

VisualThink-VLA: Visual Intermediate Reasoning for Effective and Low-Latency Vision-Language-Action Policies

Mingjian Gao¹, Wenqiao Zhang^{1*}, Yuqian Yuan¹, Yang Dai¹, Binhe Yu¹, Zheqi Lv², Haoyu Zheng¹, Jiaqi Zhu³, Zhiqi Ge¹, Zixuan Wan⁴, Siliang Tang¹, Yueting Zhuang¹

¹Zhejiang University ²Cornell University

³National University of Singapore ⁴Xi'an University of Electronic Science and Technology

Abstract

Recent work has begun to equip vision-language-action (VLA) policies with explicit intermediate reasoning. In embodied control, however, textual chain-of-thought is a poor fit: irrelevant or weakly textual information can interfere with action prediction, while autoregressive text decoding adds too much latency for real-time closed-loop execution. We present **VISUALTHINK-VLA**, a visual intermediate-reasoning framework for accurate, low-latency VLA policies. We define this term operationally as constructing typed visual evidence, selecting a task-dependent budgeted route, and composing the selected channels into compact states before action decoding. Our bootstrapping philosophy is to *guide action with effective visual thinking*: VISUALTHINK-VLA bootstraps action prediction through a compact visual-evidence interface that preserves spatial precision and supports low-overhead decoding. Besides, to further improve performance and efficiency, VISUALTHINK-VLA adopts a tailored selective routing mechanism to learn the visual evidence tokens, enabling low-latency inference while preserving high-capacity specialization. We also introduce **VisualEvidence-Kit**, a supervision-and-audit resource centered on a **VisualEvidence-Agent** that constructs 754.7k VLA instructions in the **VisualEvidence-Set** for route supervision and counterfactual faithfulness tests. Across multiple benchmarks and real-robot evaluation, VISUALTHINK-VLA achieves the highest success rate on most benchmarks while reducing the **multi-second** latency of reasoning-augmented baselines to the **sub-second** regime. On BridgeData V2, policy-side latency falls from 8.377 s with ECoT to 0.367 s (**22.8×**).

Code: <https://github.com/DCDmllm/VisualThink-VLA>

*Corresponding author.

1 Introduction

Vision-language-action (VLA) policies map visual observations and language instructions directly to robot actions, enabling general-purpose language-conditioned control across diverse embodiments and tasks (Brohan et al., 2023; Kim et al., 2024; Team et al., 2024; Black et al., 2024). Yet direct action prediction can be brittle in manipulation scenarios that require resolving distractors, grounding spatial relations, tracking motion, or maintaining progress over long horizons. To address these challenges, recent work augments VLA policies with intermediate thinking before action prediction, improving robustness under spatial ambiguity and task-level constraints (Zawalski et al., 2024; Duan et al., 2025; Zhao et al., 2025; Zheng et al., 2025; Qu et al., 2025; Chen et al., 2025; Yin et al., 2025).

Despite their promise, existing reasoning-augmented VLA policies expose an **accuracy-efficiency trade-off**. Textual chain-of-thought provides an explicit reasoning trace, but its weak visual grounding can distract action prediction and autoregressive decoding incurs substantial latency for closed-loop control (Wei et al., 2022; Kojima et al., 2022; Jacovi and Goldberg, 2020). Dense visual or spatial side information stays closer to the scene, but may also introduce interference: redundant channels can overwhelm the action decoder, irrelevant cues can compete with task-critical evidence, and noisy auxiliary perception can propagate conflicting signals. Selective use of reasoning context is therefore important for real-time control.

This motivates our central goal: identifying a *minimal yet effective visual reasoning interface* for VLA policies. We define *visual intermediate reasoning* operationally as three coupled operations before acting: constructing typed evidence from observations and the instruction, inferring a task-adaptive budgeted route, and composing the se-

lected channels into compact states for a frozen action decoder. The resulting intermediate variable is an explicit action-conditioned evidence bottleneck. It is behaviorally testable: relation and motion routes should respond to spatial demands and temporal change, routes should vary by manipulation stage, and removing or perturbing selected channels should affect decisions. “Minimal” denotes compact, selectively activated evidence states; “effective” denotes preservation of action-relevant visual structure. This operational view is consistent with visual-planning results showing that image-based intermediate computation can suit spatial decisions (Xu et al., 2026).

To turn this motivation into a control interface, we propose **VISUALTHINK-VLA** (see figure 1), a sparse visual intermediate-reasoning framework that enables a frozen VLA backbone to *think with images* through routed visual evidence. Its system-level novelty is the coupling of a semantically typed evidence bottleneck, budget-aware dense-to-sparse learning, and an auditable plug-in interface for frozen action decoders. **VISUALTHINK-VLA** first constructs a compact six-channel evidence layer, then uses a task-adaptive orchestration layer to select useful channels at each decision step. The selected evidence is injected as learned soft states before action decoding, allowing the policy to remain visually grounded while filtering unnecessary or distracting cues and supporting sub-second inference.

To preserve the benefit of richer evidence under sparse inference, **VISUALTHINK-VLA** couples task-adaptive routing with dense-to-sparse optimization. A task-adaptive router selects which channels to consult, and a routed adapter conditions the frozen backbone on the selected cues, following the broader principle of conditional computation (Shazeer et al., 2017; Fedus et al., 2022). To stabilize sparse routing, we train the adapter with soft-hard collaborative masks and distillation from **FULLSOFT**, a dense teacher that consults all four screened effective evidence channels (Hinton et al., 2015). The four-channel teacher is obtained after channel screening removes two low-utility channels, which were found to contribute little utility while increasing latency and interference risk. The sparse policy therefore inherits much of the dense teacher’s benefit while reducing active evidence and exposure to irrelevant cues.

Because routed evidence should be inspectable, we further introduce **VisualEvidence-Kit** for su-

pervision and audit. At its core, a **VisualEvidence-Agent** extracts channel-level evidence, constructs channel-grounded traces, and incorporates human review to build **VisualEvidence-Set**. The resulting **VisualEvidence-Set** contains 754.7k visual-thinking VLA instructions, forming a route-grounded collection that covers task stages, action primitives, evidence dependence, difficulty, and channel-level targets. This resource supports route supervision, trace supervision, and counterfactual faithfulness checks through structured channel-grounded records (DeYoung et al., 2020; Jacovi and Goldberg, 2020).

Finally, we evaluate whether this interface improves control while keeping reasoning lightweight. Experiments across real-world manipulation, simulation, long-horizon tasks, and closed-loop robot execution demonstrate that **VISUALTHINK-VLA** improves the success-latency trade-off over reasoning-augmented VLA baselines. It brings textual-reasoning latency back to the sub-second control regime, preserves nearly the performance of dense visual evidence while selecting only a small number of channels, and yields routed evidence that is stage-sensitive and behaviorally aligned with policy decisions.

The main contributions are threefold:

- **A visual intermediate-reasoning interface for frozen VLAs.** We formulate *think with images* as action decoding conditioned on compact routed visual cues with selective evidence exposure.
- **A task-adaptive evidence orchestration mechanism.** **VISUALTHINK-VLA** combines channel screening, sparse evidence routing, soft-hard collaborative optimization, and teacher-student distillation to improve the success-latency trade-off while filtering irrelevant channels.
- **A route-grounded supervision and audit framework.** **VisualEvidence-Kit** centers on **VisualEvidence-Agent**, which performs evidence extraction, trace construction, and human review to build **VisualEvidence-Set** for faithfulness diagnostics.

2 Related Work

VLA Reasoning Policies. Large-scale robot policies increasingly use language-conditioned sequence models trained on heterogeneous demon-

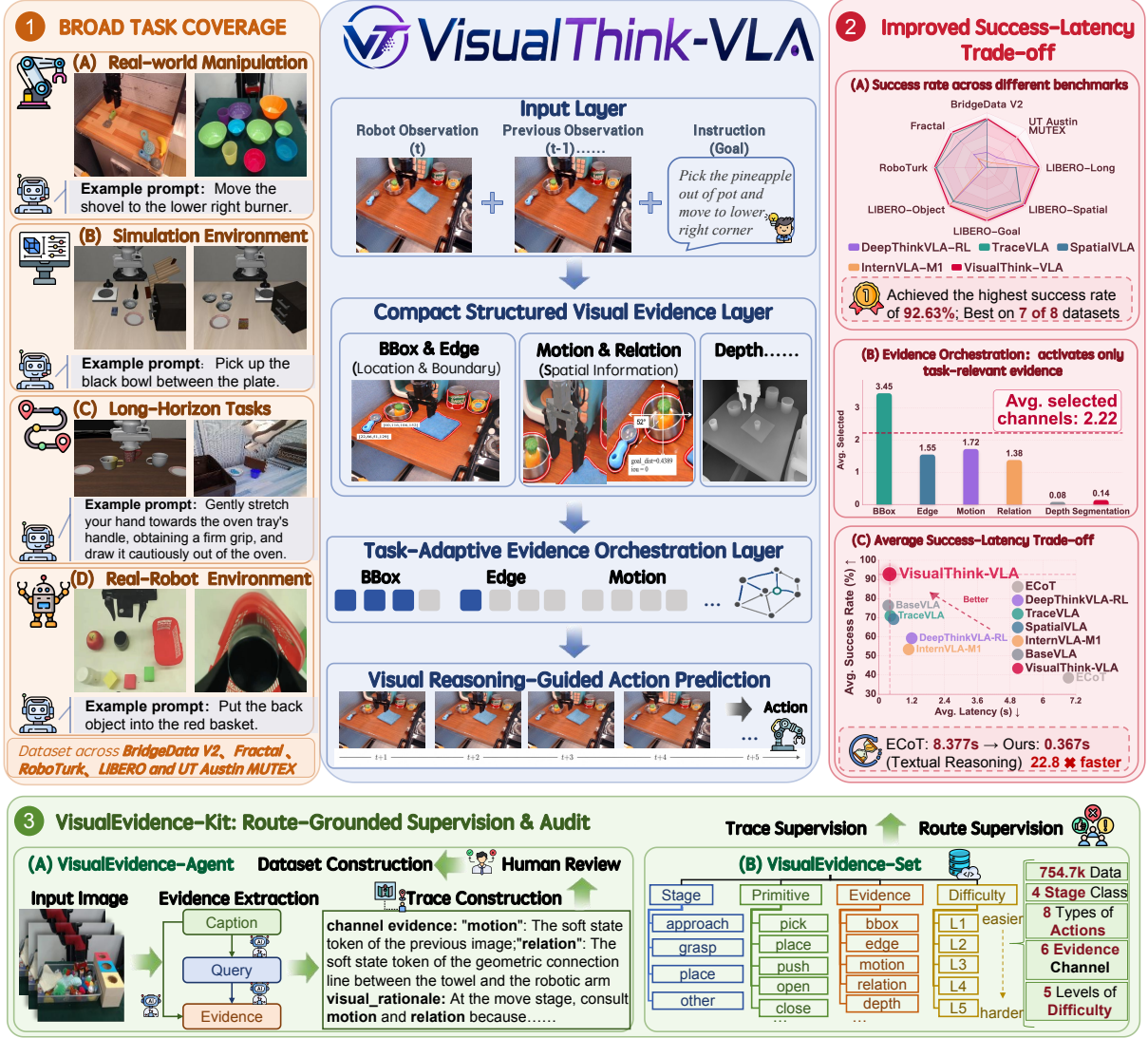


Figure 1: Overview of VISUALTHINK-VLA, covering its task coverage, the success-latency trade-off against prior VLA reasoning baselines, and the route-grounded supervision asset built with VisualEvidence-Kit.

stration corpora. Representative VLA systems include OPENVLA, Octo, GR-2, and π_0 (Kim et al., 2024; Team et al., 2024; Cheang et al., 2024; Black et al., 2024), building on broader language-conditioned robot learning over corpora such as Open X-Embodiment and RT-X (ONEILL et al., 2024; Ahn et al., 2022; Driess et al., 2023; Brohan et al., 2023). Recent reasoning-augmented VLA methods add explicit intermediate reasoning or stronger grounding, including ECoT for textual reasoning traces and TraceVLA or SpatialVLA for visual or spatial grounding (Zawalski et al., 2024; Zheng et al., 2025; Qu et al., 2025). These methods show that reasoning can help control, but they also motivate our focus on a minimal visual interface: textual CoT introduces closed-loop latency, while dense visual or spatial side information can

introduce irrelevant or redundant evidence.

Visual Evidence and Routing. Object-centric and geometry-aware representations have long been useful for robot learning and visuomotor control (Shridhar et al., 2023; Zhu et al., 2023; Chi et al., 2025; Belkhale et al., 2023). Recent multimodal and visual-planning work extends this view to grounding, segmentation, editing, and image-space planning (Ravi et al., 2025; Yuan et al., 2026; Zhong et al., 2026; Xu et al., 2026). Visual grounding primarily localizes referred entities; dense feature augmentation exposes auxiliary information to the decoder; generic conditional routing selects experts or computation paths; and textual CoT serially generates linguistic traces. VISUALTHINK-VLA predicts an explicit budget over semantically typed visual evidence and in-

Method	Location	Boundary / Geometry	Motion History	Relation	Depth	Segmentation
OPENVLA (Kim et al., 2024)	✗	✗	✗	✗	✗	✗
ECOT (Zawalski et al., 2024)	✓	✗	✗	✗	✗	✗
TraceVLA (Zheng et al., 2025)	✗	✗	✓	✗	✗	✗
SpatialVLA (Qu et al., 2025)	✗	✓	✗	✗	✓	✗
VISUALTHINK-VLA	✓	✓	✓	✓	✓	✓

Table 1: Explicit visual cues used by representative VLA methods. Depth and segmentation are included in the initial candidate bank but screened out from the default operational setting.

jects the selected non-autoregressive states into a frozen action pathway. Standard perception backends instantiate the evidence channels (Liu et al., 2024; Ravi et al., 2025; Bai et al., 2025; Radford et al., 2021; Dosovitskiy et al., 2020; Minderer et al., 2022), while dense-to-sparse learning and counterfactual audit make their route trainable and testable (Shazeer et al., 2017; Fedus et al., 2022; Hinton et al., 2015; DeYoung et al., 2020; Jacovi and Goldberg, 2020).

3 Method

3.1 Method Overview

The core design choice in VISUALTHINK-VLA is to condition the frozen VLA backbone on a small set of learned visual evidence states. As shown in Figure 2, VISUALTHINK-VLA first constructs a six-channel candidate evidence bank, screens out two low-utility channels, and then uses a task-adaptive router over the retained four effective channels. The routed evidence is mapped into visual states through the Visual State Composer before action decoding. Under this formulation, *think with images* means that action prediction is guided by compact, routed, image-grounded states.

3.2 Channel Evidence Interface

At decision step t , the policy receives the current RGB observation x_t , the previous observation x_{t-1} , and the language instruction q . From this context, we derive a six-channel candidate evidence bank

$$\mathcal{E}_t = \{e_t^c \mid c \in \mathcal{C}_{\text{cand}}\}, \quad (1)$$

where

$$\mathcal{C}_{\text{cand}} = \{\text{bbox}, \text{edge}, \text{motion}, \text{relation}, \text{depth}, \text{segment}\}. \quad (2)$$

Each e_t^c is a compact evidence vector. The construction depends on the observation pair and, for

instruction-grounded channels, on q :

$$e_t^c = g_c(x_{t-1}, x_t, q). \quad (3)$$

The six candidate channels cover complementary visual information: bbox encodes object location, edge captures boundary geometry, motion summarizes short-horizon change, relation represents instruction-grounded spatial relations, depth provides monocular geometry, and segment marks object regions. Compact vector representations give the VLA a lightweight visual interface. This decomposition is aligned with recent object-centric and spatio-temporal grounding studies that separate object identity, region reference, and temporal change (Ravi et al., 2025; Yuan et al., 2025b,a). The detailed construction of each channel is described in the appendix. However, channel screening identifies depth and segment as low-utility cues in our benchmark setting: they are rarely selected, bring marginal success gains, and increase side-perception overhead. We therefore define the operational evidence set and bank as

$$\mathcal{C}_{\text{op}} = \{\text{bbox}, \text{edge}, \text{motion}, \text{relation}\}, \quad (4)$$

$$\mathcal{E}_t^{\text{op}} = \{e_t^c \mid c \in \mathcal{C}_{\text{op}}\}. \quad (5)$$

Accordingly, **FULLSOFT** sets the route mask to 1 over $\mathcal{E}_t^{\text{op}}$, while **VISUALTHINK-VLA** uses a router-selected sparse mask over the same four-channel bank.

3.3 Task-Adaptive Orchestration

A task-adaptive router predicts channel probabilities

$$m_t^{\text{soft}} = r_\phi(x_{t-1}, x_t, q, \mathcal{E}_t^{\text{op}}), \quad (6)$$

and a hardening operator $H(\cdot)$ converts them into an inference-time route mask

$$m_t^{\text{hard}} = H(m_t^{\text{soft}}), \quad m_t^{\text{hard}} \in \{0, 1\}^{|\mathcal{C}_{\text{op}}|}. \quad (7)$$

Each operational channel is encoded as a 64-d feature. The router combines these features

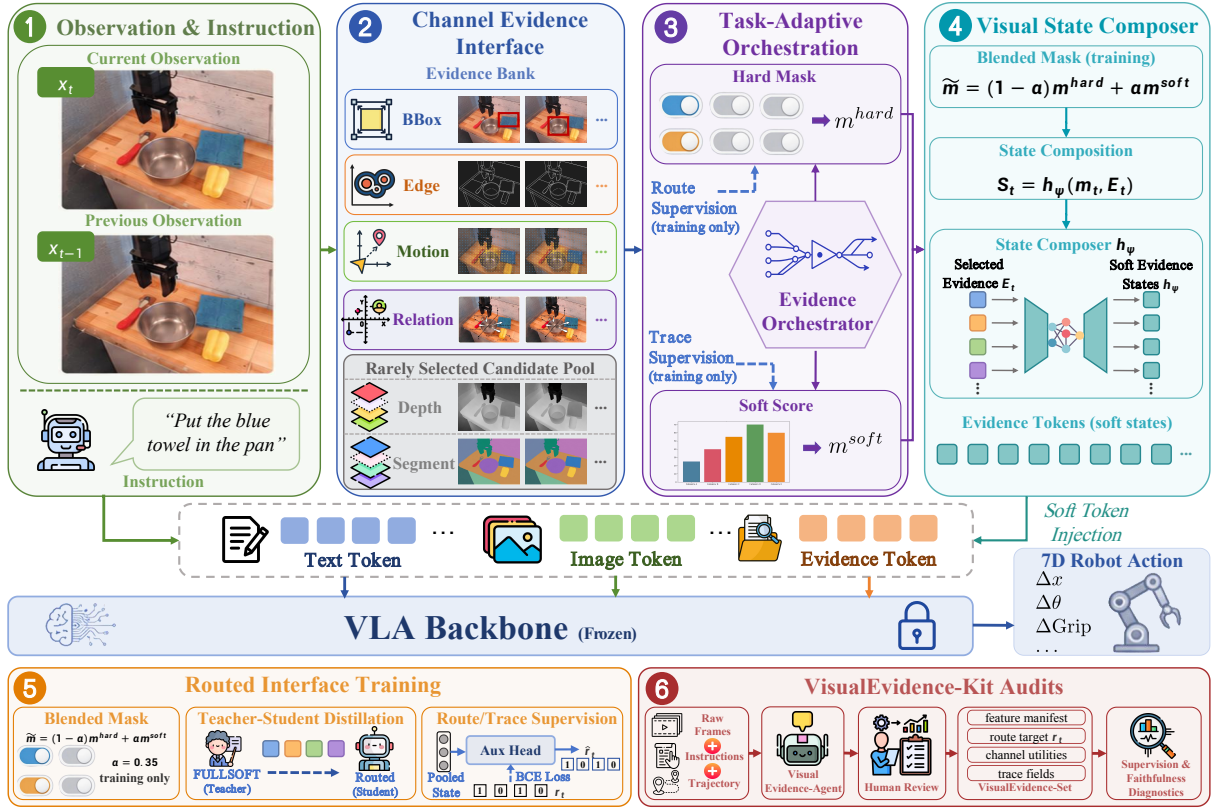


Figure 2: Overview of the VISUALTHINK-VLA pipeline. Dashed arrows denote training-only supervision.

with a 128-d inexpensive visual-text context; its sequence-text encoder uses 96-d token embeddings and a 128-d hidden state. A two-layer MLP with GELU feeds a channel-utility head (sigmoid scores) and a budget head that predicts $k_t \in \{1, \dots, 4\}$. Accordingly, H performs budget-aware top- k_t selection. Training uses a straight-through hard mask, whereas deployment uses the predicted hard mask only. Given a mask m , we denote the routed evidence as

$$\mathcal{E}_t^{\text{op}}(m) = \{m_{t,c} e_t^c \mid c \in \mathcal{C}_{\text{op}}\}. \quad (8)$$

At inference time, VISUALTHINK-VLA uses a single task-adaptive router across datasets. The hard mask m_t^{hard} activates only the evidence channels selected for the current decision. This is the main efficiency mechanism: all four operational channels are available to the interface, but the action decoder is exposed only to routed evidence states.

This separation also clarifies the relation among route variables. The predicted route is m_t^{soft} , the inference-time selection is m_t^{hard} , and the VisualEvidence-Set target r_t used below is a supervised route label. The model is trained to align

its predicted route with r_t , but at inference time it uses only its own router prediction.

3.4 Visual State Composer

The Visual State Composer is the concrete module denoted by h_ψ . Its SoftEvidenceAdapter is a Linear-GELU-Linear projection with a 768-d intermediate layer, channel-type and absent-channel embeddings, and two global evidence tokens. It projects the routed channel vectors into a small set of learned visual states,

$$S_t = h_\psi(\mathcal{E}_t^{\text{op}}(m)), \quad (9)$$

and inserts them as conditioning states before action decoding. The frozen VLA backbone then predicts the action as

$$a_t = f_\theta(x_t, q, S_t), \quad \theta \text{ frozen}. \quad (10)$$

Because the composer consumes already-routed evidence, it provides a lightweight learned interface that directly produces action-conditioning states at inference time.

3.5 Training the Routed Interface

Pure hard routing was brittle in difficult manipulation settings. During training, we therefore use a

blended route mask

$$\bar{m}_t = (1 - \alpha)m_t^{\text{hard}} + \alpha m_t^{\text{soft}}, \quad (11)$$

with $\alpha = 0.35$, and compute $S_t = h_\psi(\mathcal{E}_t^{\text{op}}(\bar{m}_t))$. At inference time, $\bar{m}_t$ is replaced by m_t^{hard} .

To transfer the capacity of dense evidence integration into the sparse route, we train the routed adapter with logits distillation from the four-channel FULLSOFT teacher (Hinton et al., 2015). Let p_T^τ and p_S^τ denote teacher and student action-token distributions at temperature τ . The dynamic loss is

$$\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{action}} + \lambda_{\text{distill}} \tau^2 \text{KL}(p_T^\tau \| p_S^\tau), \quad (12)$$

with $\lambda_{\text{distill}} = 0.2$ and $\tau = 1.5$.

VisualEvidence-Kit provides a supervised route target $r_t \in \{0, 1\}^{|\mathcal{C}_{\text{op}}|}$. A training-only auxiliary head predicts $\hat{r}_t$ from the pooled evidence state and route features. The final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dyn}} + \lambda_{\text{trace}} \mathcal{L}_{\text{BCE}}(\hat{r}_t, r_t). \quad (13)$$

This route/trace supervision regularizes the evidence pathway during training. Deployment uses observations, the instruction, operational evidence, and the router’s predicted hard mask. The route target r_t , utility ranks, trace fields, and auxiliary head serve as training and audit signals.

4 VisualEvidence-Kit

To train and examine a policy that reasons through sparse visual evidence, we construct **VisualEvidence-Kit**, a route-grounded supervision and audit resource for VLA control. VisualEvidence-Kit contains two coupled components: a **VisualEvidence-Agent** that derives and governs evidence instructions from manipulation trajectories, and the resulting **VisualEvidence-Set** that supports route supervision and counterfactual auditing. Its instructions explicitly describe the visual channels relevant to each action decision and the routed policy’s dependence on them. This design is also informed by data-centric learning work that uses adaptive pseudo-labeling and active annotation to handle imperfect supervision (DeYoung et al., 2020; Zhang et al., 2022).

4.1 VisualEvidence-Agent

The VisualEvidence-Agent converts raw frames and trajectory metadata into governed route-grounded instructions through four stages.

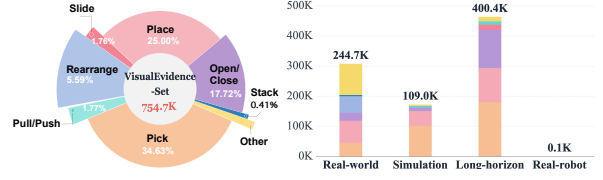


Figure 3: Data statistics of VisualEvidence-Set

(i) *Evidence extraction*: applies the candidate channel extractors to construct a feature manifest for each decision context.

(ii) *Route and utility assessment*: replaces each channel in the dense teacher with its learned absent-channel representation, measures normalized action degradation, and adjusts it by measured channel cost. A channel enters r_t only when adjusted utility exceeds both 10^{-4} and 25% of the best-channel utility; stage priors only break ties, and utility rank follows the adjusted score. Post-training interventions on the routed student are kept separate from teacher-derived training utilities.

(iii) *Trace construction*: records the manipulation stage, primitive, evidence dependence, difficulty, and selected evidence as a structured channel-grounded trace.

(iv) *Human review* verifies visible grounding, action/stage consistency, and channel support; ambiguous samples are adjudicated or sent to review/drop. As illustrated in Figure 9, this separates noisy supervision from high-confidence training or audit targets (Zhang et al., 2023, 2024). The six-reviewer protocol and dataset roles are detailed in the appendix.

4.2 VisualEvidence-Set

The resulting VisualEvidence-Set organizes heterogeneous manipulation information around the evidence demands of control. Each instruction contains the observation, instruction context, feature manifest, supervised route target r_t , counterfactual channel utilities, and channel-grounded trace fields. We further categorize the instructions along complementary axes, including scene domain, action category, manipulation stage, and routed visual evidence type, following the broader trend of heterogeneous knowledge organization in large vision-language systems (Bai et al., 2025; Lin et al., 2025). As summarized in Figure 3, the dataset displays the distribution of instructions for different action types and scene domains. In addition, each instruction is annotated by visual fea-

tures, task stage, and difficulty level, enabling both router training and diagnostic evaluation across evidence requirements.

To support different use cases, we curate the constructed instructions into increasingly stringent subsets. *Full-Clean* is used for corpus-level statistics and weighted training after basic consistency filtering; *HQ-Trace* retains reliable structured traces for route- and trace-supervised refinement; and *Gold-Faithfulness* serves as a high-reliability subset of 754.7k instructions for counterfactual audit experiments.

VisualEvidence-Kit has three distinct roles: route/trace supervision, governed dataset construction, and counterfactual audit through evidence removal or perturbation. Route labels regularize training; schema alignment measures annotation adherence, and intervention on a separately held-out human-audited subset measures behavioral faithfulness. Inference uses the policy’s predicted route over operational evidence. The workflow and statistics are detailed in Appendix C.

5 Experimental Setup

Benchmark Roles. The multi-dataset control benchmark evaluates VISUALTHINK-VLA on BridgeData V2, Fractal, RoboTurk, LIBERO, and UT Austin MUX (Walke et al., 2023; O’Neill et al., 2024; Mandlekar et al., 2018; Liu et al., 2023; Shah et al., 2023). VisualEvidence-Kit provides the route-grounded audit layer, and the real-robot study evaluates closed-loop deployment on a separate tabletop platform.

Baselines. The external comparisons cover textual traces (ECOT), image-grounded reasoning (TraceVLA, SpatialVLA), and stronger reasoning-policy variants. Populated entries in the main comparison are same-setting evaluations in our pipeline using released weights/checkpoints and compatible action interfaces; blank entries identify unavailable comparable combinations. Within our method family, we report a frozen BASEVLA baseline, a dense FULLSOFT reference, and the sparse VISUALTHINK-VLA variant.

Compared Systems. **BASEVLA-only** runs the frozen pretrained OPENVLA backbone, a 7B-parameter VLA model, with the original image and instruction inputs. **FULLSOFT** enables all channels in the four-channel operational bank for every sample. **VISUALTHINK-VLA** uses a task-adaptive router to activate a sparse subset of chan-

nels.

Real-Robot Protocol. The real-robot study uses a desk-mounted **PIPER NERO (7F)** 7-DoF arm with a fixed external RGB camera (Kim et al., 2024; Zawalski et al., 2024; Black et al., 2024; Intelligence et al., 2025; Shukor et al., 2025). The policy predicts end-effector-style actions, and a robot-side controller converts them into executable commands. The task suite covers **language-specified multi-object pick-place, relation-sensitive placement, contact-sensitive object reorientation, and two-stage compositional manipulation.**

Metrics. The primary control metric is `success_rate`. The main benchmark’s `avg_step_latency_s` is policy-side batch-1 wall-clock time after warm-up: router/composer computation plus VLA action decoding. We separately report evidence extraction and full end-to-end online latency in the appendix. We also report `avg_completion_time_s` for real-robot evaluation and route-grounded audit metrics for VisualEvidence-Kit.

6 Results and Analysis

Primary Control Comparison. Table 2 shows that VISUALTHINK-VLA achieves a favorable accuracy-latency trade-off across diverse control settings. Compared with the matched BASEVLA re-evaluation, VISUALTHINK-VLA improves success on seven of eight benchmarks, with the largest gains on BridgeData V2, the LIBERO family, and UT Austin MUX. It also reduces BridgeData V2 latency from 8.377s with ECOT to 0.367s while improving success. Compared with the dense FULLSOFT teacher, VISUALTHINK-VLA preserves comparable accuracy while reducing latency on every benchmark and limiting channel exposure to task-relevant evidence.

Backbone Portability.¹ VISUALTHINK-VLA provides a plug-in routed visual-evidence layer for VLA backbones. On the held-out VisualEvidence-Set test split, the same interface yields consistent gains with modest policy-side latency increases across OpenVLA and three stronger backbones whose action-generation mechanisms and deployment recipes differ.

Internal Interface Comparison.² The con-

¹Detailed backbone-portability results are provided in Appendix A.2.

²Detailed internal-interface results are provided in Appendix A.3.

Method	Real-World Scenes						Simulation Scenes						Long-Horizon Tasks			
	BridgeData V2		Fractal		RoboTurk		LIBERO-Object		LIBERO-Goal		LIBERO-Spatial		LIBERO-Long		UT Austin MUTEX	
	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓
<i>Textual CoT / autoregressive reasoning before action</i>																
ECoT	85.09	8.377	32.87	6.702	15.78	6.397	-	-	-	-	-	-	-	-	20.32	6.255
<i>Reinforced or hybrid thinking-action decoding</i>																
DeepThinkVLA-RL	27.93	1.204	33.67	1.219	2.12	1.323	98.58	1.359	96.32	1.277	93.97	1.224	95.02	1.202	26.09	0.772
<i>Spatial or image-grounded thinking</i>																
TraceVLA	86.87	0.404	86.25	0.402	93.27	0.416	85.68	0.425	75.24	0.407	85.90	0.428	52.62	0.440	1.10	0.403
SpatialVLA	86.57	0.589	67.85	0.582	91.50	0.585	87.50	0.481	79.38	0.479	86.86	0.490	53.25	0.498	0.65	0.544
InternVLA-MI	11.18	1.153	23.46	1.153	8.51	1.151	96.04	1.021	92.52	1.021	92.26	1.022	90.65	1.021	12.37	1.144
<i>OpenVLA-family re-evaluation</i>																
BASEVLA	75.37	0.345	87.45	0.345	95.56	0.342	89.30	0.343	80.83	0.343	84.60	0.342	54.86	0.302	41.09	0.349
FULLSOFT	88.45	0.447	90.38	0.448	96.32	0.445	97.78	0.445	97.14	0.445	96.39	0.445	96.24	0.497	77.10	0.551
VISUALTHINK-VLA	89.49	0.367	90.82	0.367	96.10	0.415	97.74	0.385	97.05	0.345	96.69	0.356	95.87	0.421	77.26	0.451

Table 2: Same-setting performance comparison using released weights/checkpoints and compatible action interfaces. Blank entries indicate unavailable comparable checkpoints or protocols. Latency is policy-side. ↑ and ↓ indicate better directions.

trolled comparison isolates two effects. BASEVLA→FULLSOFT measures the benefit of visual evidence; FULLSOFT→VISUALTHINK-VLA measures the efficiency of sparse routing. VISUALTHINK-VLA preserves comparable success to the dense teacher (90.10% vs. 89.83%) while reducing policy-side latency (0.395 s vs. 0.470 s). Prompt-text and heavy dense comparisons further separate visual-state evidence from textualization and unfiltered side cues.

The screening results in Figure 5 further show that the four-channel operational subset retains the main control gains of the initial six-channel bank. Its compact visual-state interface achieves a favorable performance–runtime balance compared with dense screened candidates and prompt-text evidence.

Real-Robot Evaluation. Table 3 shows that the trade-off transfers to live closed-loop control. VISUALTHINK-VLA outperforms BASEVLA on all four task families and improves over FULLSOFT on three of them. The only exception is contact-sensitive reorientation, where always-on four-channel dense evidence is slightly stronger. Nevertheless, VISUALTHINK-VLA remains more efficient than FULLSOFT in completion time (25.6 s vs. 30.2 s) while selecting only 1.83 channels on average.

VisualEvidence-Kit Analysis. VisualEvidence-Kit tests whether routed evidence is inspectable and behaviorally relevant. On the held-out audit set, VISUALTHINK-VLA achieves the strongest route-grounded audit score (0.840) and utility overlap (0.984); prompt-text evidence reaches 0.391 on the corresponding audit score. Stage-wise routing is also sensitive to task phase: bbox dominates during approach, motion rises sharply

at grasp, and edge remains consistently active.

Figure 7 reports the full stage-conditioned routing pattern behind these audit results. The figure supports the same conclusion as the aggregate scores: the router changes evidence use with the phase of manipulation.

7 Ablation Studies

Feature Ablations. Figure 4 evaluates the retained four-channel model on BridgeData V2 and LIBERO-Long, which stress cluttered manipulation and long-horizon control, respectively. Removing each retained channel reduces success relative to the full model, with task-dependent effect sizes. BridgeData V2 is more sensitive to removing bbox, motion, and relation cues, whereas LIBERO-Long shows the largest drops after removing edge cues or retaining only narrow channel pairs. These results demonstrate the complementary contributions of the routed evidence channels.

Orchestration Ablations. Orchestration ablations show that the final sparse routing recipe improves both trainability and efficiency.

Figure 6 separates the effects of training recipe, gate design, and blend coefficient. Soft-hard collaborative routing improves over hard sparse masks, increasing success from 82.86% to 84.19% on BridgeData V2 and from 85.62% to 93.82% on LIBERO-Long. Adding teacher-student optimization further raises success to 89.46% and 94.87%, respectively. The task-adaptive router then maintains comparable success to the sequence-text gate while selecting fewer channels on average, supporting the efficiency of the final orchestration design.

Route-Trace Ablations. Route and trace supervi-

Task family	Evidence	Metric	BASEVLA	FULLSOFT	VISUALTHINK-VLA
Multi-object pick-place	bbox + distractors	Success (%) $\uparrow$	48.6	73.4	75.6
Relation-sensitive placement	relation	Success (%) $\uparrow$	45.6	63.7	67.2
Contact-sensitive reorientation	edge/motion	Success (%) $\uparrow$	63.4	85.8	83.5
Two-stage compositional task	stage-wise routing	Success (%) $\uparrow$	26.5	57.4	59.2
All tasks	efficiency	Avg. completion time (s)	22.7	30.2	25.6
		$\downarrow$			
All tasks	route sparsity	Avg. selected ch.	0	4.0	1.83

Table 3: Focused real-robot closed-loop evaluation on the four-task tabletop suite. Each task family is evaluated over 50 trials. Up arrows indicate higher-is-better metrics, and down arrows indicate lower-is-better metrics.

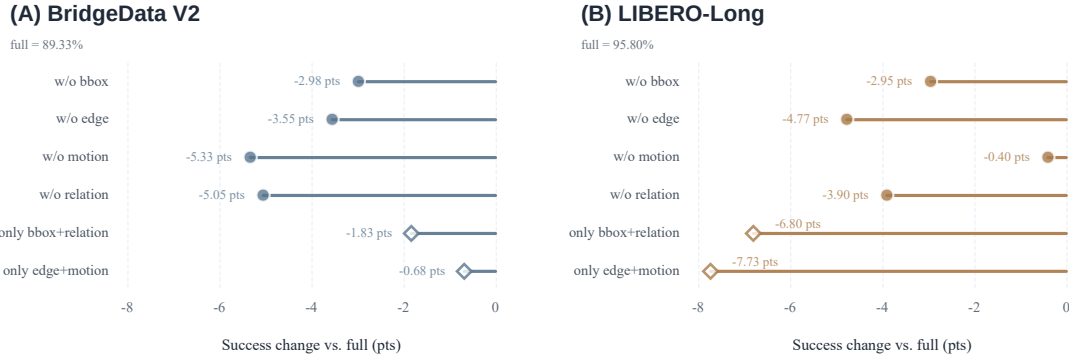


Figure 4: Feature ablations on BridgeData V2 and LIBERO-Long. The plot reports changes in success rate relative to the full retained-channel model; “w/o” denotes removing the specified evidence channel, while “only” denotes retaining only the indicated channel pair.

sion mainly improve faithfulness, with moderate success changes. Removing route supervision reduces route alignment from 0.929 to 0.758, while removing trace supervision reduces utility mention from 0.984 to 0.738. Replacing the channel-grounded target with a free-form rationale baseline causes the strongest degradation, lowering route alignment to 0.052 and utility mention to 0.039.

Together, Table 9 indicates that structured, channel-grounded supervision improves inspectable intermediate reasoning in VLA control.

8 Conclusion

VISUALTHINK-VLA introduces a sparse image-grounded evidence interface for frozen VLA policies. In the OPENVLA instantiation studied here, the routed variant preserves much of the benefit of dense soft evidence while reducing control-time overhead, and it remains effective in a real-robot closed-loop setting. The results indicate that, for embodied control, compact structured evidence in visual space is a more practical intermediate reasoning substrate than long textual reasoning traces.

Limitations

The current implementation relies on manually designed evidence channels and external perception modules; scaling the interface to new embodiments and open-world tasks may require different channels, extractors, or routing budgets. In particular, the observed low utility of depth and segmentation is specific to our current extractors, representations, and benchmark distribution. Other signals such as tactile feedback, force sensing, audio, or longer-term memory remain future work. The current portability evidence covers the four evaluated backbones and their action parameterizations and deployment stacks. Finally, broader real-world use requires appropriate monitoring and task-specific safety constraints.

Acknowledgments

This work was supported by the National Key R&D Program of China (2025ZD0123100), the National Natural Science Foundation of China (62272411 and 624B2027), the Zhejiang Provincial Natural Science Foundation (LRG25F020001), and the Key R&D Program of Zhejiang Province (2025C01030).

References

- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Suneel Belkhal, Yuchen Cui, and Dorsa Sadigh. 2023. Hydra: Hybrid robot actions for imitation learning. In *Conference on Robot Learning*, pages 2113–2133. PMLR.
- Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. Pi-0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.
- Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. 2024. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*.
- Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. 2025. Internv1a-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*.
- Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2025. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- Zhekai Duan, Yuan Zhang, Shikai Geng, Gaowen Liu, Joschka Boedecker, and Chris Xiaoxuan Lu. 2025. Fast ecot: Efficient embodied chain-of-thought via thoughts reuse. *arXiv preprint arXiv:2506.07639*.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. 2025. Pi-0.5: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sankeeti, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.

- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.
- Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. 2018. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR.
- Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer.
- Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.
- Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128.
- Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. 2023. Mutex: Learning unified policies from multimodal task specifications. *arXiv preprint arXiv:2309.14320*.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarsz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2023. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR.
- Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. 2025. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.
- Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on robot learning*, pages 1723–1736. PMLR.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. 2026. Visual planning: Let’s think only with images. In *International Conference on Learning Representations*, volume 2026, pages 36706–36739.
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911.
- Cheng Yin, Yankai Lin, Wang Xu, Sikyuen Tam, Xiangrui Zeng, Zhiyuan Liu, and Zhouping Yin. 2025. Deepthinkvla: Enhancing reasoning capability of vision-language-action models. *arXiv preprint arXiv:2511.15669*.
- Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. 2025a. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18970–18980. IEEE.
- Yuqian Yuan, Wenqiao Zhang, Xin Li, Shihao Wang, Kehan Li, Wentong Li, Jun Xiao, Lei Zhang, and Beng Chin Ooi. 2025b. Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity. *arXiv preprint arXiv:2510.23603*.

Yuqian Yuan, Wenqiao Zhang, Juekai Lin, Yu Zhong, Mingjian Gao, Binhe Yu, Yunqi Cao, Wentong Li, Yueting Zhuang, and Beng Chin Ooi. 2026. Lmms meet object-centric vision: Understanding, segmentation, editing and generation. *arXiv preprint arXiv:2604.11789*.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.

Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. 2023. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1423–1432. IEEE.

Wenqiao Zhang, Zheqi Lv, Hao Zhou, Jia-Wei Liu, Juncheng Li, Mengze Li, Yunfei Li, Dongping Zhang, Yueting Zhuang, and Siliang Tang. 2024. Re-visiting the domain shift and sample uncertainty in multi-source active domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16751–16761.

Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. 2022. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20634–20644. IEEE.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. 2025. Cotvla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713. IEEE.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. 2025. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, volume 2025, pages 54277–54296.

Yu Zhong, Tianwei Lin, Ruike Zhu, Yuqian Yuan, Haoyu Zheng, Liang Liang, Wenqiao Zhang, Feifei Shao, Haoyuan Li, Wanggui He, et al. 2026. Unified personalized understanding, generating and editing. *arXiv preprint arXiv:2601.06965*.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. 2023. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pages 1199–1210. PMLR.

Appendix

This appendix provides supporting implementation and evaluation details. We document the VisualEvidence-Set governance pipeline, describe the evidence-channel construction, spell out the real-robot task suite, and collect qualitative and audit details.

A Additional Experimental Results

A.1 End-to-End Latency Accounting

The main tables use *policy-side step latency*, which includes router/composer computation and VLA action decoding. Table 4 reports the full online path on VisualEvidence-Set. The four evidence extractors total 1.09 ms per step; adding them to 0.021 s routing/composition and 0.374 s action decoding gives 0.396 s end-to-end latency. Thus, the low-latency claim is supported by the full pipeline, while the main-table metric isolates policy-side differences across methods.

A.2 Backbone Portability

Table 5 reports the full backbone-portability comparison referenced in the main paper. Every method uses the same held-out VisualEvidence-Set test split, so the success and policy-side batch-1 latency changes are directly comparable across backbone and action-generation designs.

A.3 Effect Isolation and Statistical Reliability

The controlled comparisons in Table 6 separate evidence availability from routing efficiency. BASEVLA versus FULLSOFT isolates the value of exposing visual evidence; FULLSOFT versus VISUALTHINK-VLA isolates sparse routing against the all-channel teacher; prompt text versus routed soft tokens isolates autoregressive textualization; and heavy dense cues versus routed soft tokens isolates unfiltered side perception. As reported in Table 8, the sparse route preserves the dense-evidence gain while reducing latency and the number of active channels.

Because the success difference between FULLSOFT and VISUALTHINK-VLA is modest, Table 7 reports five independent seeds. The routed variant has a higher mean on all five aggregate benchmark groups, while the standard deviations show that the contribution is principally a more favorable success-latency trade-off.

Method	BBox	Edge	Motion	Relation	Router + composer	Action decoding	Full online
VISUALTHINK-VLA	0.42 ms	0.11 ms	0.21 ms	0.35 ms	0.021 s	0.374 s	0.396 s

Table 4: End-to-end online latency breakdown on VisualEvidence-Set. Evidence extraction is included in the full online measurement.

Backbone	Base policy		+VISUALTHINK-VLA		Change	
	Succ. (%)	Lat. (s)	Succ. (%)	Lat. (s)	Δ Succ.	Δ Lat.
OPENVLA (Kim et al., 2024)	71.73	0.340	84.28	0.374	+12.55	+0.034
π_0 (Black et al., 2024)	43.28	0.381	54.49	0.422	+11.21	+0.041
$\pi_{0.5}$ (Intelligence et al., 2025)	73.94	0.449	82.02	0.486	+8.08	+0.037
GR00T-N1.7 (Bjorck et al., 2025)	65.06	0.428	79.35	0.460	+14.29	+0.032

Table 5: Backbone portability on the held-out VisualEvidence-Set test split; every row uses the same split and reports policy-side batch-1 latency.

Evidence-channel screening. Figure 5 reports the channel-usage and cost–accuracy diagnostics behind the operational evidence bank. Under the current extractors, representation, and benchmark distribution, depth and segmentation have low utility. In representative failures, monocular depth is too coarse for fine contact decisions; segmentation can be redundant with bbox/edge/relation and introduce noisy object partitions. Their low selection frequency and cost–accuracy profiles support the retained four-channel default bank.

Interface comparison. Table 8 provides the full internal interface comparison summarized in the main text. It contrasts routed soft tokens with prompt-text evidence, heavy dense perception, and dense soft tokens across the benchmark families.

Routing and supervision ablations. Figure 6 separates the effects of the training recipe, gate design, and blend coefficient. The teacher–student recipe and task-adaptive routed gate provide the final accuracy–sparsity trade-off used by VISUALTHINK-VLA.

Figure 7 complements the aggregate audit scores with a stage-conditioned view of route use. The changing channel frequencies across approach, grasp, transport, and placement demonstrate stage-adaptive evidence selection.

Finally, Table 9 quantifies the contribution of route, trace, and utility supervision. Removing the structured targets weakens route alignment and utility mention, while the free-form target produces the largest degradation.

A.4 Counterfactual Faithfulness Beyond Schema Agreement

Route and trace targets regularize training, while behavioral faithfulness is measured through interventions on a separately held-out, human-audited subset. Table 10 shows task-dependent effects: removing the top routed channel causes clear drops on BridgeData V2 and UT Austin MUTEX, whereas some LIBERO subsets have small or even negative changes. Non-routed removals have a smaller mean absolute effect, and relation perturbations change both success and action distributions. These tests complement the in-schema trace/utility scores in Table 13.

B Benchmark Success–Latency Frontiers

Figure 8 visualizes the main-paper benchmark suite as success–latency frontiers. Textual reasoning baselines can improve selected tasks but remain far outside the sub-second control regime, while the sparse routed interface stays close to the low-latency OpenVLA-family region and preserves most of the gains of dense visual evidence.

C VisualEvidence-Kit Workflow

The VisualEvidence-Agent builds the VisualEvidence-Set in a staged manner: it first derives the candidate evidence channels from raw observations, then aggregates route targets and utility signals, and finally sends the generated records to a manual review and annotation stage before exporting the governed benchmark entries with channel-grounded trace fields.

Comparison	Isolated factor	Main conclusion
BASEVLA vs. FULLSOFT	Visual-evidence availability	Visual evidence improves action prediction in the controlled comparison.
FULLSOFT vs. VISUALTHINK-VLA	Sparse route vs. all channels	Sparse routing preserves dense-evidence performance while reducing latency and active channels.
Prompt text vs. routed tokens	Textualized vs. visual-state evidence	Soft visual states are more efficient and better aligned with control.
Heavy dense vs. routed tokens	Dense side cues vs. compact evidence	Dense cues can add interference and runtime cost.

Table 6: Effect-isolation interpretation of the controlled internal comparisons.

Method	BridgeData V2	Fractal	RoboTurk	LIBERO family	UT Austin MUTEX	Average
FULLSOFT	88.47±0.16	90.25±0.19	95.98±0.44	96.83±0.07	76.71±0.54	89.65
VISUALTHINK-VLA	89.42±0.15	90.68±0.34	96.26±0.55	96.88±0.07	78.13±0.67	90.27

Table 7: Success rate (% , mean $\pm$ standard deviation) over five independent random seeds.

Figure 9 summarizes this workflow and the resulting record schema, including the evidence manifest, route mask, and structured trace fields used for supervision and audit.

Figure 10 and table 11 summarize how raw route-grounded traces are governed before use. Full-Clean removes invalid or inconsistent records and supports broad dataset statistics and weighted training. HQ-Trace keeps higher-quality channel-grounded traces for trace-supervised refinement. Gold-Faithfulness is the strictest subset and is used for high-confidence counterfactual audit experiments.

The governance pipeline starts with 942.8k raw records and produces 923.7k Full-Clean, 877.8k HQ-Trace, and 754.7k Gold-Faithfulness records, with 45.9k assigned to review/drop. Automatic consistency filtering and direct human verification are tracked separately. Full-Clean supports weighted training, HQ-Trace supports route/trace refinement, and the held-out Gold-Faithfulness portion supports the main counterfactual analysis.

Evidence-channel construction. The six-channel candidate bank is extracted offline before policy inference. **bbox** uses open-vocabulary detection with Grounding DINO / OWL-ViT-style localization to encode coarse object location and target extent (Liu et al., 2024; Minderer et al., 2022). **edge** uses SAM2 mask proposals and boundary rasterization to capture object contours and local shape discontinuities (Ravi et al., 2025). **motion** uses short-horizon frame differencing over (x_{t-1}, x_t) together with SAM2-based temporal mask propagation and lightweight code-based motion rendering to capture immediate scene change (Ravi et al., 2025). **relation** encodes instruction-grounded spatial geometry with Qwen2.5-VL parsing and deterministic relation

rendering, such as left-of, in-front-of, or inside relations (Bai et al., 2025). **depth** uses Depth Anything V2 to estimate dense monocular depth and pools it into a compact geometry cue (Yang et al., 2024). **segment** uses detector-prompted SAM2 masks to encode object regions and coarse scene partitioning (Liu et al., 2024; Ravi et al., 2025). At policy inference time, VISUALTHINK-VLA consumes the compact evidence vectors and router-selected soft states.

D Real-Robot Task Suite

The main paper reports only the aggregate real-robot result table because the primary point is to compare deployment-time behavior across BASEVLA, FULLSOFT, and VISUALTHINK-VLA. This section spells out the tabletop task families behind that summary. The task design follows the mechanism of the method: one family stresses target localization under distractors, one isolates instruction-grounded spatial relations, one emphasizes contact-sensitive pose change, and one tests stage-dependent evidence usage within a short compositional task. Table 12 lists representative instructions for each family.

The deployment platform is a desk-mounted **PIPER NERO (7F)** 7-DoF arm with **1.5 kg** payload, **580 mm** reach, and ± 0.1 mm repeatability. The arm is controlled from a host PC through the vendor **CAN** interface, while RGB observations are captured by a fixed external camera.

These task definitions provide a compact description of the real-robot evaluation suite.

E Qualitative Case Studies

Figure 11 provides qualitative examples of the routed visual evidence used in representative closed-loop cases.

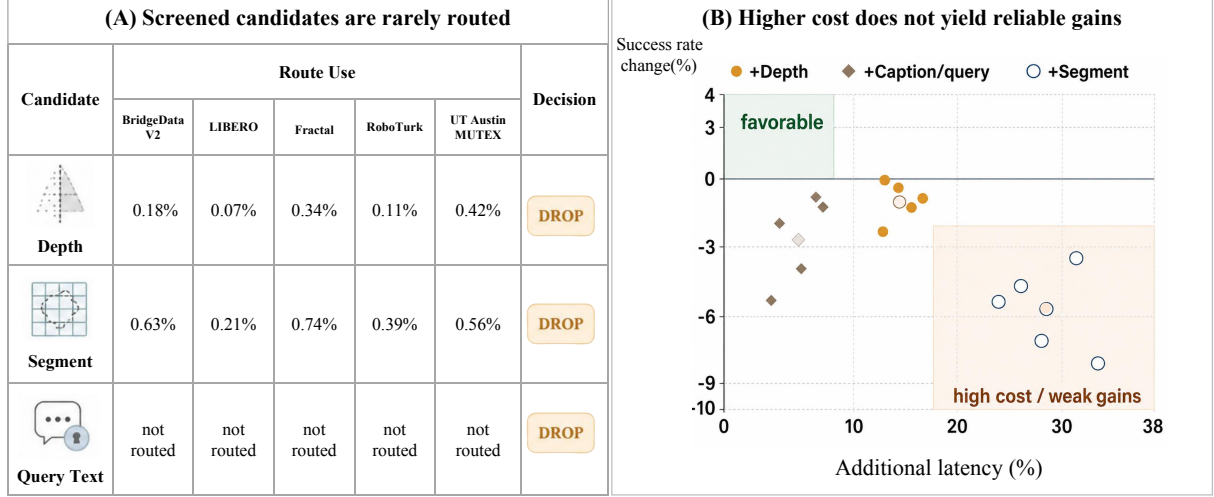


Figure 5: Evidence-channel screening for the retained routed evidence bank.

Variant	Evidence path	Real-World Scenes						Long-Horizon & Simulation Tasks				Average	
		BridgeData V2		Fractal		RoboTurk		LIBERO family		UT Austin MUTEX		Succ. (%) ↑	Lat. (s) ↓
		Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓	Succ. (%) ↑	Lat. (s) ↓		
BASEVLA-only	none	75.37	0.345	87.45	0.345	96.56	0.342	77.40	0.333	41.09	0.349	75.57	0.343
Prompt-text evidence	prompt text	87.49	1.384	86.82	1.389	94.90	1.448	95.07	1.453	71.76	1.467	87.21	1.428
Heavy dense	dense side cue	86.69	0.536	88.22	0.542	93.20	0.607	91.67	0.613	72.71	0.663	86.50	0.592
FULLSOFT	dense soft tokens	88.45	0.447	90.38	0.448	96.32	0.445	96.89	0.458	77.10	0.551	89.83	0.470
VISUALTHINK-VLA	routed soft tokens	89.49	0.367	90.82	0.367	96.10	0.415	96.84	0.377	77.26	0.451	90.10	0.395

Table 8: Internal interface comparison. ↑ indicate higher-is-better metrics, and ↓ indicate lower-is-better metrics.

Variant	Success	Route aln.	Util. ment.	Δ TopRemove
VISUALTHINK-VLA	90.46%	0.929	0.984	-
w/o route sup.	87.40%	0.758	0.742	-3.06 pts
w/o trace sup.	87.33%	0.882	0.738	-3.13 pts
w/o utility rank	88.47%	0.847	0.793	-1.99 pts
free-form tgt.	82.43%	0.052	0.039	-4.97 pts

Table 9: Interpretability ablations on the VisualEvidence-Set trace-control slice. Here, *Route aln.* denotes route alignment, *Util. ment.* denotes utility mention, and Δ TopRemove denotes the performance change after removing the top-ranked routed channel.

The case studies illustrate how routed visual evidence changes action-relevant perception. In the lever-rotation example, the relation cue makes the target pose explicit: its removal leads the base policy to rotate in the wrong direction, while the routed policy identifies the lever’s orientation relative to the plane and selects the correct downward rotation. In the bowl-placement example, edge evidence sharpens the contact and grasp boundary, helping the policy close on the bowl securely. In the oven-tray example, bbox evidence localizes the handle under visual clutter and prevents early stopping caused by target mislocalization. Across the three cases, the selected channels correspond

to the active task demand: relation for pose and spatial orientation, edge for precise contact geometry, and bbox for target localization.

F VisualEvidence-Kit Audit and Faithfulness Details

VisualEvidence-Kit provides a supervision-and-audit layer for routed evidence. VisualEvidence-Agent builds the VisualEvidence-Set used for the in-schema trace-quality summary in Table 13, complementing the stage-conditioned routes in Figure 7 and behavioral interventions in Table 10.

Taken together with Figure 7 and table 10, Table 13 separates annotation adherence from behavioral faithfulness. Trace alignment and utility overlap are in-schema diagnostics; channel removal and relation perturbation test whether routed evidence changes the trained policy’s behavior. The governance taxonomy uses four compact axes: stage, primitive, evidence dependence, and difficulty.

G Human Review Protocol

Human review in VisualEvidence-Kit focuses on route-grounded records, evidence-channel labels, stage consistency, and trace quality. The

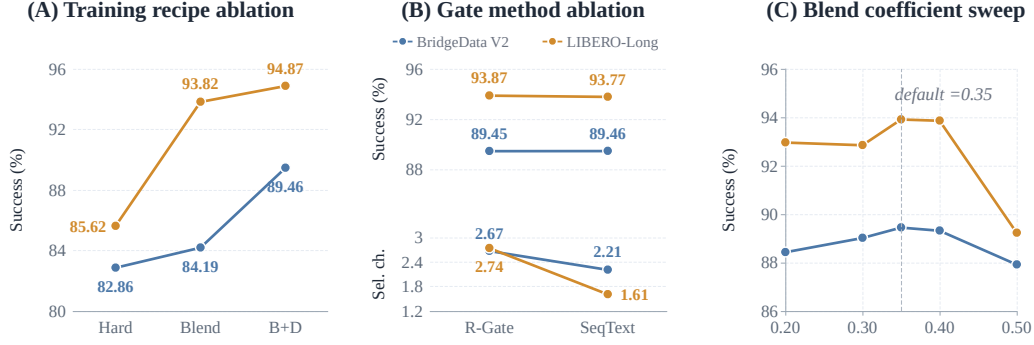


Figure 6: Evidence-orchestration ablations on BridgeData V2 and LIBERO-Long, where B+D denotes blend plus distillation, R-Gate denotes routed evidence gate, SeqText denotes the sequence-text gate baseline, and Sel. ch. denotes the average number of selected evidence channels.

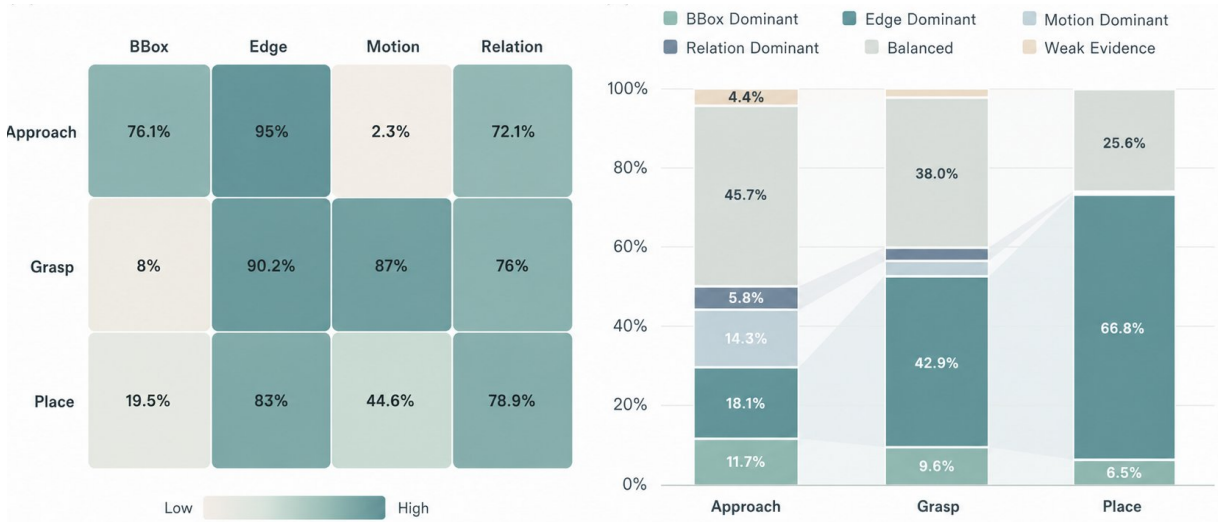


Figure 7: Post-hoc analysis of stage-wise evidence-routing patterns on VisualEvidence-Set.

team comprises six laboratory reviewers familiar with robot manipulation, followed by two senior-author audit passes. Guidelines require routes to be visibly supported, action-relevant, and stage-consistent; hallucinated, ambiguous, or unsupported traces are rejected. Conflicts use majority vote, with unresolved cases adjudicated by a senior author or assigned to review/drop. No crowd-working platform was used and no monetary payment was provided.

Route/trace overlap serves as an in-schema diagnostic because training-label construction, policy training, and audit share the same schema. The separately held-out, human-audited counterfactual subset supports the behavioral faithfulness analysis. Teacher-derived utilities used for supervision are distinct from interventions measured on the trained student.

The review process collected robot observa-

tions, task instructions, generated evidence labels, and audit fields, with no personal data from reviewers or external human subjects. Reviewers were informed that the reviewed records would be used for research dataset construction and evaluation. The study concerns robot-manipulation records and does not involve human-subject behavioral data or intervention with human participants; accordingly, no separate ethics-board protocol was sought.

H Artifact Licenses

We use external datasets, VLA backbones, and perception tools only under their original release conditions. The public robot-learning benchmarks used for evaluation, including BridgeData V2, Open X-Embodiment/Fractal, RoboTurk, LIBERO, and UT Austin MUTEX, are cited in the main experimental section and remain gov-

Audit condition	Result / diagnostic
Remove top routed channel	Success drop: 4.35 points on BridgeData V2 and 9.23 points on UT Austin MUTEX; several LIBERO subsets show smaller or negative changes.
Remove non-routed channel	Mean absolute success change is 0.42 points, versus a 1.45-point mean effect for top-channel removal.
Replace route target with free-form rationale	Route alignment and utility mention fall to 0.052 and 0.039, respectively.
Counterfactual relation perturbation	On the relation-sensitive subset, swapping left/right or near/far cues changes success by -2.1 points and normalized action L_1 by $+0.006$.

Table 10: Behavioral counterfactual audits of the routed evidence pathway.

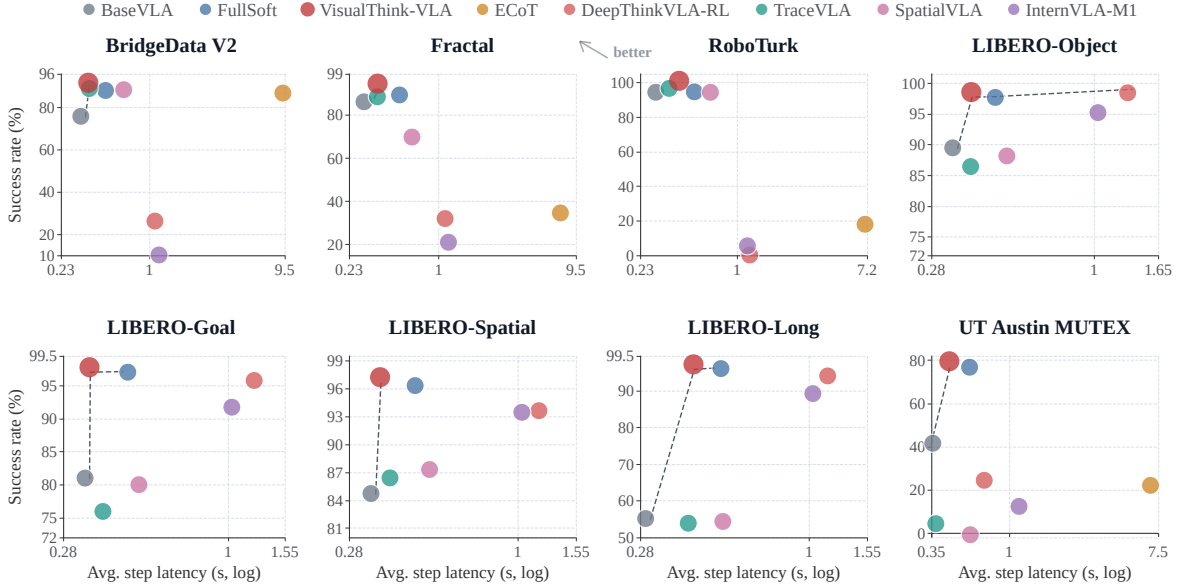


Figure 8: Success–latency trade-off across the eight main benchmarks.

erned by their respective dataset licenses and terms of use. The VLA backbones and comparison systems, including OPENVLA, Octo, SmolVLA, and related reasoning-augmented VLA methods, are used as research artifacts according to the conditions specified by their original authors. The perception back-ends used to construct visual evidence channels, including Grounding DINO, OWL-ViT, SAM2, Qwen2.5-VL, and Depth Anything V2, are likewise cited and used under their original model or code licenses.

Our use of these artifacts is consistent with their intended research use. External datasets are used for benchmark evaluation, VLA backbones are used for policy comparison and adaptation, and perception models are used to derive auxiliary visual evidence for research analysis. VisualEvidence-Kit and VisualEvidence-Set are intended only for VLA supervision, audit, and diagnostic evaluation. Any derivatives should remain within the access conditions of the original

datasets and models from which they are derived.

The data used and created in this work consists of robot manipulation observations, task instructions, routed evidence labels, and audit records. The VisualEvidence-Set contains no personal conversations, user profiles, human-subject records, or social-media text, and its records do not identify individual people. The dataset contains route-grounded manipulation annotations and audit fields.

I AI Assistant Use

We used AI assistants, including Codex and GPT-5.5, to support manuscript editing, language polishing, and local engineering assistance during paper preparation.

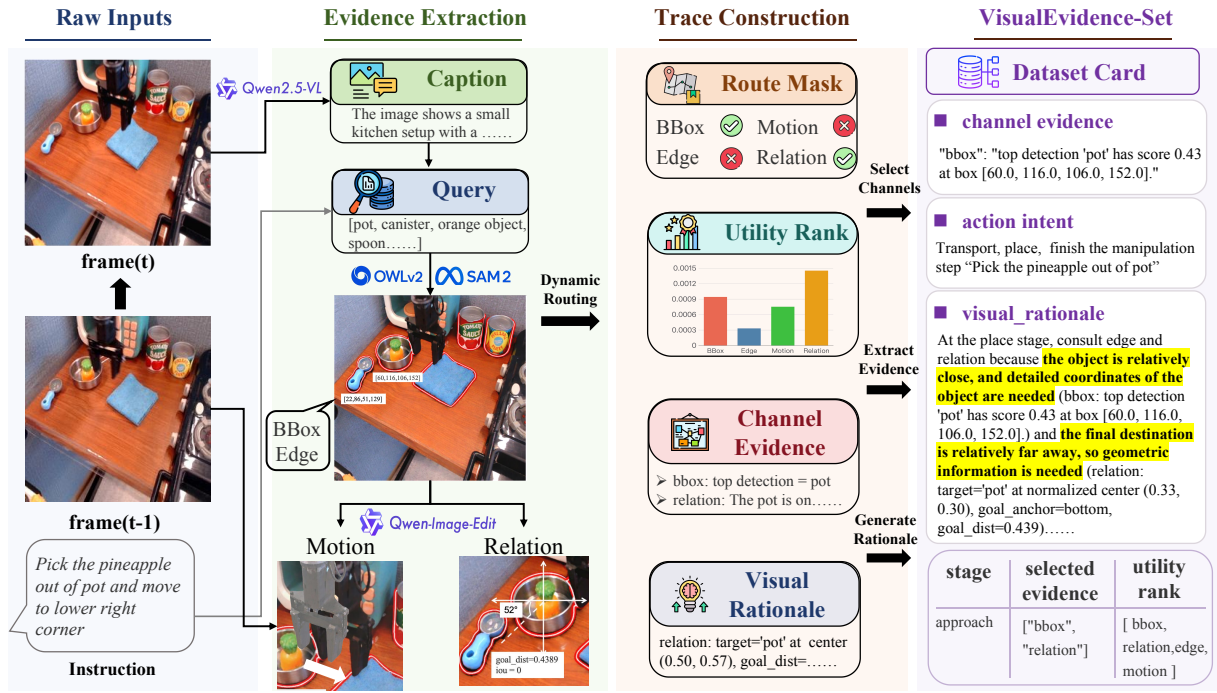


Figure 9: VisualEvidence-Kit workflow and VisualEvidence-Set record schema.

Source	Raw traces	Full-Clean	HQ-Trace	Gold-Faithfulness	Review/drop
BridgeData V2	210,622	203,948	188,736	146,680	15,212
Fractal	89,675	86,104	78,442	52,318	7,662
LIBERO family	273,465	271,982	267,541	253,806	4,441
RoboTurk	7,127	6,842	6,091	4,376	751
UT Austin MUTEX	361,883	354,726	336,918	297,462	17,808
VIOLA	64	61	55	48	6

Table 11: VisualEvidence-Set governance readout. Full-Clean is used for broad statistics and weighted training, HQ-Trace for trace-supervised refinement, and Gold-Faithfulness for high-confidence audit experiments.

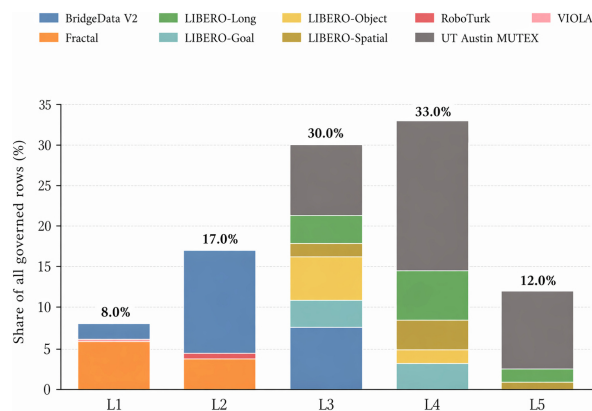


Figure 10: VisualEvidence-Set governance pipeline for sorting raw traces into progressively stricter subsets.

Task family	Representative instructions
Multi-object pick-place	Pick the block from the cluttered area and place it into the bowl.
Relation-sensitive placement	Move the block to the left side of the cup and stop there.
Contact-sensitive reorientation	Flip the mug upright and release it once the base is stable.
Two-stage compositional manipulation	Move the block to the drawer area, then open the drawer and finish placement.

Table 12: Real-robot task definitions for the tabletop deployment study. Each task family lists two representative instructions; the detailed stressors and success criteria are described in the surrounding text.

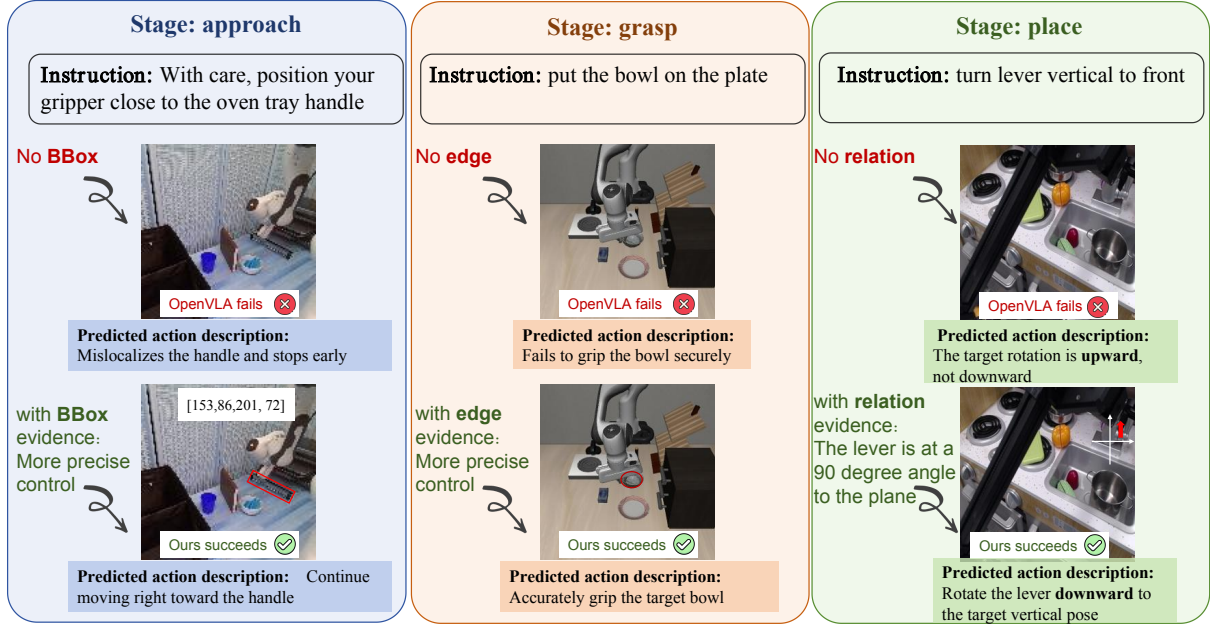


Figure 11: Qualitative case studies of routed visual evidence in representative manipulation tasks.

Method	Trace	Route	Util.	Sel. ch.
Prompt-text evidence	0.391	✗	0.768	4.00
Heavy dense perception	0.441	✓	0.755	2.24
FULLSOFT	0.791	✓	0.961	4.00
VISUALTHINK-VLA	0.840	✓	0.984	2.22

Table 13: VisualEvidence-Kit supervision and audit summary. Here, *Trace* denotes the channel-grounded trace alignment score, *Route* indicates whether explicit route information is exposed, *Util.* denotes utility overlap, and *Sel. ch.* denotes the average number of selected channels.