm{x}} \to \mathbb{R}^{n_\mathrm{y}}$ are the delayed sensor outputs (IMU, images, etc), accounting for processing latency common in vision-based feedback. The functions $E : \mathbb{R}^{n_\mathrm{x}} \to \mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$, $F : \mathbb{R}^{n_\mathrm{x}} \to \mathbb{R}^{n_\mathrm{y}\times n_\mathrm{y}}$ capture process noise and measurement noise on the system dynamics and observations, respectively.

Assumption 1. *We assume bounded process and measurement noise (arising, e.g., from VFM error) with $w_t \in \mathcal{B}^{n_\mathrm{x}}$ and $e_t \in \mathcal{B}^{n_\mathrm{y}}$, with a system-specific scaling via $E(x_k)$ and $F(x_k)$.*

Assumption 2. *We assume the dynamics f are twice continuously differentiable.*

We aim to guarantee that all reachable states and controls satisfy the constraint

$$\mathcal{S} = \{(x, u) \in \mathbb{R}^{n_\mathrm{x}+n_\mathrm{u}} |\ c_i^\top (x, u) + b_i \leq 0,\ i = 1, ..., n_\mathrm{c}\} \tag{2}$$

despite the process noise and measurement noise. To ensure robust constraint satisfaction[1] over the prediction horizon, we seek an output-feedback policy that leverages the history of measurements to compute the control at time k. To obtain this policy, we formulate the following constrained non-convex optimization problem for obtaining an output-feedback policy $\boldsymbol{\pi} := (\pi_0(\cdot), \ldots, \pi_{T-1}(\cdot))$:

$$\begin{aligned}
\min_{\boldsymbol{\pi}} \quad & J_\pi(\boldsymbol{\pi}) && \text{(3a)}\\
\text{s.t.} \quad & x_{k+1} = f(x_k, u_k) + E(x_k) w_k, \quad k \in \mathbb{N}_{0:T-1}, && \text{(3b)}\\
& y_{k+1} = h(x_k) + F(x_k) e_k, \quad k \in \mathbb{N}_{0:T-1}, && \text{(3c)}\\
& x_0 \in \mathcal{X}_0, && \text{(3d)}\\
& u_k = \pi_k(y_1, \ldots, y_k), \quad \forall k \in \mathbb{N}_{0:T-1}, && \text{(3e)}\\
& (x_k, u_k) \in \mathcal{S}, \forall w_k \in \mathcal{B}^{n_x}, \forall e_k \in \mathcal{B}^{n_y}, k \in \mathbb{N}_{0:T-1}, && \text{(3f)}
\end{aligned}$$

where $J_\pi(\cdot)$ is the objective and $\mathcal{X}_0$ is the set of possible initial states. The optimization in (3) is difficult for three reasons:

- The optimization ranges over all causal maps (3e) from measurement histories to controls, and hence is infinite-dimensional unless parameterized.
- The constraints (3f) must hold for each of the infinitely-many possible disturbances w_k, e_k in the sets $\mathcal{B}^{n_x}$, $\mathcal{B}^{n_y}$.
- The dynamics and observation functions f, h are non-linear and high-dimensional. In particular, the observations are assumed to be high-dimensional images, i.e., $y_k \in \mathbb{R}^{n_y}$ with n_y in the order of 10^5 or more.

In this paper, we develop a tractable pipeline, VISION-SLS, to address these challenges. We will first summarize background preliminaries on SLS in Sec. IV and overview our method in Sec. V, which learns reduced observations and robustly synthesizes safe nonlinear output-feedback control (Sec. VI) via an efficient solver based on Riccati recursions and sequential convex programming (Sec. VII).

IV. Preliminaries

In this section, we review the SLS framework for LTV systems [2], which we use as a subroutine in Sec. VI-A to design a dynamic output-feedback controller for the nonlinear system (1). We consider a one-step measurement-delayed LTV system to capture perception latency typical of vision-based feedback

$$\begin{aligned}
x_0 &= \bar{x}_0 + \Xi \tilde{w},\\
x_{k+1} &= A_k x_k + B_k u_k + E_k w_k, \qquad (4)\\
y_{k+1} &= C_k x_k + F_k e_k,
\end{aligned}$$

where $\tilde{w} \in \mathcal{B}^{n_x}$ denotes a normalized initial state uncertainty, and $\Xi \in \mathbb{R}^{n_x \times n_x}$ denotes the corresponding scaling matrix. The corresponding nominal dynamics evolve according to

$$\bar{x}_{k+1} = A_k \bar{x}_k + B_k \bar{u}_k, \quad \bar{y}_{k+1} = C_k \bar{x}_k. \qquad (5)$$

We define the stacked variables $\mathbf{x} = (x_0, \ldots, x_T)$, $\mathbf{u} = (u_0, \ldots, u_{T-1})$, $\mathbf{y} = (y_1, \ldots, y_T)$, $\mathbf{w} = (\tilde{w}, w_0, \ldots, w_{T-1})$, and $\mathbf{e} = (e_0, \ldots, e_{T-1})$, and block-diagonal matrices as $\mathbf{A} = \operatorname{blkdiag}(A_0, \ldots, A_{T-1})$, and similarly for $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{F}$, and $\mathbf{E} = \operatorname{blkdiag}(\Xi, E_0, \ldots, E_{T-1})$. We further define deviations $\Delta\mathbf{x} := \mathbf{x} - \bar{\mathbf{x}}$, $\Delta\mathbf{u} := \mathbf{u} - \bar{\mathbf{u}}$, $\Delta\mathbf{y} := \mathbf{y} - \bar{\mathbf{y}}$, and by stacking the dynamics over the time horizon, the closed-loop error system is written compactly as

$$\begin{aligned}
\Delta\mathbf{x} &= \mathbf{Z}\mathbf{A}\Delta\mathbf{x} + \mathbf{Z}\mathbf{B}\,\Delta\mathbf{u} + \mathbf{E}\mathbf{w}, && \text{(6a)}\\
\Delta\mathbf{y} &= \mathbf{Z}\mathbf{C}\Delta\mathbf{x} + \mathbf{F}\mathbf{e}. && \text{(6b)}
\end{aligned}$$

We consider an output-feedback controller of the form

$$\mathbf{u} = \bar{\mathbf{u}} + \mathbf{K}(\mathbf{y} - \mathbf{Z}\mathbf{C}\bar{\mathbf{x}}). \qquad (7)$$

Optimizing $\mathbf{K}$ directly is difficult since enforcing robust constraint satisfaction for the realized trajectory is nonconvex in $\mathbf{K}$. In contrast, SLS designs the closed-loop response maps $\boldsymbol{\Phi}$ directly, mapping disturbances to state and input deviations

$$\begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{u} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\Phi}^{\mathrm{xw}} & \boldsymbol{\Phi}^{\mathrm{xe}} \\ \boldsymbol{\Phi}^{\mathrm{uw}} & \boldsymbol{\Phi}^{\mathrm{ue}} \end{bmatrix}}_{\boldsymbol{\Phi}} \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix}, \qquad (8)$$

where the matrices $\boldsymbol{\Phi}^{\mathrm{xw}}$, $\boldsymbol{\Phi}^{\mathrm{xe}}$, $\boldsymbol{\Phi}^{\mathrm{uw}}$, and $\boldsymbol{\Phi}^{\mathrm{ue}}$ are block-lower-triangular. Crucially, this output-feedback parameterization enables an efficient *convex* controller synthesis with robust constraint satisfaction, as we will see next.

Proposition 1 (Adapted from [2, 48])**.** *The error LTV system (6) in closed-loop with the controller (7) can be written as Eq. (8)* [2] *if and only if $\boldsymbol{\Phi}$ satisfies*

$$\begin{aligned}
\begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A} & -\mathbf{Z}\mathbf{B} \end{bmatrix} \boldsymbol{\Phi} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix},\\
\boldsymbol{\Phi} \begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A} \\ -\mathbf{Z}\mathbf{C} \end{bmatrix} &= \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}.
\end{aligned} \qquad (9)$$

The feedback gains $\mathbf{K}$ in (7) can be recovered from $\boldsymbol{\Phi}$ via

$$\mathbf{K} = \boldsymbol{\Phi}^{\mathrm{ue}} - \boldsymbol{\Phi}^{\mathrm{uw}} (\boldsymbol{\Phi}^{\mathrm{xw}})^{-1} \boldsymbol{\Phi}^{\mathrm{xe}}. \qquad (10)$$

Proof: See, e.g., [48, Prop. 1]. ■

We show that the associated linear–quadratic–Gaussian (LQG) problem [2, 40] can be written as the convex program

$$\begin{aligned}
\min_{\boldsymbol{\Phi}} \quad & \left\| \begin{bmatrix} \mathbf{Q}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}^{\mathrm{xw}} & \boldsymbol{\Phi}^{\mathrm{xe}} \\ \boldsymbol{\Phi}^{\mathrm{uw}} & \boldsymbol{\Phi}^{\mathrm{ue}} \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} \right\|_{\mathrm{F}}^2 \\
& + \left\| P^{1/2} \begin{bmatrix} \boldsymbol{\Phi}^{\mathrm{xw}}_T & \boldsymbol{\Phi}^{\mathrm{xe}}_T \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} \right\|_{\mathrm{F}}^2 && \text{(11a)}\\
\text{s.t.} \quad & \begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A} & -\mathbf{Z}\mathbf{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}^{\mathrm{xw}} & \boldsymbol{\Phi}^{\mathrm{xe}} \\ \boldsymbol{\Phi}^{\mathrm{uw}} & \boldsymbol{\Phi}^{\mathrm{ue}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}, && \text{(11b)}\\
& \begin{bmatrix} \boldsymbol{\Phi}^{\mathrm{xw}} & \boldsymbol{\Phi}^{\mathrm{xe}} \\ \boldsymbol{\Phi}^{\mathrm{uw}} & \boldsymbol{\Phi}^{\mathrm{ue}} \end{bmatrix} \begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A} \\ -\mathbf{Z}\mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, && \text{(11c)}
\end{aligned}$$

with $\mathbf{Q} = \operatorname{blkdiag}(Q, \ldots, Q)$ and $\mathbf{R} = \operatorname{blkdiag}(R, \ldots, R)$ the state and input cost matrices, respectively, with $Q \succeq 0$ and $R \succ 0$, and P the terminal cost corresponding to the last time step T, and last block-row $\boldsymbol{\Phi}^{\bullet}_T$ of $\boldsymbol{\Phi}^{\bullet}$. As for typical LQG design, this cost function in (11) corresponds to the expected value of the variance of the quadratic performance objective

[1] For simplicity, we present linear constraints; nonlinear constraints (e.g., obstacle avoidance) are handled via sequential linearizations, similarly in Section VII and [43]. Likewise, terminal constraints can also be imposed.

[2] In the stacked matrices $\boldsymbol{\Phi}^{\mathrm{xw}}$ and $\boldsymbol{\Phi}^{\mathrm{uw}}$, the block at index $j = 0$ corresponds to the initial condition, while indices $j \geq 1$ correspond to process disturbances injected at time $j - 1$.

under stochastic process and measurement noise [2]. In the following, we formulate the corresponding robust constraint satisfaction conditions for the output-feedback setting.

Proposition 2. *Consider (6) in closed loop with (7), and let* $\mathbf{\Phi}^{\mathrm{xw}}, \mathbf{\Phi}^{\mathrm{xe}}, \mathbf{\Phi}^{\mathrm{uw}}, \mathbf{\Phi}^{\mathrm{ue}}$ *satisfy (9). Then robust satisfaction of (2) for all k and all admissible disturbances holds if and only if*

$$\sum_{j=0}^{k} \|c_i^\top \mathbf{\Phi}^{\mathrm{w}}_{k,j} E_j\|_1 + \|c_i^\top \mathbf{\Phi}^{\mathrm{e}}_{k,j} F_j\|_1 + c_i^\top (z_k, v_k) + b_i \leq 0. \quad (12)$$

for $i = 1, \dots, n_c$ *with* $\mathbf{\Phi}^{\mathrm{w}}_{k,j} = (\mathbf{\Phi}^{\mathrm{xw}}_{k,j}, \mathbf{\Phi}^{\mathrm{uw}}_{k,j})$ *and* $\mathbf{\Phi}^{\mathrm{e}}_{k,j} = (\mathbf{\Phi}^{\mathrm{xe}}_{k,j}, \mathbf{\Phi}^{\mathrm{ue}}_{k,j})$.

Proof: The proof follows directly from the state-feedback result in [32]; we provide the full proof in Appendix A. ∎

V. Method Overview

In this section, we summarize how our method, VISION-SLS, solves the nonlinear output-feedback problem (3) (also see Fig. 2) after we introduce two approximations to make (3) tractable. First, instead of planning directly from high-dimensional images, we map each observation $y \in \mathbb{R}^{n_{\mathrm{y}}}$ to a low-dimensional reduced measurement $y^r \in \mathbb{R}^{n_r}$. Concretely, we compose a pre-trained DINO feature extractor [28] with a learned projection to obtain

$$\mathrm{p}^{\mathrm{dino}}(\cdot) : \mathbb{R}^{n_{\mathrm{y}}} \to \mathbb{R}^{n_r}$$

and a corresponding state-based reduced observation model

$$h^r(\cdot) : \mathbb{R}^{n_{\mathrm{x}}} \to \mathbb{R}^{n_r},$$

so that the reduced observation can be written as

$$y^r_{k+1} = \mathrm{p}^{\mathrm{dino}}(y_k) = h^r(x_k) + F^r(x_k) e_k. \quad (13)$$

Intuitively, $h^r(x_k)$ captures the control-relevant information in the image, while $F^r(x_k)e_k$ is an inflated, state-dependent uncertainty term that overbounds both sensor noise and the error introduced by compressing the image into a low-dimensional observation. This lets us plan efficiently in a reduced feature space, rather than directly in pixel space (see Sec. VI-C2).

Remark 1. *Although tailored for camera sensors, this work could be implemented with other sensors, such as LiDAR (see Sec. VIII), depth sensors, or joint encoders.*

Second, since optimizing over general causal policies is intractable, we restrict (3e) to causal *affine*, time-varying feedback policies around a jointly-optimized nominal state/control/measurement trajectory $(\boldsymbol{z}, \boldsymbol{v})$ [3], i.e.,

$$u_k = \bar{u}_k + K^0_k(x_0 - \bar{x}_0) + \textstyle\sum_{j=1}^{k-1} K_{k,j}(y^r_{j+1} - h^r(\bar{x}_j)), \quad (14)$$

with optimized gains $K_{k,j}$ (Sec. VI-A). The additional term $K^0_k(x_0 - \bar{x}_0)$ explicitly accounts for uncertainty in the initial condition $\bar{x}_0$, which cannot in general be inferred from the measurement history and therefore enters the closed-loop system as an independent disturbance channel. This form admits fast evaluation at runtime via matrix–vector products.

[3] We use the notation $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ to denote the nominal trajectory of the LTV dynamics, in contrast to $(\boldsymbol{z}, \boldsymbol{v})$ which denote the nominal trajectory of the nonlinear dynamics.

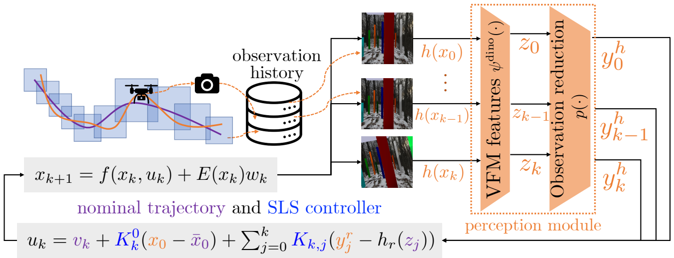

Fig. 2. **Online SLS execution:** Images are mapped to low-dimensional observations $p(\psi^{\mathrm{dino}}(y))$, approximated by a learned linear map $h^r(x) = Cx$ with bounded error. These observations feed into an SLS-based dynamic output-feedback controller, which combines the nonlinear nominal trajectory $(\boldsymbol{z}, \boldsymbol{v})$ and feedback gains K to generate safe inputs.

We detail our method in Sec. VI and VII. In Sec. VI-A, we develop a nonlinear SLS approach to the output-feedback policy in (3e) by combining nonlinear trajectory optimization with an SLS parameterization of the local closed-loop error dynamics, with robust constraint satisfaction (Sec. VI-B). In Sec. VI-C1, we use pre-trained visual features to learn a low-dimensional observation map h^r, together with error bounds in Sec. VI-C2, so that pixel observations can be used while retaining robust constraint satisfaction guarantees in Sec. VI-B. In Sec. VII, we propose an efficient iterative convexification algorithm to efficiently solve the resulting nonconvex program.

VI. Robust Nonlinear Control Design

In this section, we build upon the LTV SLS framework from Propositions 1–2 to synthesize a dynamic output-feedback controller for the nonlinear system (1) (Sec. VI-A) with robust constraint satisfaction guarantees (Sec. VI-B). Specifically, we jointly optimize the nominal trajectory $(\boldsymbol{z}, \boldsymbol{v})$, the closed-loop responses $\mathbf{\Phi}$ of the induced LTV system, and modeling-error overbounds for robust constraint satisfaction of the nonlinear system. Finally, we tractably implement the proposed output-feedback SLS framework for visual measurements by learning a reduced observation model (Sec. VI-C1).

A. Closed-loop error dynamics

Our robust control design leverages a nominal system

$$z_{k+1} = f(z_k, v_k), \quad k = 0, \dots, T-1 \quad (15)$$

where $z_k \in \mathbb{R}^{n_{\mathrm{x}}}$ is the nominal state, and $v_k \in \mathbb{R}^{n_{\mathrm{u}}}$ is the nominal input. We approximate the tracking error dynamics

$$\Delta x_{k+1} \doteq f(x_k, u_k) + E(x_k)w_k - f(z_k, v_k) \quad (16)$$

around the nominal trajectory $(\mathbf{z}, \mathbf{v})$ using a first-order Taylor expansion. This yields the LTV error system

$$\Delta x_{k+1} = A(z_k, v_k)\Delta x_k + B_k(z_k, v_k)\Delta u_k + \sigma(\tau_k, z_k, v_k), \quad (17)$$

$$\text{with } A_k := \frac{\partial f(x,u)}{\partial x}\Big|_{(z_k, v_k)}, \quad B_k := \frac{\partial f(x,u)}{\partial u}\Big|_{(z_k, v_k)}, \quad (18)$$

and the term $\sigma(\tau_k, z_k, v_k)$ captures the modelling error with $\tau \in \mathbb{R}$ satisfying

$$\left\| \begin{bmatrix} \Delta x \\ \Delta u \end{bmatrix} \right\|_\infty \leq \tau, \quad \Delta x = x - z, \quad \Delta u = u - v. \quad (19)$$

Assumption 3. *For each $k \in \mathbb{N}_{0:T-1}$ and any deviations $(\Delta x, \Delta u)$ satisfying $\|(\Delta x, \Delta u)\|_\infty \leq \tau_k$, the dynamics remainder is bounded as*

$$|f(z_k + \Delta x,\, v_k + \Delta u) - f(z_k, v_k) \\ - A_k \Delta x - B_k \Delta u| \leq \sigma(\tau_k, z_k, v_k). \tag{20}$$

Remark 2. *The function $\sigma(\tau_k, z_k, v_k)$ can be selected to capture the linearization error of the dynamics f around the nominal trajectory, as in [25, 24], or learned directly via data [34].*

Similarly, the reduced measurement model (13) with measurement noise e_k can be expanded around the nominal trajectory. In the following, we use a linear reduced observation model

$$h^r(x) = C^r x, \qquad C^r \in \mathbb{R}^{n_r \times n_x}, \tag{21}$$

which yields the reduced measurement error model

$$\Delta y^r_{k+1} = C^r \Delta x_k + F^r(z_k) e_k + \eta(\tau_k). \tag{22}$$

where $\eta(\tau_k)$ captures error due to nonlinearities in $F^r(\cdot)$.

Remark 3. *The reduced observation model h^r in (13) may be any differentiable nonlinear function. A key strength of the linear reduced observation model (22) is that the learned perception map $\mathrm{p}(\cdot)$ absorbs the main nonlinearities of visual perception offline, so that online planning can rely on a simple linear reduced model, with the remaining mismatch captured by the calibrated state-dependent error bound in Sec. VI-C2.*

To certify the nonlinear output-feedback loop, we require that the residual reduced-observation error admits a state- and tube-dependent overbound. This is stated in Assumption 4 below and is implemented in Sec. VI-C2 via empirical calibration.

Assumption 4. *For each $k \in \mathbb{N}_{0:T-1}$ and any deviations $(\Delta x, \Delta u)$ satisfying $\|(\Delta x, \Delta u)\|_\infty \leq \tau_k$, the reduced observation mismatch in (22) is bounded by the calibrated state-dependent overbound $\Upsilon(\tau_k, z_k)$, i.e.,*

$$\left|F^r(z_k) e_k + \eta(\tau_k)\right| \leq \Upsilon(z_k, \tau_k) e_k, \tag{23}$$

where the absolute value is understood component-wise.

In Section VI-C2, we work with a reduced image y^r and a corresponding reduced observation model (22). The function F is defined to explicitly capture the approximation error introduced by this reduction, as described in (13).

We then define the closed-loop response variables $\mathbf{\Phi}$ and enforce the SLS constraints (9). In this formulation, the system matrices $\mathbf{A}(\boldsymbol{z}, \boldsymbol{v})$ and $\mathbf{B}(\boldsymbol{z}, \boldsymbol{v})$ depend on the nominal decision variables $\boldsymbol{z}$ and $\boldsymbol{v}$. Specifically, the deviation variables satisfy

$$\begin{bmatrix} \mathbf{\Delta x} \\ \mathbf{\Delta u} \end{bmatrix} = \mathbf{\Phi} \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Upsilon} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix}, \tag{24}$$

i.e., the closed-loop response of the linearized dynamics (17). Here, $\mathbf{\Sigma} = \mathrm{blkdiag}(\Xi, \Sigma_0, \ldots, \Sigma_{T-1})$ captures state-dependent disturbance due to the function $\sigma(\tau_k, z_k, v_k)$, with

$$\Sigma_j = \sigma(\tau_j, z_j, v_j) I_{n_x}, \quad j \in \mathbb{N}_{0:T-1}, \tag{25}$$

including the initial condition uncertainty Ξ. The vector $\boldsymbol{\tau} = (\rho, \tau_0, \ldots, \tau_{T-1})$ collects the tube radii, and $\mathbf{\Upsilon}$ is defined analogously to capture the observation uncertainty.

B. Nonlinear robust constraint satisfaction

Using the LTV parameterization and error overbounds, we can robustly enforce constraints on the nonlinear system (1).

Proposition 3. *Consider the output–feedback SLS responses $\mathbf{\Phi}^{\mathrm{xw}}, \mathbf{\Phi}^{\mathrm{xe}}, \mathbf{\Phi}^{\mathrm{uw}}, \mathbf{\Phi}^{\mathrm{ue}}$ satisfying (9) for the LTV error model and suppose that Assumptions 1, 3 and 4 hold. Then the closed-loop trajectories of the nonlinear system (1) satisfy the state and input constraints robustly if the following tightened inequalities hold:*

$$\textstyle\sum_{j=0}^{k} \left\|c_i^\top \mathbf{\Phi}^{\mathrm{w}}_{k,j} \Sigma_j\right\|_1 + \left\|c_i^\top \mathbf{\Phi}^{\mathrm{e}}_{k,j} \Upsilon_j\right\|_1 + c_i^\top (z_k, v_k) + b_i \leq 0. \tag{26a}$$

$$\textstyle\sum_{j=0}^{k} \left\|\hat{e}_l^\top \mathbf{\Phi}^{\mathrm{w}}_{k,j} \Sigma_j\right\|_1 + \left\|\hat{e}_l^\top \mathbf{\Phi}^{\mathrm{e}}_{k,j} \Upsilon_j\right\|_1 \leq \tau_k, \tag{26b}$$

for all $k \in \mathbb{N}_{0:T-1}$, $i = 1, \ldots, n_{\mathrm{c}}$, with $\Sigma_j = \sigma(\tau_j, z_j, v_j) I_{n_{\mathrm{x}}}$, for $l = 1, \ldots, n_{\mathrm{x}} + n_{\mathrm{u}}$ where $\hat{e}_l$ denote standard basis vectors.

Proof: The proof follows from [32, 25]; see App. C. ■

Collecting the nominal dynamics (15), the controller parameterization (9) based on the linearized dynamics (17) and observation model, and the robust constraint satisfaction in (26a)–(26b), we pose the following finite-horizon nonlinear output-feedback SLS synthesis problem in Eq. (27).

$$\min_{\substack{\mathbf{\Phi}, \boldsymbol{\tau} \\ \boldsymbol{z}, \boldsymbol{v}}} J(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\tau}, \mathbf{\Phi}) \tag{27a}$$

$$\text{s.t.} \quad z_{k+1} = f(z_k, v_k), \quad k \in \mathbb{N}_{0:T-1}, \tag{27b}$$

$$\begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A}(\boldsymbol{z}, \boldsymbol{v}) & -\mathbf{Z}\mathbf{B}(\boldsymbol{z}, \boldsymbol{v}) \end{bmatrix} \mathbf{\Phi} = [\mathbf{I} \;\; \mathbf{0}], \tag{27c}$$

$$\mathbf{\Phi} \begin{bmatrix} \mathbf{I} - \mathbf{Z}\mathbf{A}(\boldsymbol{z}, \boldsymbol{v}) \\ -\mathbf{Z}\mathbf{C}(\boldsymbol{z}, \boldsymbol{v}) \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, \tag{27d}$$

$$\sum_{j=0}^{k} \left\|c_i^\top \mathbf{\Phi}^{\mathrm{w}}_{k,j}\, \Sigma_j\right\|_1 + \left\|c_i^\top \mathbf{\Phi}^{\mathrm{e}}_{k,j}\, \Upsilon_j\right\|_1 + c_i^\top \begin{bmatrix} z_k \\ v_k \end{bmatrix} + b_i \leq 0, \\ i = 1, \ldots, n_{\mathrm{c}}, \quad k \in \mathbb{N}_{0:T-1}, \tag{27e}$$

$$\textstyle\sum_{j=0}^{k} \left\|\hat{e}_l^\top \mathbf{\Phi}^{\mathrm{w}}_{k,j}\, \Sigma_j\right\|_1 + \left\|\hat{e}_l^\top \mathbf{\Phi}^{\mathrm{e}}_{k,j}\, \Upsilon_j\right\|_1 \leq \tau_k, \\ l = 1, \ldots, n_{\mathrm{x}} + n_{\mathrm{u}}, \quad k \in \mathbb{N}_{0:T-1}. \tag{27f}$$

The formulation (27) integrates all components of the robust MPC design:

- The nominal trajectory is propagated according to the nonlinear dynamics model (27b).
- The propagation of model and measurement disturbance is captured via (27c)–(27d), ensuring that the response matrices $\mathbf{\Phi}$ correspond to the closed-loop response of the LTV system linearized around the nominal trajectory.
- Robust constraint satisfaction is enforced by the tightened constraints (27e), which guarantee that constraints (2) hold for all disturbances and modeling errors.

As per Prop. 3, the disturbance-feedback gains resulting from solving (27) ensures that the uncertain nonlinear system (27b) remains within the constraints $\mathcal{S}$ over the prediction horizon.

C. Perception function

We now describe how to build the reduced observation map in (13) from camera output via pre-trained visual features.

Our goal is to compress high-dimensional images into a low-dimensional, control-relevant observation, while fitting a calibrated overbound on the error caused by this reduction, so that robust constraint satisfaction can still be enforced via Prop. 3.

1) Learning a lower-dimensional visual representation: We seek a reduced representation that retains only the information in the image needed for control. Since directly planning from high-dimensional visual features is intractable, we instead project them into a low-dimensional observation space that is as informative as possible about the underlying state. To quantify this, we use a dynamics-aware observability metric. Based on the dynamics f and the reduced observation model h^r, we define the local discrete-time nonlinear observability matrix at $\hat{x}$ over an input sequence $\boldsymbol{u} := \{u_0, \ldots, u_n\}$ as

$$\mathcal{O}(\hat{x}, \boldsymbol{u}) \doteq \begin{bmatrix} h^r(f(\hat{x}, u_0)) \\ h^r(f(f(\hat{x}, u_0), u_1)) \\ \vdots \\ h^r(f(f(\ldots), u_{n-1}), u_n)) \end{bmatrix}, \tag{28}$$

where $\mathcal{O}$ quantifies how well x can be inferred from outputs: higher (Jacobian) rank implies better observability [17, Ch. 3], implemented here via the minimum singular value $\underline{\sigma}$.

Remark 4. *If $\nabla_x \mathcal{O}|_{x=\hat{x}}$ is full rank, the system is strongly observable, i.e., x is locally recoverable from $n+1$ outputs.*

Remark 5. *Dimensionality reduction via reconstruction (e.g., autoencoders [5]) often retains visually salient but control-irrelevant distractors [44], making overbounds loose.*

Given DINO features $\psi^{\text{dino}}(y_i)$, we learn a projection $\mathrm{p}(\cdot)$ from the feature space to a low-dimensional observation space $\mathbf{y}_r$ and a state-based reduced observation model $h^r(\cdot)$. The projection is trained so that the reduced observation $\mathrm{p}(\psi^{\text{dino}}(y_i))$ matches the state-based target $h^r(x_i)$. To make this reduced observation useful for control, we encourage it to preserve observability of the state x via the metric above: that is, over dataset $\mathbf{D}$, we train $\mathrm{p}(\cdot)$ and h^r by minimizing

$$\ell(\theta_p, \theta_h) = \| \, \mathrm{p}(\psi^{\text{dino}}(y_i); \theta_p) - h^r(x_i; \theta_h) \| - \lambda \underline{\sigma}(\mathcal{O}(\hat{x}, \boldsymbol{u})) \tag{29}$$

where θ_p are the parameters of $p(\cdot)$ and θ_h parameterizes h^r (in our implementation, the entries of C^r). The scalar $\lambda > 0$ balances feature matching against observability. The dimension of the reduced observation h^r and the horizon length n used in $\mathcal{O}$ are chosen as hyperparameters, and $\hat{x}$ and $\boldsymbol{u}$ are randomly sampled during training.

2) Calibration of the perception reduction error: After training, we fix $\mathrm{p}(\cdot)$ and $h^r(\cdot)$, and then calibrate a state-dependent overbound on the residual reduction error introduced by replacing the original high-dimensional observation with the reduced observation $y^r = \mathrm{p}(\psi^{\text{dino}}(y))$. This residual captures the mismatch between the learned reduced observation and its state-based model $h^r(x)$.

To empirically preserve robust constraint satisfaction in (13), we overbound this mismatch by a state-dependent uncertainty set. Specifically, from a dataset $\{(x_i, y_i)\}$ we form residuals

$$r_i \;=\; \left\| \, \mathrm{p}(\psi^{\text{dino}}(y_i)) - h^r(x_i) \right\|_\infty. \tag{30}$$

To define a state-dependent uncertainty set of the form $F^r(x)\mathcal{B}$ that bounds this mismatch, we empirically calibrate a scalar envelope $b(x)$ by fitting a polynomial via the convex program

$$\begin{aligned} \min_{\beta,\, \xi_i \geq 0} \quad & \sum_i \left(\beta^\top m(x_i)\right) + \gamma \|\beta\|_2^2 + \mu \sum_i \xi_i \\ \text{s.t.} \quad & \beta^\top m(x_i) \;\geq\; r_i - \xi_i \quad \forall i, \end{aligned} \tag{31}$$

where $\gamma, \mu \geq 0$ trade off smoothness and outlier violations, $m(\cdot)$ is a monomial basis, and $b(x) := \beta^\top m(x)$ is the calibrated scalar envelope. We then set $F^r(x) = b(x)\, I$ so that

$$\mathrm{p}(\psi^{\text{dino}}(y)) - h^r(x) \;\in\; F^r(x)\, \mathcal{B}_\infty. \tag{32}$$

This yields a differentiable, state-dependent uncertainty set that is cheap to evaluate in constrained trajectory optimizers.

Remark 6. *The calibrated scalar overbound is intended to certify perception error over the operating region covered by the calibration data, and is therefore most reliable near that distribution. Adapting the certificate to a similar environment only requires recalibrating the scalar bound $b(x)$, rather than retraining the full perception-control pipeline. In our experiments (Section VIII), this recalibration is data-efficient: even a small number of calibration samples already provides high empirical coverage. The learned perception map plays a different role: because it is built on DINO features that capture transferable geometric structure, it may generalize beyond the calibration environment, see, e.g, [47], while the bound is used only for safety certification. Alternatively, Lipschitz-type bounds [20, 11] and bounds given by neural network verifiers [45] can provide rigorous overbounding guarantees, but typically lead to nonsmooth or overly conservative constraint tightenings. Alternatively, distribution-free approaches such as conformal prediction [34] or extreme value theory [21] can provide statistical guarantees beyond the sampled data.*

VII. Scalable Implementation via SCP

In this section, we discuss how VISION-SLS efficiently solves (27) for high-dimensional robotics problems. We first discuss an efficient Riccati-based solver for the LTV SLS output feedback problem (11) (Sec. VII-A), which will be used as a subroutine in a sequential convex programming (SCP) solver (Sec. VII-B).

A. Efficient Solver for SLS-based Output-Feedback LQG

Although the LTV SLS output-feedback problem (11) is convex and therefore amenable to off-the-shelf general-purpose convex optimization solvers [18], the large number of decision variables and constraints causes these solvers to scale poorly with the horizon and system dimension in SLS problems [23]. This can become prohibitive for large-scale robotics problems. To overcome this challenge, we demonstrate that the unconstrained linear time-varying problem (11) admits an efficient double-Riccati structure, akin to state-feedback [23]. In Section VII, this structure is used as a computational building block inside the nonlinear SCP algorithm.

Proposition 4. *Problem (11) has an optimal solution given by T parallel backward Riccati recursions*

$$\begin{aligned} S_{T,j} &= P, \\ K_{k,j} &= -\left(R + B_k^\top S_{k+1,j} B_k\right)^{-1} \left(B_k^\top S_{k+1,j} A_k\right), \\ S_{k,j} &= Q + A_k^\top S_{k+1,j} A_k + \left(A_k^\top S_{k+1,j} B_k\right) K_{k,j}, \end{aligned} \tag{33}$$

and independently T parallel backward Kalman recursions

$$\begin{aligned} \Pi_{k,0} &= \Xi\Xi^\top, \\ L_{k,j+1} &= -\left(F_j F_j^\top + C_j \Pi_{k,j} C_j^\top\right)^{-1} C_j \Pi_{k,j} A_j^\top, \\ \Pi_{k,j+1} &= E_j E_j^\top + A_j \Pi_{k,j} A_j^\top + A_j \Pi_{k,j} C_j^\top L_{k,j+1}, \end{aligned} \tag{34}$$

where S and Π are the cost-to-go and estimation error covariance matrices, respectively, defined recursively. Given the controller gains $K_{k,j}$ and observer gains $L_{k,j}$, compute the closed-loop propagation operators in parallel via the two T forward passes

$$\begin{aligned} &\bar{\mathbf{\Phi}}^{\mathrm{x}}_{j,j} = I, \\ &\bar{\mathbf{\Phi}}^{\mathrm{u}}_{k,j} = K_{k,j} \bar{\mathbf{\Phi}}^{\mathrm{x}}_{k,j}, \quad \bar{\mathbf{\Phi}}^{\mathrm{x}}_{k+1,j} = (A_k + B_k K_{k,j}) \bar{\mathbf{\Phi}}^{\mathrm{x}}_{k,j}, \end{aligned} \tag{35}$$

and

$$\begin{aligned} &\hat{\mathbf{\Phi}}^{\mathrm{x}}_{k,k} = I, \\ &\hat{\mathbf{\Phi}}^{\mathrm{y}}_{k,j} = \hat{\mathbf{\Phi}}^{\mathrm{x}}_{k,j} L_{k,j}^\top, \quad \hat{\mathbf{\Phi}}^{\mathrm{x}}_{k,j} = \hat{\mathbf{\Phi}}^{\mathrm{x}}_{k,j+1} A_j + \hat{\mathbf{\Phi}}^{\mathrm{y}}_{k,j+1} C_j, \end{aligned} \tag{36}$$

for $j = 0, \ldots, T,\ k = j, \ldots, T-1$ and $k = 0, \ldots, T,\ j = 0, \ldots, k-1$ respectively.

Finally, assemble the matrices using $\mathbf{M} = \mathbf{I} - \mathbf{Z}\mathbf{A}$,

$$\mathbf{\Phi}^{\mathrm{xw}} = \bar{\mathbf{\Phi}}^{\mathrm{x}} + \hat{\mathbf{\Phi}}^{\mathrm{x}} - \bar{\mathbf{\Phi}}^{\mathrm{x}} \mathbf{M} \hat{\mathbf{\Phi}}^{\mathrm{x}}, \quad \mathbf{\Phi}^{\mathrm{uw}} = \bar{\mathbf{\Phi}}^{\mathrm{u}} (I - \mathbf{M} \hat{\mathbf{\Phi}}^{\mathrm{x}}), \tag{37a}$$

$$\mathbf{\Phi}^{\mathrm{xe}} = \hat{\mathbf{\Phi}}^{\mathrm{y}} - \bar{\mathbf{\Phi}}^{\mathrm{x}} \mathbf{M} \hat{\mathbf{\Phi}}^{\mathrm{y}}, \quad \mathbf{\Phi}^{\mathrm{ue}} = -\bar{\mathbf{\Phi}}^{\mathrm{u}} \mathbf{M} \hat{\mathbf{\Phi}}^{\mathrm{y}}. \tag{37b}$$

Proof: See Appendix B. ∎

In the nonlinear constrained setting, we use this double-Riccati structure as an efficient computational primitive inside SCP; we discuss the resulting nonlinear algorithm in Sec. VII-B.

B. Implementation via SCP

To solve (27), we adopt an SCP approach [38] that, at each iteration, constructs a convex second-order cone program (SOCP) by linearizing the dynamics and convexifying the cost and constraint tightening around the current nominal trajectory $(\boldsymbol{z}, \boldsymbol{v})$. Each subproblem is then solved using a structure-exploiting SLS method based on [23], together with the double-Riccati construction of Proposition 4 to construct feasible output-feedback responses Φ and their reachable sets. In practice, this amounts to combining the control-side state-feedback SLS solution with a time-varying Kalman recursion for state estimation. Because the extra estimation step only requires a standard Riccati recursion, the per-iteration complexity remains close to that of a single state-feedback fast-SLS MPC step. The runtime of this SCP procedure can be further reduced via GPU parallelization [16]. The resulting scheme is efficient and maintains feasibility, though we note the coupling does not yield optimality guarantees: unlike the state-feedback fast-SLS solver [23], the dual-based cost update used within the unconstrained Riccati recursions (11) is not based on the KKT structure of the constrained problem (analogous to [23, Eq. 21]), and is hence a heuristic. Future work will aim to uncover the banded numerical structure of constrained SLS-based LQG. The full procedure is in Alg. 1.

Algorithm 1: SCP-based Output-Feedback SLS

Repeat until convergence (SCP loop):

1) **Linearize** dynamics/constraints around $(\boldsymbol{z}, \boldsymbol{v})$; quadratize cost to obtain convex SOCP subproblem.
2) **Solve linear approximation:**
   a) solve QP: update $(\Delta\boldsymbol{z}, \Delta\boldsymbol{v})$,
   b) compute $\bar{\mathbf{\Phi}}^{\mathrm{x}}, \bar{\mathbf{\Phi}}^{\mathrm{u}}$ via fast-SLS [23] (parallel Riccati recursions).
   c) compute $\hat{\mathbf{\Phi}}^{\mathrm{x}}, \hat{\mathbf{\Phi}}^{\mathrm{y}}$ (parallel Kalman recursions).
   d) compute $\mathbf{\Phi}^{\mathrm{xw}}$, $\mathbf{\Phi}^{\mathrm{uw}}$, $\mathbf{\Phi}^{\mathrm{xe}}$, $\mathbf{\Phi}^{\mathrm{ue}}$ via (37)
   e) compute constraint tightening (26a)–(26b)
   f) set $(\boldsymbol{z}, \boldsymbol{v}) \leftarrow (\boldsymbol{z}, \boldsymbol{v}) + (\Delta\boldsymbol{z}, \Delta\boldsymbol{v})$

VIII. Results

We evaluate VISION-SLS, first comparing against a recent safe perception-based control algorithm [42] and then evaluating on five systems with uncertain dynamics and uncertain partial observations: the classic light-dark benchmark [19], a 4D car with overhead RGB images, a 10D quadrotor with onboard RGB images, a humanoid with partial state sensing, and a real-world TurtleBot with onboard RGB images. We demonstrate (i) robust constraint satisfaction via tube tightening, (ii) information-gathering motion emerging from the SLS design even for large state dimensions and RGB observations. For the car and quadrotor, we use DINOv2-vitg14 and DINOv2-vitl14 [28] as the pre-trained VFM, leveraging 512x512 RGB images rendered in PyBullet, respectively with 3×10^5 and 2×10^6 datapoints to train $\mathrm{p}(\cdot)$ and C^r, and fit a polynomial error bound with (31). To implement Alg. 1, we use CasADi [3], using IPOPT [39] to find an initial guess for $(\mathbf{z}, \mathbf{v})$ (Step 1 of Alg. 1) and OSQP [35] to solve the QP in Step 2a of Alg. 1. Alg. 1 terminates if $\|(\Delta z, \Delta v)\|_2 \leq 0.001$ or after 20 iterations.

We baseline Alg. 1 against 1) a certainty equivalent (CE) controller [13, 36], which assumes the separation principle, akin to iLQG, by optimizing a reference and controller assuming perfect observations, i.e., with post-hoc tubes, which corresponds to the first iteration of Alg. 1 2) a non-robust (NR) belief-space controller inspired by [19], which performs information gathering but does not perform constraint tightening to robustly ensure safety under uncertainty, and 3-4) two recent safe perception-based control methods [37, 42]. We define success as reaching the terminal set without collision and report success rate (SR) and constraint violation rate (CVR) of our method and baselines over 15 random initial states each with 50 rollouts under random disturbances in Table I.

Baseline [42]. We compare VISION-SLS with a recent safe perception-based control method [42] in their F1/10th driving setting [42, Sec. 5], which assumes a 3D vehicle state $[p_x, p_y, \theta] \in \mathbb{R}^3$ with translation (p_x, p_y) and orientation θ. We use the code of [42] to generate a dataset to train our perception module and error bound. Notably, [42] leverages a 64-

	Light-dark			4D Car		
	Ours	CE	NR	Ours	CE	NR
SR	**100%**	90.13%	87.73%	**98.53%**	93.47%	48.13%
CVR	**0%**	9.87%	12.27%	**1.6%**	6.53%	55.87%
	Quadrotor			Humanoid		
	Ours	CE	NR	Ours	CE	NR
SR	**100%**	100%	29.47%	**100%**	100%	2%
CVR	**0%**	0%	92.8%	**0%**	0%	98%

TABLE I

COMPARING VISION-SLS, CERTAINTY EQUIVALENT (CE), NON-ROBUST (NR) IN SUCCESS RATE (SR), CONSTRAINT VIOLATION RATE (CVR).

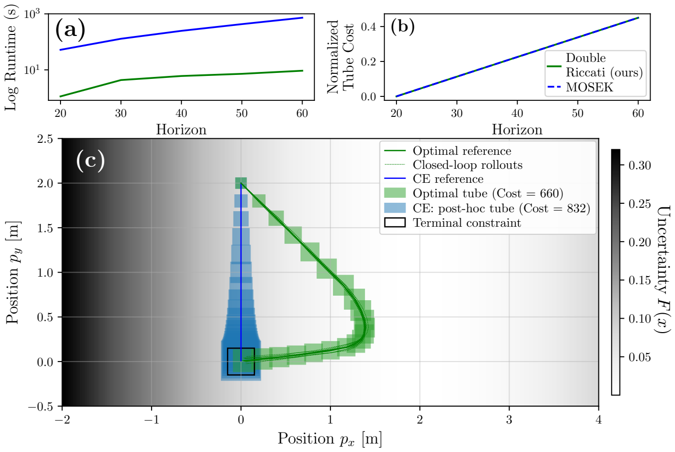

Fig. 3. On the light-dark domain, our method visits the low perception-error region, where lighter shades indicate lower error (c), reducing final state uncertainty. The naïve CE controller incurs more uncertainty over the horizon. Our method is much faster to solve than a baseline convex optimization-based solver [2, 48] that uses MOSEK [4] (a) and yields the same optimal cost (b).

dimensional LiDAR observation instead of pixel observations; our method can flexibly be applied to this setting by learning a projection function $p : \mathbb{R}^{64} \to \mathbb{R}^3$, setting $h^r(x) = x$, and calibrating the error bound $F^r(x_k)$ using (31). VISION-SLS achieves 100% safety, higher than the 93% reported by [42], as it jointly computes a trajectory and robust controller with hard safety guarantees given a valid overbound F^r.

Light-Dark. We first show the information-gathering abilities, runtime scaling with horizon length, and optimality of Prop. 4 in the light-dark domain [19], a benchmark for partially-observed control. We consider a 2D single integrator with state $x = (p_x, p_y)$, with state, input, and terminal state constraints (see App. D2). Alg. 1 converges after 17 SCP iterations in 1.171 seconds. By penalizing the tube volume in (27a), our method steers the system to lower perception-error regions (Fig. 3c), which reduces tube growth and enables information-gathering while satisfying the terminal constraints. In contrast, the CE baseline experiences larger perception error, causing larger tubes and terminal constraint violation. All 10 simulated closed-loop rollouts stay in their computed reachable sets, as expected. Moreover, we show that our method scales much better than a naïve SLS output feedback controller synthesized via convex optimization [2, 48] that does not use the double Riccati solver (Fig. 3a) while achieving the same cost (Fig. 3b), empirically validating the optimality result of Prop. 4. Our method always reaches the terminal set while CE fails in 9.87% and NR fails in 12.27% of all trials (see Table I).

4D Car with Overhead RGB Images. In the vision-based driving task in Fig. 4, we steer a 4D car with state (p_x, p_y, p_θ, v) (position, rotation, linear velocity) [11] to a goal state over a horizon $T = 30$ with a timestep of 0.15s, enforcing state and input constraints, avoiding collisions with other cars modeled as circular obstacles, while minimizing the propagated uncertainty along the trajectory. More details are in App. D3. To clearly show information-gathering for vision-based control, a heavy visual occlusion (0.97 opacity) is placed in the middle of the environment, increasing the worst-case measurement error for those states. In Fig. 4, we visualize the learned state-dependent polynomial overbound $F^r(x)$ as a grayscale heatmap over the (p_x, p_y) workspace. The calibration of the perception error bounds are given in App. D5. The coverage of error bound over-approximation (tested at 500 randomly sampled states) vs. number of calibration points, shown in Table II, demonstrates the data-efficiency of our perception calibration procedure. Notably, our method exceeds 90% empirical coverage with only 50 samples. The optimal plan produced by Alg. 1 (computed in 3.584s over 20 SCP iterations) is shown in green in Fig. 4. The resulting trajectory avoids the occluded region to actively reduce perception uncertainty, whereas the CE baseline drives through the occlusion and incurs larger tube widths (Fig. 4d). Closed-loop rollouts using DINO-v2 features remain in the computed reachable tubes and satisfy all constraints despite heavy occlusion (Fig. 4c). In contrast, the non-robust baseline violates safety under disturbance, highlighting the necessity of constraint tightening for vision-based control.

In this setting, we also compare our method against a variant of [37] in which we replace the NeRF model with a convolutional neural network trained on our perception dataset to predict the minimum distance to obstacles (to better match our constrained single-camera-view setting). The SR and CVR of the variant of [37] are 65.47% and 27.93% respectively, indicating our method outperforms all baselines with 98.53% SR and 1.6% CVR. The variant of [37] performs poorly here because its formulation makes it difficult to sample safe actions for nonlinear dynamics. Moreover, [37] does not formally account for robust constraint satisfaction under disturbances, and thus cannot counteract the error of the distance predictor.

calibration points	50	100	250	500	1500
empirical coverage	91.4%	92.2%	97.4%	99%	99%

TABLE II

ERROR BOUND $F^r(x)$ COVERAGE VS. NUMBER OF CALIBRATION POINTS.

10D Quadrotor with Onboard RGB Images. We evaluate on a vision-based flying task (Fig. 5), steering a 10D quadrotor model [11] in a cluttered 3D environment to a goal under partial observability (see App. D4). The resulting prediction tubes robustly avoid obstacles despite perception uncertainty, whereas both the certainty-equivalent controller and the non-robust information-gathering baseline collide with obstacles, even though their nominal trajectories are collision-free. In addition, tight state constraints (e.g., pitch) are respected throughout the horizon due to robust constraint tightening. Our method succeeds and never violates constraints in all trials. The CE baseline also succeeds in all trials, but produces larger tubes because it cannot perform information gathering. In contrast, the NR baseline succeeds only in 29.47% trials (see Table

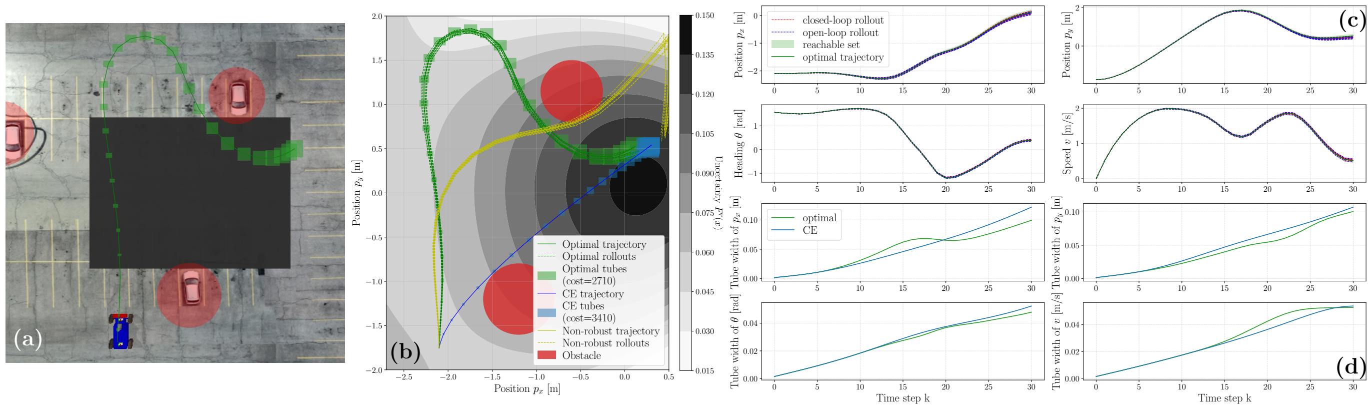

Fig. 4. (a) Car in an occluded parking lot with a top-down camera avoiding red obstacles. (b) Grayscale shows the learned uncertainty for the reduced observation (darker = higher uncertainty/occlusion). Our controller tightens constraints using the propagated set and yields a collision-free trajectory. In contrast, the certainty-equivalent trajectory stays longer in the high-uncertainty occluded region. (c) Closed-loop rollouts in the occluded scene, overlaying the nominal plan, the SLS-induced uncertainty set (for constraint tightenings), and sample closed-loop rollouts. (d) Comparing reachable tube size per state dimension. We minimize a weighted tube-volume cost, with a stronger penalty on the terminal tube, encouraging information gathering to reduce final-time uncertainty.

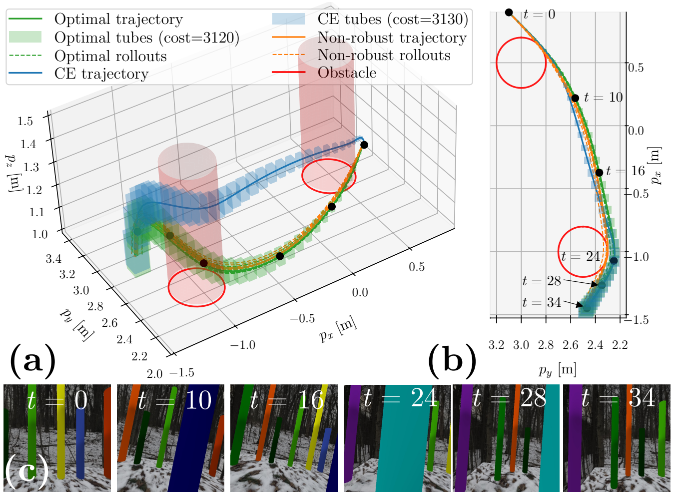

Fig. 5. Controlling a quadcopter with an onboard camera, avoiding obstacles despite perception error and dynamics uncertainty. (a) Our closed-loop rollouts stay within the computed tubes, reaching the target with smaller uncertainty bounds compared to the CE baseline, maintaining safety while a non-robust information-gathering policy does not. (b) A top-down view of (a). (c) Snapshots of the images collected onboard to implement feedback control.

I). Overall, these results show robust constraint satisfaction for vision-based control tasks on higher-dimensional dynamics.

Humanoid with Partial State Measurement. We demonstrate applicability to high-dimensional systems with partial state measurements on a Unitree G1 humanoid [22] with 59 states and 23 inputs modeled in Pinocchio [8]. We consider a vision-free setting to isolate scalability with respect to the state dimension from scalability with respect to the observation dimension, since the former primarily impacts online control synthesis whereas the latter is handled offline by the perception reduction module (see Remark 1). Hence, only joint angles and base pose are observed (no vision), and state and input constraints must be satisfied (see App. D6 for details and all reachable tubes). The task is robust tracking of a backflip trajectory under disturbance (Fig. 1a). In Fig. 1b, open-loop rollouts (red) exit the reachable tubes and violate constraints (e.g., quaternion limits), whereas all closed-loop rollouts (blue) remain within the tubes and satisfy constraints, demonstrating effective output-feedback optimization and joint reachability–trajectory optimization for partially observed large-scale dynamics. Our method always steers the humanoid to the terminal set without constraint violation despite the uncertainties and partial observation. In this example, the CE baseline also yields the same solution because the perception uncertainty is uniformly bounded over the space. However, the NR baseline performs far worse than VISION-SLS, achieving only 2% SR and 98% CVR (Table I), because even small disturbances can be amplified by the complex nonlinear humanoid dynamics.

Hardware Experiment. We evaluate our method on a physical TurtleBot 4 platform operating in an office environment (Fig. 1d). The robot is tasked with reaching a goal position under state and input constraints, including collision avoidance with furniture. See App. D7 for detailed experiment details. Using onboard visual feedback, the closed-loop controller robustly tracks the planned trajectory across 3 real-world runs (Fig. 1c). Fig. 1e shows example onboard images used during execution. All executions stay within the computed reachable tubes and satisfy collision-avoidance constraints, demonstrating safe vision-based output-feedback on hardware.

IX. Conclusion

We develop a tractable framework for safe nonlinear output-feedback planning and control from pixels. The key ingredients are (i) a learned low-dimensional observation map that turns high-dimensional images into reduced measurements with calibrated error bounds, and (ii) an SLS-based parameterization of dynamic output-feedback as causal affine policies around a nominal trajectory with robust constraint satisfaction. After training the perception map and error bounds from data, the full nonlinear output-feedback synthesis problem is solved efficiently via SCP and a Riccati-based SLS solver. Empirically, the method enables information-gathering behavior and robust constraint satisfaction on high-dimensional, underactuated systems with realistic visual observations. Overall, the results suggest that pairing visual representation learning with SLS is a practical path to scalable safe visuomotor control. Future work includes total occlusion handling, learned dynamics, and an optimal solver for the constrained version of (11).

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *CoRR*, abs/2112.05814, 2021.

[2] James Anderson, John C. Doyle, Steven H. Low, and Nikolai Matni. System level synthesis. *Annu. Rev. Control.*, 47:364–393, 2019.

[3] Joel AE Andersson, Joris Gillis, Greg Horn, James B Rawlings, and Moritz Diehl. Casadi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36, 2019.

[4] MOSEK ApS. *The MOSEK optimization toolbox for Python manual. Version 11.1.*, 2019. URL http://docs.mosek.com/9.0/toolbox/index.html.

[5] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors, *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pages 37–50. JMLR.org, 2012.

[6] Ershad Banijamali, Rui Shu, Mohammad Ghavamzadeh, Hung Bui, and Ali Ghodsi. Robust locally-linear controllable embedding. In *AISTATS*, 2018.

[7] Dimitri P Bertsekas. Dynamic programming and suboptimal control: A survey from adp to mpc. *European journal of control*, 11(4-5):310–334, 2005.

[8] Justin Carpentier, Guilhem Saurel, Gabriele Buondonno, Joseph Mirabel, Florent Lamiraux, Olivier Stasse, and Nicolas Mansard. The pinocchio c++ library: A fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 614–619. IEEE, 2019.

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[10] Glen Chou and Russ Tedrake. Synthesizing stable reduced-order visuomotor policies for nonlinear systems via sums-of-squares optimization. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 624–631. IEEE, 2023.

[11] Glen Chou, Necmiye Ozay, and Dmitry Berenson. Model error propagation via learned contraction metrics for safe feedback motion planning of unknown systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3576–3583. IEEE, 2021.

[12] Glen Chou, Necmiye Ozay, and Dmitry Berenson. Safe output feedback motion planning from images via learned perception modules and contraction theory. In *International Workshop on the Algorithmic Foundations of Robotics*, pages 349–367. Springer, 2022.

[13] Sarah Dean and Benjamin Recht. Certainty equivalent perception-based control. In *Learning for Dynamics and Control*, pages 399–411. PMLR, 2021.

[14] Sarah Dean, Andrew Taylor, Ryan Cosner, Benjamin Recht, and Aaron Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. In *Conference on Robot Learning*, pages 654–670. PMLR, 2021.

[15] Sampada Deglurkar, Michael H. Lim, Johnathan Tucker, Zachary N. Sunberg, Aleksandra Faust, and Claire J. Tomlin. Compositional learning-based planning for vision pomdps. In *Learning for Dynamics and Control Conference, L4DC 2023*, volume 211 of *Proceedings of Machine Learning Research*, pages 469–482. PMLR, 2023.

[16] Jeffrey Fang and Glen Chou. Safe large-scale robust nonlinear mpc in milliseconds via reachability-constrained system level synthesis on the gpu. *arXiv preprint arXiv:2604.07644*, 2026.

[17] Walter D. Grossman. Observers for discrete-time nonlinear systems. In *PhD thesis, New Jersey Institute of Technology*, 1999.

[18] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2026. URL https://www.gurobi.com.

[19] Robert Platt Jr., Russ Tedrake, Leslie Kaelbling, and Tomás Lozano-Pérez. Belief space planning assuming maximum likelihood observations. In *R:SS*, 2010.

[20] Craig Knuth, Glen Chou, Necmiye Ozay, and Dmitry Berenson. Planning with learned dynamics: Probabilistic guarantees on safety and reachability via lipschitz constants. *IEEE Robotics and Automation Letters*, 6(3): 5129–5136, 2021.

[21] Craig Knuth, Glen Chou, Jamie Reese, and Joseph Moore. Statistical safety and robustness guarantees for feedback motion planning of unknown underactuated stochastic systems. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12700–12706, 2023.

[22] Scott Kuindersma, Robin Deits, Maurice Fallon, Andrés Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous robots*, 40(3):429–455, 2016.

[23] Antoine P. Leeman, Johannes Köhler, Florian Messerer, Amon Lahr, Moritz Diehl, and Melanie N. Zeilinger. Fast system level synthesis: Robust model predictive control using Riccati recursions. *IFAC-PapersOnLine (Conference on Nonlinear MPC)*, 58(18):173–180, 2024. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2024.09.027.

[24] Antoine P Leeman, Johannes Köhler, and Melanie N Zeilinger. Guaranteed robust nonlinear mpc via disturbance feedback. *arXiv preprint arXiv:2509.18760*, 2025.

[25] Antoine P. Leeman, Johannes Köhler, Andrea Zanelli, Samir Bennani, and Melanie N. Zeilinger. Robust non-

linear optimal control via system level synthesis. *IEEE Transactions on Automatic Control*, 70(7):4780–4787, 2025. doi: 10.1109/TAC.2025.3552482.

[26] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[27] Florian Messerer, Katrin Baumgärtner, and Moritz Diehl. A dual-control effect preserving formulation for nonlinear output-feedback stochastic model predictive control with constraints. *IEEE Control Systems Letters*, 7:1171–1176, 2022.

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. doi: 10.48550/arXiv.2304.07193.

[29] Sadra Sadraddini and Russ Tedrake. Robust output feedback control with guaranteed constraint satisfaction. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, pages 1–10, 2020.

[30] Michel Schubiger, Goran Banjac, and John Lygeros. Gpu acceleration of admm for large-scale quadratic programming. *Journal of Parallel and Distributed Computing*, 144:55–67, 2020.

[31] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Ving: Learning open-world navigation with visual goals. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 13215–13222. IEEE, 2021.

[32] Jerome Sieber, Samir Bennani, and Melanie N Zeilinger. A system level approach to tube-based model predictive control. *IEEE Control Systems Letters*, 6:776–781, 2021.

[33] Sumeet Singh, Mo Chen, Sylvia L Herbert, Claire J Tomlin, and Marco Pavone. Robust tracking with model mismatch for fast and safe planning: an sos optimization approach. In *International Workshop on the Algorithmic Foundations of Robotics*, pages 545–564. Springer, 2018.

[34] Anutam Srinivasan, Antoine Leeman, and Glen Chou. Safety beyond the training data: Robust out-of-distribution mpc via conformalized system level synthesis. *arXiv preprint arXiv:2602.12047*, 2026.

[35] Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. Osqp: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[36] Emanuel Todorov and Weiwei Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

[37] Mukun Tong, Charles Dawson, and Chuchu Fan. Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields, 2023. URL https://arxiv.org/abs/2209.12266.

[38] Robin Verschueren, Gianluca Frison, Dimitris Kouzoupis, Jonathan Frey, Niels van Duijkeren, Andrea Zanelli, Branimir Novoselnik, Thivaharan Albin, Rien Quirynen, and Moritz Diehl. acados—a modular open-source framework for fast embedded optimal control. *Mathematical Programming Computation*, 14 (1):147–183, 2022.

[39] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

[40] Yuh-Shyang Wang and Nikolai Matni. Localized lqg optimal control for large-scale systems. In *2016 American Control Conference (ACC)*, pages 1954–1961. IEEE, 2016.

[41] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin A. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Neural Information Processing Systems*, pages 2746–2754, 2015.

[42] Shuo Yang, George J. Pappas, Rahul Mangharam, and Lars Lindemann. Safe perception-based control under stochastic sensor uncertainty using conformal prediction, 2023. URL https://arxiv.org/abs/2304.00194.

[43] Shuyu Zhan, Chih-Yuan Chiu, Antoine P Leeman, and Glen Chou. Robustly constrained dynamic games for uncertain nonlinear dynamics. *arXiv preprint arXiv:2509.16826*, 2025.

[44] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *9th International Conference on Learning Representations, ICLR*, 2021.

[45] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Neural Information Processing Systems*, 31, 2018.

[46] Marvin Zhang, Sharad Vikram, Laura M. Smith, Pieter Abbeel, Matthew J. Johnson, and Sergey Levine. SOLAR: deep structured representations for model-based reinforcement learning. In *ICML*, 2019.

[47] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[48] Hongyu Zhou and Vasileios Tzoumas. Safe control of partially-observed linear time-varying systems with minimal worst-case dynamic regret, 2023.