ViPFormer: Visual-Physics Reasoning for Safe Depalletizing via Segmentation-Guided Transformer

Shan Huang[1,†], Yutian Zhang[2,†], Man Him Cheng[1], Guanqiu Guo[1], Lirong Che[1],
Tian Gao[1], Junbo Tan[1], Xueqian Wang[1,*]
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Shenzhen Graduate School, Peking University, Shenzhen, China
[†]These authors contributed equally to this work. [*]Corresponding author.
Emails: {huang-s25, zheng-wq25, ggq25, clr24}@mails.tsinghua.edu.cn,
yutianzhang25@stu.pku.edu.cn, 22S061022@stu.hit.edu.cn,
{tjblql, wang.xq}@sz.tsinghua.edu.cn

***Abstract*—Automated depalletizing in unstructured warehousing environments faces a critical safety challenge: removing a key load-bearing box can trigger catastrophic stack collapse. Existing approaches primarily rely on geometric perception or static stability analysis, remaining "physics-blind" to the latent dynamic consequences of robotic interactions. To bridge this gap, we propose ViPFormer (Visual-Physics Reasoning Transformer), a novel end-to-end reasoning model that transitions depalletizing from passive perception to predictive physical reasoning. ViPFormer explicitly localizes potential instability risks and predicts the safety of the remaining stack after object removal, utilizing a segmentation-guided mechanism for precise spatial reasoning. To circumvent the high cost and risks of collecting real-world data with physical consequence labels, we introduce SafeDepal-200k, a large-scale synthetic benchmark featuring diverse stacking configurations and automatic counterfactual annotations. Extensive experiments demonstrate that ViPFormer learns generalized physical intuition, achieving an 89.58% F1-score on the simulation benchmark. Despite the domain gap, our method achieves a Sim-to-Real transfer with a 77.62% F1-score on a physical robot without fine-tuning, proving effective as a safety shield for autonomous logistics. Code and dataset are publicly available at https://github.com/YellowThree-HS/ViPFormer-Depalletizing.**

***Index Terms*—Robotic Depalletizing, Visual-Physics Reasoning, Consequence Prediction, Sim-to-Real Transfer, Physics-Aware Manipulation, Synthetic Benchmarks**

I. Introduction

With the exponential growth of e-commerce, Automated Depalletizing has become a pivotal component in modern logistics [1], [2]. Unlike structured assembly lines, contemporary warehouses operate with heterogeneous pallets, where non-uniform boxes are stacked in unstructured configurations [3]. In such chaotic scenarios, the primary challenge transcends efficiency; it lies in ensuring *Operational Safety*.

Removing a critical load-bearing box can trigger avalanches, causing damaged goods and threatening worker safety. Therefore, endowing robots with the capability to foresee physical stability is a fundamental prerequisite for autonomy [4]–[8]. However, existing approaches largely overlook this critical physical dimension.

As illustrated in Fig. 1, existing strategies rely primarily on geometric perception [9]–[12]. Early heuristics typically assess only graspability, disregarding the remaining stack's stability [13]. While recent learning-based approaches introduce static analysis, they only judge *current* stability, offering no guidance on the *future* consequences of an action [5], [14]. Crucially, both paradigms remain "Physics-Blind," lacking the intuition to anticipate latent risks [15], [16], often leading to collapse despite geometrically perfect grasps.

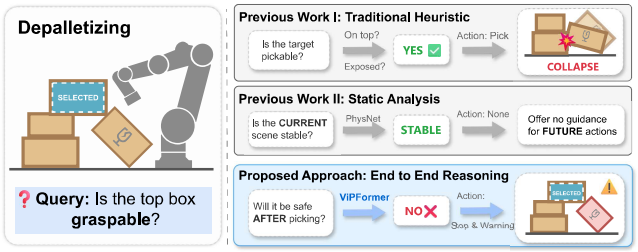

Fig. 1. Comparison between traditional geometric policies and our physics-aware approach. **Top:** Traditional heuristic methods only consider graspability, ignoring the remaining stack. **Middle:** Static analysis methods identify current stability but fail to predict future consequences. **Bottom:** Our **ViPFormer** possesses prescriptive physical reasoning. By predicting the consequence of removing a specific box , it acts as a safety shield, preventing the robot from executing dangerous actions that would destabilize the stack.

To bridge this gap, we propose **ViPFormer**, a novel Visual-Physics Reasoning Transformer designed to transition depalletizing from passive perception to predictive reasoning. Distinct from prior works, ViPFormer is the first method dedicated to Consequence Prediction in this domain. We construct an end-to-end pipeline with a Segmentation-Guided mechanism, leveraging global attention [17], [18] to predict stability and localize *Affected Areas*. As shown in Fig. 1, our model answers: "*If I remove this box, will the stack remain safe?*" enabling the robot to filter out dangerous actions.

To address the scarcity of real-world risk data, we introduce **SafeDepal-200k**, a large-scale benchmark collected in NVIDIA Isaac Sim [19]. It contains diverse stacking configurations with counterfactual labels, designed to enforce physical rule learning over texture memorization. Trained on this benchmark, ViPFormer demonstrates robust Sim-to-Real transfer without fine-tuning [20]. In physical experiments, it effectively functions as a safety shield, significantly outper-

forming geometric baselines, including those utilizing depth sensors, by successfully avoiding collapse-inducing actions.

The main contributions of this work are summarized as follows:

- **ViPFormer Model:** We propose the first visual-physics reasoning model tailored for depalletizing. By leveraging a segmentation-guided mechanism, the model transitions robotic perception from passive observation to predictive reasoning, explicitly localizing pixel-wise instability risks to prevent catastrophic collapse.
- **SafeDepal-200k Benchmark:** To address the scarcity of real-world risk data, we construct a large-scale synthetic benchmark in Isaac Sim. It provides diverse stacking configurations with automatically generated counterfactual labels, specifically designed to enforce physical rule learning over texture memorization.
- **Sim-to-Real Deployment:** Extensive experiments demonstrate the model's generalized physical intuition and robustness. ViPFormer achieves an **89.58% F1-score** on the simulation benchmark and maintains a **77.62% F1-score** in zero-shot real-world transfer, proving its practical value in autonomous logistics.

II. RELATED WORK

A. Geometric Grasping Approaches

Robotic depalletizing has traditionally been treated as a geometric matching problem. Classical analytical approaches rely on explicit 3D object models and force closure principles to determine stable grasp configurations [21], [22]. With the advent of deep learning, data-driven approaches have achieved significant success in unstructured environments. Prominent frameworks, such as Dex-Net [9] and GraspNet [10], utilize Convolutional Neural Networks (CNN) to predict grasp quality directly from depth images or point clouds, focusing on local surface geometry and antipodal constraints [11], [13], [23].

However, these methods operate in a physics-agnostic manner. They primarily assess whether an object is reachable and graspable, implicitly assuming that geometric accessibility equates to operational safety. This assumption is often invalid in heterogeneous stacking scenarios, where removing a key load-bearing box—even one with a perfect grasp score—can trigger a gravity-induced collapse.

B. Intuitive Physics and Visual Foresight

To address the blindness of geometric methods, the domain of Intuitive Physics has explored learning physical laws from visual observations [5], [6], [24], [25].

Early works, such as PhysNet [26] and others [27], [28], train deep networks to predict the stability of block towers from static images. However, as shown in Fig. 1 (Middle), these methods perform Static Analysis. They are limited to answering "*Is the current scene stable?*" but offer no guidance on "*What happens after an action?*". Furthermore, they typically produce a single binary label for the entire scene, lacking the spatial granularity to inform the robot *which* specific object is the hazard.

TABLE I
COMPARISON OF REASONING CAPABILITIES BETWEEN EXISTING DEPALLETIZING METHODOLOGIES AND OUR APPROACH.

Methodology	Physics	Action?	Output	Granularity
Geometric Grasping [9]–[11]	None	Yes	Grasp Quality	Local (Point)
Static Analysis [26]–[28]	Static	No	Stability Label	Global (Image)
Visual Foresight [8], [29]	Dynamic	Yes	Future Video	Pixel (All)
ViPFormer (Ours)	**Dynamic**	**Yes**	**Safety Map**	**Pixel (Mask)**

A parallel line of research attempts to predict future video frames conditioned on robot actions [8], [29]. While theoretically generalizable, pixel-level video generation is computationally intensive and prone to accumulating blurring artifacts over long horizons [30]. This makes it impractical for high-speed industrial depalletizing where precise, real-time decision-making is required [31].

As summarized in Table I, there remains a distinct cognitive gap between coarse *static classification* and computationally heavy *video prediction*. Current methodologies lack the ability to explicitly reason about the *future stability* of a stack post-interaction in a spatially precise and action-conditioned manner. To achieve safe depalletizing in unstructured environments, a system is required that can not only predict *if* a collapse will occur but also segment *where* the instability originates, bridging the gap between passive perception and active physical understanding.

III. METHODOLOGY

As illustrated in Fig. 2, our architecture processes visual observations to simultaneously predict global stack stability and localize potential instability risks.

A. Overview

The proposed model operates on a dual-task paradigm: *Global Stability Prediction* and *Affected Mask Prediction*. By leveraging pixel-level segmentation cues, the model focuses on physically critical regions. The system takes an RGB image and a target mask as primary inputs, with an optional auxiliary depth stream in the geometry-aware variant.

B. Visual Encoding Stream

To handle the multi-modal nature of the input and investigate the impact of geometric cues, we instantiate our framework in two variants: the **ViPFormer-RGB** and the geometry-aware **ViPFormer-RGBD**. We employ a parallel stream architecture based on ResNet-50 backbones [32] to process these modalities.

1) RGB-Mask Stream: The primary stream processes the concatenation of the RGB image $I_{\text{rgb}} \in \mathbb{R}^{3\times H\times W}$ and the binary target mask $M \in \mathbb{R}^{1\times H\times W}$, forming the input tensor $X_{\text{rgb}} \in \mathbb{R}^{4\times H\times W}$.

To accommodate the 4-channel input, we modify the first convolutional layer of the ResNet-50 backbone. The backbone extracts high-level feature maps:

$$F_{\text{rgb}} = \text{ResNet}(X_{\text{rgb}}), \quad F_{\text{rgb}} \in \mathbb{R}^{C_{\text{res}}\times H'\times W'} \tag{1}$$

where $C_{\text{res}} = 1024$, and the feature map spatial dimensions are $H' = H/16$ and $W' = W/16$.

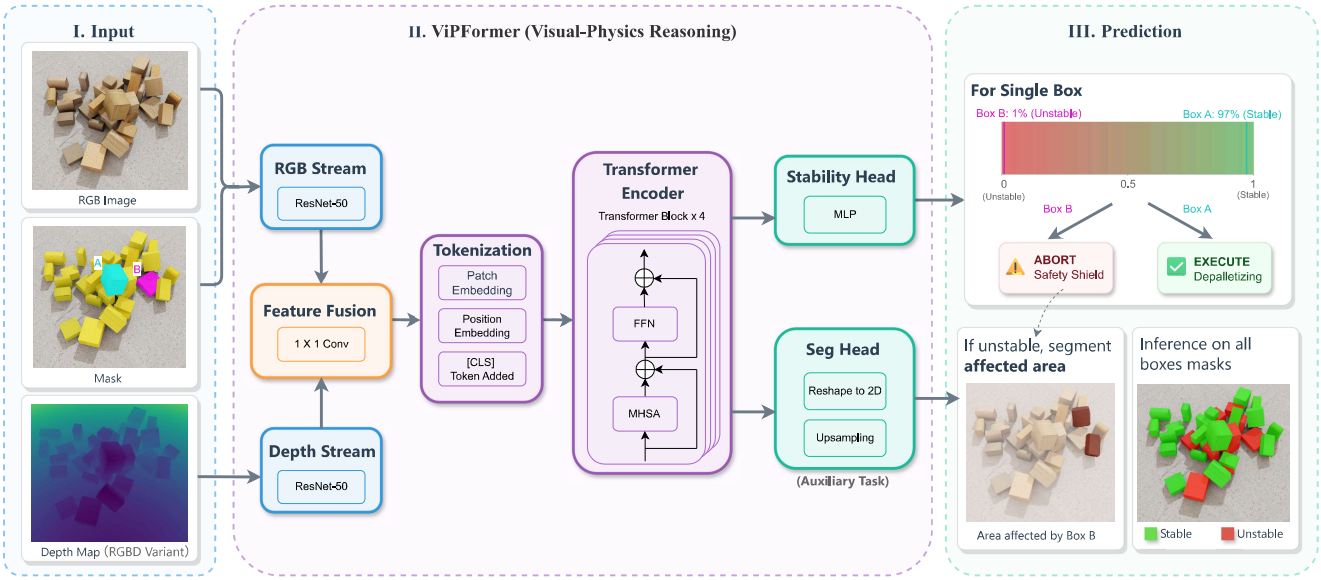

Fig. 2. **Overview of the ViPFormer Architecture.** The framework primarily takes an RGB image and a query mask (target box) as input. In the RGB-D variant, a depth map is additionally incorporated. A ResNet backbone (configured as a dual-stream architecture with feature fusion for the RGB-D variant) extracts visual features, which are then tokenized for a Transformer Encoder to capture long-range physical dependencies between objects. The model branches into two outputs: a *Stability Head* that predicts the probability of a safe removal, and a *Segmentation Head* (an auxiliary task) that localizes the specific area affected by a potential collapse.

2) Depth Stream and Multi-Modal Fusion (RGBD Variant): To enhance physical reasoning with explicit geometric information, the ViPFormer-RGBD variant introduces an auxiliary depth stream. The depth map $D \in \mathbb{R}^{1\times H\times W}$ is processed by a separate ResNet-50 encoder tailored for single-channel input.

The features from both streams are fused via concatenation followed by a projection block ψ (consisting of a 1×1 convolution, Batch Normalization, and ReLU) to reduce dimensions and harmonize features. Formally, the visual representation F_{vis} is defined as:

$$F_{\text{vis}} = \begin{cases} \psi([F_{\text{rgb}}, F_{\text{depth}}]) & \text{for ViPFormer-RGBD} \\ \psi(F_{\text{rgb}}) & \text{for ViPFormer-RGB} \end{cases} \tag{2}$$

The final visual representation F_{vis} has dimensions (D_{emb}, H', W'), serving as the input for the Transformer.

C. Transformer Reasoning Body

The core reasoning module is a Transformer Encoder [17] designed to capture long-range physical dependencies between objects.

1) Tokenization: The feature map F_{vis} is flattened into a sequence of patch tokens $Z_{\text{patch}} \in \mathbb{R}^{L\times D_{\text{emb}}}$, where $L = H'\times W'$. A learnable classification token (`[CLS]`) is prepended to aggregate global context for stability prediction. To retain spatial structural information essential for the segmentation task, learnable position embeddings E_{pos} are added:

$$Z_0 = [z_{\text{cls}}; z^1_{\text{patch}}; z^2_{\text{patch}}; \ldots; z^L_{\text{patch}}] + E_{\text{pos}} \tag{3}$$

2) Encoder Blocks: The sequence Z_0 is processed by a stack of $N = 4$ Transformer blocks. Each block consists of Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN), with Layer Normalization (LN) and residual connections. The self-attention mechanism enables the model to reason about the causal relationships between the masked target object and other stack components.

D. Segmentation-Guided Prediction Heads

The output of the Transformer encoder, Z_N, is processed by two specialized heads.

1) Global Stability Head: This head predicts the probability of stack collapse. We extract the `[CLS]` token z^0_N from the final layer, which serves as a global summary of the scene's physical state. It is passed through a Multi-Layer Perceptron with a sigmoid activation:

$$P_{\text{stable}} = \sigma(\text{MLP}(z^0_N)) \tag{4}$$

where $P_{\text{stable}} \in [0, 1]$ indicates the confidence score (0: Unstable, 1: Stable).

2) Affected Mask Prediction Head: This auxiliary head localizes the regions affected by potential instability. We take the sequence of patch tokens $\{z^1_N, \ldots, z^L_N\}$ excluding the classification token.

The patch tokens are reshaped into a 2D feature map $\mathcal{R}(Z^{1:L}_N)$ and processed by a lightweight upsampling decoder $\mathcal{D}$ to project features back to the pixel space:

$$M_{\text{affected}} = \mathcal{D}(\mathcal{R}(Z^{1:L}_N)) \tag{5}$$

where $\mathcal{R}(\cdot)$ denotes the reshaping operation from sequence to grid form (D_{emb}, H', W').

E. Objective Function

The network is trained end-to-end using a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{aux}}\mathcal{L}_{\text{seg}} \tag{6}$$

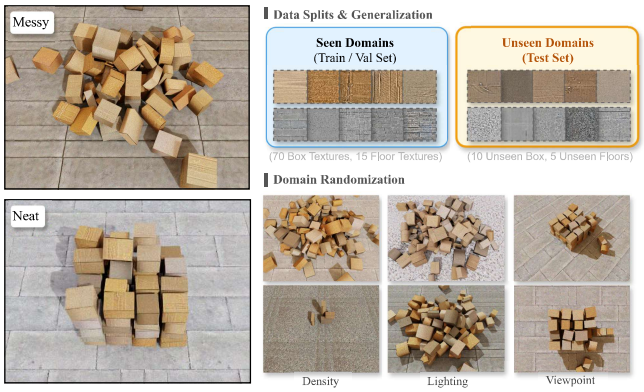

Fig. 3. **Overview of the SafeDepal-200k Benchmark.** The dataset covers diverse topological configurations, ranging from chaotic "Messy" piles to structured "Neat" stacks. To evaluate zero-shot physical generalization, we enforce a strict split between **Seen Domains** (used for Training and Validation) and **Unseen Domains** (reserved exclusively for Testing). Extensive Domain Randomization ensures robustness against visual and physical variations.

We employ Binary Cross-Entropy (BCE) for the primary stability classification $\mathcal{L}_{\text{cls}}$, and a pixel-wise BCE loss $\mathcal{L}_{\text{seg}}$ for the segmentation task.

To enforce the role of segmentation as a supporting objective, we set the balancing hyperparameter $\lambda_{\text{aux}} = \mathbf{0.8}$. This configuration ensures that the auxiliary task effectively regularizes feature learning by encouraging spatial awareness, while keeping the optimization landscape dominated by the primary stability prediction goal.

IV. The SafeDepal-200k Benchmark

To address the scarcity of datasets coupling visual perception with physical stability reasoning, we introduce **SafeDepal-200k**, a large-scale synthetic benchmark. Unlike existing grasping datasets focusing solely on geometric feasibility [33], [34], SafeDepal-200k provides ground-truth causal labels for the dynamic consequences of object removal (statistics in Table II).

A. Data Generation and Randomization

To bridge the sim-to-real gap, we employ extensive Domain Randomization [20]. We simulate two scene topologies: chaotic, interlocking *Messy* piles, and structured *Neat* stacks. By randomizing physical parameters and visual attributes (Table II), we force the model to learn intrinsic physical rules rather than overfitting to specific object appearances.

B. Simulation Dataset Splits

As illustrated in Fig. 3, we partition the dataset based on seen and unseen textures rather than random splitting.

- **Seen Domains (Train/Val):** Constructed using a large pool of box and floor textures to facilitate representation learning.
- **Unseen Domains (Test):** Constructed using a strictly held-out set of textures. This design ensures that evaluation metrics reflect the model's ability to generalize to novel visual appearances, testing true physical understanding.

TABLE II
Statistics of the SafeDepal-200k Synthetic Benchmark

Property	Specification
Dataset Overview	
Total Samples	200,000+
Scenarios	Messy, Neat
Objects per Scene	1–100
Sensor Configuration	
Viewpoints	9 Fixed RGB-D Cameras
Camera Positions	8 Surrounding + 1 Top-down
Image Resolution	640×480
Modalities	RGB, Depth, Instance Seg.
Physical Randomization	
Box Size (Messy)	5–20 cm per dimension
Box Size (Neat)	$12 \times 10 \times 8$ cm $\pm$ 2.5 cm
Box Mass	0.3–2.0 kg
Friction Coeff.	0.2–1.0
Visual Randomization	
Box Textures	80 (70 Train / 10 Test)
Floor Textures	20 (15 Train / 5 Test)
Lighting Intensity	300–1200
Color Temperature	4000–8000 K

C. Automatic Counterfactual Labeling

Obtaining ground truth for stability is non-trivial in the real world. We leverage simulation to perform Counterfactual Labeling [35]. For every scene, we determine the stability of each object by virtually removing it and simulating the resulting dynamics. If the removal causes significant displacement in the remaining structure, the object is labeled as Unsafe. This process generates precise ground-truth labels and segmentation masks without manual intervention.

D. Real-World Evaluation Set

To rigorously evaluate the Sim-to-Real transfer capability of our model, we collected a dedicated Real-World Benchmark.

- Acquisition Setup: Data was collected using an Intel RealSense D435i RGB-D camera. We constructed 62 diverse stacking scenarios, ranging from structured stacks to precarious, unstructured piles. In total, the dataset contains 737 box-removal pairs, serving as a challenging testbed for real-world deployment.
- Expert Annotation Protocol: Unlike simulation, real-world data lacks automatic causal labels. To ensure high-quality ground truth, we employed a rigorous Multi-Expert Arbitration Protocol. Two human experts independently assessed the stability of each object based on physical principles. In cases of disagreement, a third expert adjudicated to reach a consensus.

E. Evaluation Metrics

To align with the high-stakes nature of robotic safety, we propose metrics that prioritize risk avoidance over simple accuracy.

- Metric Selection Rationale: Given the asymmetric cost of errors—where a *False Safe* prediction leads to physical collapse while a *False Dangerous* merely lowers efficiency—we prioritize Precision on the "Safe" class. We formulate this metric to explicitly highlight the penalty on dangerous errors:

$$\text{Precision}_{\text{safe}} = \frac{N_{\text{True Safe}}}{N_{\text{True Safe}} + \underbrace{N_{\text{False Safe}}}_{\text{Risk Source}}} \tag{7}$$

  Minimizing the risk term ($N_{\text{False Safe}}$) in the denominator ensures that the robot only acts when high safety is guaranteed.
- Affected Area Segmentation: In simulation, we quantify fine-grained spatial reasoning using mIoU and Boundary F-measure on instability masks. For the real-world benchmark, where pixel-perfect ground truth for dynamics is unavailable, we report standard classification metrics (Safety Precision, Accuracy, F1).

V. EXPERIMENTS

We evaluate ViPFormer against heuristic rules and ablation variants to validate two hypotheses: 1) stability prediction requires learning contact physics beyond geometric heuristics, and 2) segmentation-based reasoning enhances accuracy and Sim-to-Real robustness.

A. Experimental Setup

Consistent with Section IV, we prioritize Precision (Safe) to minimize physical failures.

1) Datasets and Protocols: We utilize the SafeDepal-200k benchmark and the newly collected Real-World Evaluation Set introduced in Section IV. The evaluation is conducted on two distinct test sets:

- Simulation Test Set (Unseen Textures): Consists of 52,335 scenarios with entirely novel textures (box and floor) never seen during training. This serves as a critical test for physical rule learning versus texture memorization.
- Real-World Test Set: Comprises 737 image-mask pairs collected from 62 physical stacking scenarios, annotated via the multi-expert arbitration protocol. This evaluates the model's robustness to the Sim-to-Real gap.

2) Robotic System Configuration: To validate real-world applicability, we deployed the models on a physical depalletizing platform (see Fig. 4a). The setup comprises:

- Manipulator: A Universal Robots UR5e robotic arm equipped with a pneumatic suction cup.
- Perception: An Intel RealSense D435i RGB-D camera mounted overhead to capture the workspace.
- Compute: All inference is performed on a workstation with a single NVIDIA RTX 4090 GPU.

3) Implementation Details: We implemented our model using PyTorch and conducted all experiments on a high-performance computing cluster equipped with 8 NVIDIA H100 GPUs. The network is trained end-to-end using the AdamW optimizer with a cosine annealing learning rate scheduler. The RGB backbone is initialized with ImageNet pre-trained weights.

B. Baselines and Model Variants

We design a comparative study involving a heuristic baseline and four deep learning variants to isolate the impact of our contributions.

1) Heuristic Rule Baseline: To determine if stability can be solved by simple physical intuition without deep learning, we designed a rule-based baseline grounded in the hypothesis that *objects with high exposure on the top layer are likely safe*. The rule is defined as:

$$\text{Pred} = \begin{cases} \text{Safe} & \text{if } V_{ratio} \geq \tau \text{ AND } z \geq \text{median}(Z_{all}) \\ \text{Unsafe} & \text{otherwise} \end{cases} \tag{8}$$

where V_{ratio} is the visibility ratio of the object, z is its vertical position, and $\tau = 0.6$ is the optimal threshold determined via grid search on the validation set.

2) ViPFormer Ablation Variants: To strictly validate the hypothesis that segmentation aids physical reasoning, we compare the full framework against variants where the auxiliary segmentation head is removed.

1) **ViPFormer (w/o Seg):** The ablation variant consisting of the encoder and the stability classification head only. It relies solely on global semantic features without explicit structural supervision. We test this on both RGB and RGB-D inputs.
2) **ViPFormer (Full):** The proposed framework incorporating the auxiliary segmentation head to enforce boundary-aware feature learning (*segguided*).

C. Quantitative Results

1) Comparison with Heuristic Rules: As shown in Table III, the Heuristic Rule Baseline achieves competitive Precision (77.49%) but significantly lower Recall (55.26%) and Accuracy (63.74%) compared to learning-based methods.

This highlights the inherent complexity of the task: simple geometric statistics (like visibility) cannot capture intricate physical dependencies such as inter-object friction, center-of-mass shifts, and potential chain reactions. Deep learning approaches, by contrast, outperform the rule baseline by over **+20%** in accuracy, proving that end-to-end feature learning is essential for understanding physical stability.

2) Ablation Analysis on Visual-Physics Reasoning: Comparing the variants in Table III, we observe:

The ablation study confirms the critical role of the auxiliary task. Across both modalities, ViPFormer (Full) consistently outperforms the ViPFormer (w/o Seg) counterpart. Specifically, in the RGB-D setting, adding the segmentation head improves Precision (Safe) from 59.59% to **63.67%** in the real

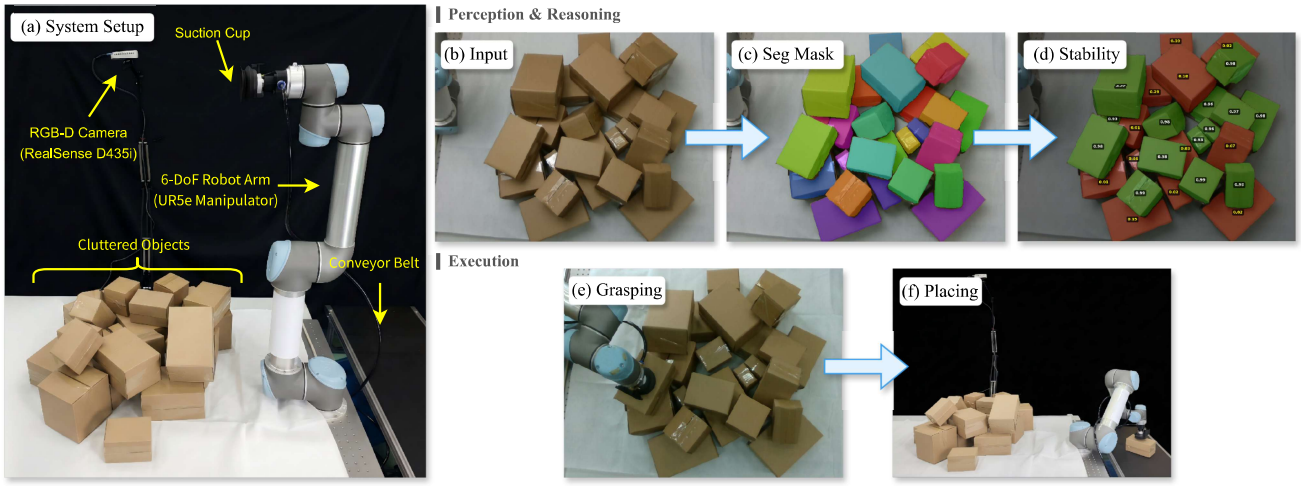

Fig. 4. **Real-World System Deployment and Inference Pipeline.** (a) **System Setup:** The physical platform setup featuring a 6-DoF UR5e manipulator equipped with a suction cup and an Intel RealSense D435i RGB-D camera overlooking cluttered objects. (b) **Input:** The raw RGB image captured by the camera. (c) **Seg Mask:** Instance segmentation masks generated by SAM3 [36]. (d) **Stability:** Stability prediction results where green labels indicate safe, graspable targets and red labels denote unstable ones. (e) **Grasping:** The robot executes a suction grasp on a selected high-stability target. (f) **Placing:** The target is transported and placed onto the conveyor belt.

TABLE III
QUANTITATIVE COMPARISON ON SIMULATION AND REAL-WORLD BENCHMARKS. WE COMPARE THE GEOMETRIC RULE BASELINE AND CONDUCT AN ABLATION ON SEGMENTATION GUIDANCE (W/O SEG VS. FULL). (PREC./REC.: PRECISION/RECALL ON SAFE CLASS).

Method	Input	Simulation Test Set (Unseen Textures)						Real-World Test Set (737 Samples)		
		Prec. (↑)	Rec. (↑)	F1 (↑)	Acc (↑)	mIoU (↑)	Bound-F (↑)	Prec. (↑)	Acc (↑)	F1 (↑)
Rule Baseline	Geometry	0.7749	0.5526	0.6451	0.6374	-	-	-	-	-
ViPFormer (w/o Seg)	RGB	0.8442	0.9187	0.8798	0.8503	-	-	0.6197	0.7327	0.7652
ViPFormer (Full)	RGB	0.8447	0.9245	0.8828	0.8536	**0.5854**	**0.5153**	0.6211	0.7327	0.7635
ViPFormer (w/o Seg)	RGB-D	0.8563	0.9240	0.8888	0.8622	-	-	0.5959	0.7042	0.7459
ViPFormer (Full)	RGB-D	**0.8644**	**0.9296**	**0.8958**	**0.8710**	0.5853	0.4981	**0.6367**	**0.7503**	**0.7762**

world. This indicates that without the explicit constraint to predict object masks, the network (w/o Seg) tends to overfit to noisy depth values, whereas the full model learns robust structural features.

The integration of depth (RGB-D variants) provides a clear performance boost in simulation (Accuracy 87.10% vs. 85.36%). However, even without depth, ViPFormer (RGB) achieves competitive results, demonstrating that our architecture can effectively infer 3D spatial relationships from 2D cues when depth sensors are unavailable.

D. Real-World Deployment and Analysis

1) Sim-to-Real Gap: Real-world results in Table III reveal a susceptibility to sensor noise: *ViPFormer (w/o Seg) (RGB-D)* lags behind its RGB counterpart (Precision: 59.59% vs. 61.97%). This indicates that direct fusion of noisy depth maps causes negative transfer.

In contrast, **ViPFormer (Full) (RGB-D) achieves the highest real-world Precision of 63.67%**, significantly outperforming ViPFormer (w/o Seg) (RGB-D) by **+4.08%**. This suggests that our segmentation-guided mechanism effectively filters sensor noise. By focusing on reconstructing clean instance masks (auxiliary task), the model learns to prioritize structural integrity over raw, noisy depth values, enabling Sim-to-Real transfer.

2) System Demonstration: Fig. 4 visualizes the inference pipeline. Given a cluttered scene (Fig. 4b), the model successfully identifies safe targets (green) while flagging potentially unstable boxes (red). The successful execution of grasping and placement (Fig. 4e-f) confirms that ViPFormer provides sufficiently accurate stability rankings to guide physical interaction in dynamic industrial environments.

VI. CONCLUSION

In this work, we proposed **ViPFormer**, a unified visual-physics reasoning framework that transitions robotic depalletizing from passive perception to active physical understanding. By integrating a novel Segmentation-Guided mechanism, the model explicitly learns the dependency between local object removal and global stack stability, effectively filtering out actions that would trigger collapses.

To enable rigorous evaluation, we introduced the **SafeDepal-200k** benchmark. Experiments on strictly unseen visual domains demonstrate that our model acquires generalized physical intuition rather than merely memorizing texture patterns. Crucially, ViPFormer significantly

outperforms geometric baselines in both simulation and real-world scenarios with robust zero-shot Sim-to-Real transfer capabilities. These results confirm that coupling high-level physical reasoning with low-level visual perception is an effective strategy for ensuring operational safety in unstructured logistics environments.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Shenzhen (No. JCYJ20230807111604008, No. JCYJ20240813112007010), the Natural Science Foundation of Guangdong Province (No. 2024A1515010003) and Cross-disciplinary Fund for Research and Innovation (No. JC2024002) of Tsinghua SIGS.

REFERENCES

[1] J. Aleotti, A. Baldassarri, M. Bonfè, M. Carricato, D. Chiaravalli, R. Di Leva, C. Fantuzzi, S. Farsoni, G. Innero, D. Lodi Rizzini *et al.*, "Toward future automatic warehouses: An autonomous depalletizing system based on mobile manipulation and 3d perception," *Applied Sciences*, vol. 11, no. 13, p. 5959, 2021.

[2] S. Kim, V.-D. Truong, K.-H. Lee, and J. Yoon, "Revolutionizing robotic depalletizing: Ai-enhanced parcel detecting with adaptive 3d machine vision and rgb-d imaging for automated unloading," *Sensors*, vol. 24, no. 5, p. 1473, 2024.

[3] A. Daios and I. Kostavelis, "Mixedpalletboxes dataset: A synthetic benchmark dataset for warehouse applications," *Applied System Innovation*, vol. 9, no. 1, p. 14, 2025.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, 2023.

[5] J. Han, W. Huang, H. Ma, J. Li, J. Tenenbaum, and C. Gan, "Learning physical dynamics with subequivariant graph neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 256–26 268, 2022.

[6] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the national academy of sciences*, vol. 110, no. 45, pp. 18 327–18 332, 2013.

[7] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv:2503.20020*, 2025.

[8] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE international conference on robotics and automation (ICRA)*, 2017.

[9] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv:1703.09312*, 2017.

[10] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[11] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[12] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[13] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[14] Y. Yang, B. Jia, P. Zhi, and S. Huang, "Physcene: Physically interactable 3d scene synthesis for embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[15] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv:2307.05973*, 2023.

[16] Y.-H. Wu, J. Wang, and X. Wang, "Learning generalizable dexterous manipulation from human grasp affordance," in *Conference on Robot Learning*, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[19] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv:2108.10470*, 2021.

[20] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2017.

[21] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[22] A. Guo, H. Jing, and C. Ge, "Design and implementation of robot depalletizing system based on 3d vision," in *Fourth International Conference on Sensors and Information Technology (ICSI)*, 2024.

[23] J. Lundell, F. Verdoja, and V. Kyrki, "Ddgc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6899–6906, 2021.

[24] Y. Li, T. Lin, K. Yi, D. Bear, D. Yamins, J. Wu, J. Tenenbaum, and A. Torralba, "Visual grounding of learned physical models," in *International conference on machine learning*, 2020.

[25] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti, "Visual interaction networks: Learning a physics simulator from video," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *International conference on machine learning*, 2016.

[27] W. Li, A. Leonardis, and M. Fritz, "Visual stability prediction and its application to manipulation." in *AAAI Spring Symposia*, vol. 5, 2017.

[28] O. Groth, F. B. Fuchs, I. Posner, and A. Vedaldi, "Shapestacks: Learning vision-based physical intuition for generalised object stacking," in *Proceedings of the european conference on computer vision (ECCV)*, 2018.

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

[30] P. Agrawal, A. V. Nair, P. Abbeel, J. Malik, and S. Levine, "Learning to poke by poking: Experiential learning of intuitive physics," *Advances in neural information processing systems*, vol. 29, 2016.

[31] M. Janner, S. Levine, W. T. Freeman, J. B. Tenenbaum, C. Finn, and J. Wu, "Reasoning about physical interactions with object-centric models," in *International conference on learning representations*, 2019.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[33] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[34] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[35] C. Glossop, W. Chen, A. Bhorkar, D. Shah, and S. Levine, "Cast: Counterfactual labels improve instruction following in vision-language-action models," *arXiv:2508.13446*, 2025.

[36] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "Sam 3: Segment anything with concepts," *arXiv:2511.16719*, 2025.