GSTVLA: Structured Gaussian Spatial Tokens for 3D Depth-Aware Vision-Language-Action Models

Md Selim Sarowar [1] **Sungho Kim*** [1]

Abstract

VLA models encode visual observations as 2D patch tokens with no intrinsic geometric structure. Dense monocular depth adds pixel-uniform scalar values, but encodes neither surface orientation nor geometric confidence and provides no intermediate spatial verification before action decoding. We introduce GST-VLA with two contributions. First, the Gaussian Spatial Tokenizer (GST) converts frozen depth and semantic patch features into 128 anisotropic 3D Gaussian primitives, each carrying a metric-residual mean, log-scale covariance, and learned opacity. Covariance eigenstructure encodes local surface orientation, opacity provides per-primitive geometric confidence, and learned spatial pooling concentrates the fixed token budget on geometrically salient regions. Second, Depth-Aware Chain-of-Thought (DA-CoT) reasoning supervises four structured intermediate spatial thoughts (3D object grounding, grasp contact geometry, pairwise metric distances, and coarse SE(3) waypoints) as explicit generation targets. A cross-attention sublayer at every VLM transformer block provides access to the raw 256-primitive Gaussian field during DA-CoT generation. A 300M-parameter flow-matching action expert decodes 7-DoF delta-action chunks via conditional ODE integration, conditioned on both VLM hidden states and DA-CoT outputs. Trained with flow, CoT, and depth objectives across progressive stages, GST-VLA achieves low stage-1 training and validation losses on LIBERO. We report preliminary offline MSE results, while simulation evaluation, real-world deployment, and ablations remain ongoing.

[1]Department of Electronics Engineering, Yeungnam University, South Korea. Correspondence to: Sungho Kim <sunghokim@yu.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

VLA models (Kim et al., 2024; Li et al., 2024; Qu et al., 2025; Yuan et al., 2025; Zhen et al., 2024) fine-tune large VLMs on demonstration data to produce robot control policies. Visual observations enter these models as 2D patch tokens: a grid of N_p fixed-resolution embeddings capturing local appearance statistics. Each patch token occupies a fixed spatial extent in pixel space irrespective of the underlying scene geometry. No patch token encodes depth, surface normal direction, or geometric confidence. When manipulation requires millimeter scale geometric accuracy, such as edge grasping, peg insertion, or thin object picking, the model must recover 3D structure implicitly within its hidden states, a computation that degrades systematically as task precision increases.

DepthVLA (Yuan et al., 2025) addresses this by adding a depth expert as a third transformer stream in a mixture-of-transformers (MoT) design. The depth stream shares attention layers with the VLM and action expert, enabling action tokens to attend to intermediate depth representations rather than a final depth map. This establishes that explicit geometric signals improve manipulation precision. However, three structural limitations remain. (i) The depth representation is pixel uniform: each token holds a scalar depth value at a fixed pixel location, distributing token budget equally across geometrically relevant and irrelevant regions. (ii) No token encodes surface orientation. Scalar depth at a point provides no information about the local tangent plane; a flat surface and a sharp edge at identical depth produce identical representations. (iii) There is no mechanism for the model to explicitly verify or articulate its 3D scene interpretation before generating actions. The spatial reasoning pathway from depth tokens to action tokens is fully implicit and non-inspectable.

GST-VLA addresses all three limitations. The Gaussian Spatial Tokenizer (GST) replaces the dense scalar depth stream with N_g=128 anisotropic 3D Gaussian primitives. Each primitive $(\mathbf{c}_k, \Sigma_k, \alpha_k)$ is a volumetric spatial token whose axis-aligned covariance $\Sigma_k = \mathrm{diag}(\exp(2\sigma_k))$ encodes surface orientation through its eigenstructure (the minimum-eigenvalue axis approximates the surface normal), whose opacity $\alpha_k \in (0,1)$ encodes geometric confidence (suppress-

ing tokens on specular or textureless surfaces where the depth estimate is unreliable), and whose spatial allocation via learned attention pooling concentrates the fixed token budget on task-relevant geometry. Depth-Aware Chain-of-Thought (DA-CoT) reasoning introduces a supervised intermediate generation stage: the VLM must produce four structured spatial thoughts (3D object centroid, grasp contact geometry, metric spatial relations, SE(3) waypoints) as explicit supervised targets before generating action-conditioning tokens. During this generation, a cross-attention sublayer at every VLM transformer block provides unfiltered access to the full 256-primitive raw Gaussian field, allowing thought level queries into specific geometric regions.

Our contributions are:

- **GST-VLA architecture.** A trainable GST module produces structured 3D Gaussian tokens from frozen depth and visual features via back projection, per-patch parameter estimation with multi-scale opacity gating, 3D Fourier positional encoding, and spatial attention pooling. A two-stage training protocol uses the composite $\mathscr{L}_{\text{flow}} + \mathscr{L}_{\text{CoT}} + \mathscr{L}_{\text{depth}}$ objective.

- **DA-CoT.** A supervised intermediate reasoning stage imposing explicit 3D geometric targets as sequential generation within the VLM.

2. Related Work

2.1. Spatial Representations in VLA Models

DepthVLA (Yuan et al., 2025) introduces a depth expert as a dedicated transformer stream within a MoT architecture, pretrained on metric depth datasets and jointly trained with the VLM and action expert. The action expert attends to intermediate depth features at every transformer layer, not merely a final depth map. SpatialVLA (Qu et al., 2025) injects 3D egocentric position encodings derived from an off-the-shelf depth estimator into patch tokens, modulating the 2D features with scalar position offsets, and introduces adaptive spatial action grids for discrete action tokenization. Both approaches share a structural property: depth enters the model as a pixel-uniform scalar, with one value per spatial location. The GST departs from this by parameterizing each spatial token as a full anisotropic Gaussian primitive with seven learned parameters (three for mean offset, three for log-scale covariance, one for opacity), and compressing N_p=256 raw tokens into N_g=128 via learned spatial attention pooling rather than uniform spatial binning. This produces tokens that encode surface orientation and geometric confidence, and that concentrate representational capacity on task-relevant regions.

2.2. 3D Gaussian Primitives Beyond Rendering

3D Gaussian Splatting (Kerbl et al., 2023) represents scenes as sets of anisotropic Gaussian primitives $\{(\mu_k, \Sigma_k, \alpha_k)\}$ optimized via differentiable rendering for novel view synthesis. GaussTR (Jiang et al., 2025) aligns Gaussian representations with foundation model features for self-supervised 3D understanding. GPSToken (Zhang et al., 2025) uses 2D Gaussian functions for spatially-adaptive image tokenization, decoupling spatial layout from texture for generation. The GST departs from all of these in objective: it does not optimize for rendering, understanding, or image tokenization, but uses the Gaussian primitive as a spatial token format consumed by a VLM for manipulation policy conditioning. The Gaussian parameterization is valued not for its rendering properties but for the geometric information it encodes per token: position, anisotropic extent, and confidence, all calibrated against metric depth through a differentiable rendering auxiliary loss.

2.3. Chain-of-Thought for Embodied Reasoning

Chain-of-thought reasoning (Zhao et al., 2025) supervises intermediate steps in LLMs. ECoT (Zawalski et al., 2024) studies spatial reasoning chains in manipulation settings. CogACT (Li et al., 2024) decouples cognition from action by conditioning a diffusion transformer on a cognition token output by the VLM. HybridVLA (Liu et al., 2025) unifies autoregressive and diffusion action generation within a single LLM but does not introduce structured spatial reasoning targets. None of these works supervise intermediate generation with explicit metric 3D coordinates, grasp contact geometry, or SE(3) waypoints derived from demonstration data. DA-CoT introduces exactly this: four structured thought components, each supervised against offline 3D annotations, each generating tokens that flow through the VLM's autoregressive decoding before action-conditioning tokens are produced.

3. Method

GST-VLA in (Fig. 2) processes an RGB observation $\mathbf{o}_t \in \mathbb{R}^{H\times W\times 3}$ (H=W=224), a language instruction ℓ, and proprioceptive state $\mathbf{s}_t \in \mathbb{R}^7$ through five sequential stages. A frozen visual encoder produces dense semantic patch features $\mathbf{F}_{\text{sem}} \in \mathbb{R}^{256\times 1152}$. A frozen monocular depth estimator produces affine-invariant metric depth $\hat{D} \in \mathbb{R}^{H\times W}$. Both encoders remain frozen throughout all training stages. The two streams are fused exclusively within the GST.

3.1. Gaussian Spatial Tokenizer

The GST is a trainable module that converts the pair $(\mathbf{F}_{\text{sem}}, \hat{D})$ into N_g structured volumetric spatial tokens. Four sequential operations construct the representation.

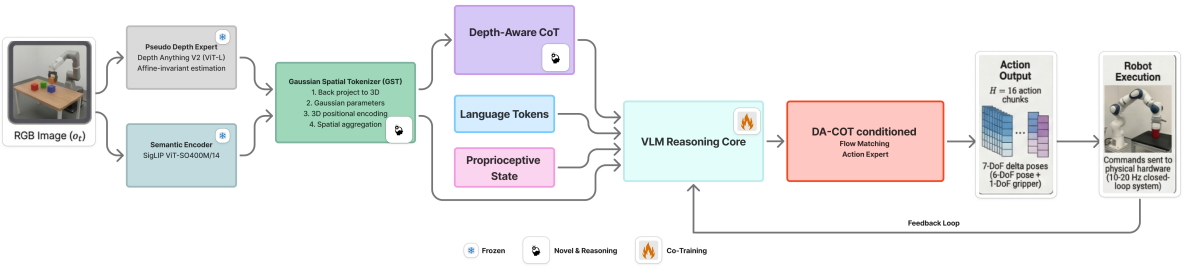

Figure 1. GST-VLA overall framework. Two frozen encoders produce semantic features and metric depth. The GST module lifts these into Ng=128 anisotropic 3D Gaussian tokens via four operations. A cross-attention projector injects spatial tokens into the VLM, where DA-CoT sublayers attend to the raw 256-primitive Gaussian field. The action expert receives dual conditioning: VLM hidden states (semantic and visual context) and DA-CoT action tokens (3D geometric reasoning).

3.1.1. Depth Back Projection to Metric 3D Anchors

Given metric depth $\hat{D}$ and calibrated camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$, each pixel (u, v) is lifted to camera frame 3D via

$$\mathbf{p}_{uv} = \hat{D}_{uv} \cdot \mathbf{K}^{-1}[u, v, 1]^\top \in \mathbb{R}^3. \tag{1}$$

For each of the $N_p = 256$ semantic patches (corresponding to a 16×16 pixel receptive field at input resolution 224×224), we compute the mean of back-projected 3D coordinates within the receptive field, yielding metric anchors $\{\mathbf{p}_k\}_{k=1}^{N_p}$. Each anchor $\mathbf{p}_k$ localizes patch k in the camera-frame 3D coordinate system. The anchor precision is bounded by the depth estimator accuracy (median relative error ∼3% on indoor scenes) and the spatial averaging within the 16×16 receptive field, which smooths sub patch depth variation. The residual offset μ_k estimated in the next step compensates for this averaging artifact.

3.1.2. Spatial Attention Pooling

The $N_p = 256$ raw spatial tokens are compressed to $N_g = 128$ output tokens via single layer cross-attention:

$$\mathbf{Z}_{\text{spatial}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{pool}}\mathbf{K}^\top}{\sqrt{d_g}}\right)\mathbf{V} \in \mathbb{R}^{N_g \times d_g}, \tag{2}$$

where $\mathbf{Q}_{\text{pool}} \in \mathbb{R}^{N_g \times d_g}$ are N_g learned pooling queries, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_p \times d_g}$ are key and value projections of the raw token set $\mathbf{T}_{\text{raw}}$. Each learned query specializes to attend to a specific geometric pattern in the raw token set. Queries that correspond to object surfaces attend to dense clusters of similar depth, high opacity, anisotropic primitives. Queries that correspond to background attend to scattered, high-σ, low-α primitives. The resulting allocation is adaptive: a scene with many small objects distributes pooling queries across multiple object surfaces; a scene with one large object and empty background concentrates queries on the object and assigns few to background.

The full raw token set $\mathbf{T}_{\text{raw}} \in \mathbb{R}^{N_p \times d_g}$ is retained separately and used as the geometric reference for DA-CoT cross-attention (Section 3.2).

3.1.3. Differentiable Depth Rendering as Geometric Regularizer

The predicted Gaussian field must remain geometrically consistent with the metric depth target. We enforce this through a differentiable rendering loss. For camera ray $\mathbf{r}$ with direction $\mathbf{d}$ and origin $\mathbf{o}$, rendered depth is

$$\hat{D}_{\text{render}}(\mathbf{r}) = \sum_{k=1}^{N_p} w_k \cdot (\mathbf{c}_k^\top \mathbf{d}), \tag{3}$$

where $w_k = \alpha_k^{(\mathbf{r})} \prod_{j<k}(1 - \alpha_j^{(\mathbf{r})})$ and the ray-specific opacity is $\alpha_k^{(\mathbf{r})} = \alpha_k \exp\left(-\frac{1}{2}\frac{(t_k - \mathbf{c}_k^\top \mathbf{d})^2}{\mathbf{d}^\top \Sigma_k \mathbf{d}}\right)$ with $t_k = \|\mathbf{c}_k - \mathbf{o}\|$. Equation (3) is fully differentiable with respect to μ_k, σ_k, and α_k, meaning that the depth reconstruction loss $\mathscr{L}_{\text{depth}}$ provides gradients that geometrically calibrate all GST parameters. This loss acts as a geometric regularizer: without it, the MLP could produce Gaussian parameters that are useful for downstream action prediction (via $\mathscr{L}_{\text{flow}}$) but geometrically inconsistent with the actual scene.

3.2. VLM Reasoning Core and DA-CoT

3.2.1. Spatial Token Injection

The pooled spatial tokens $\mathbf{Z}_{\text{spatial}} \in \mathbb{R}^{N_g \times d_g}$ are projected into the VLM hidden dimension d_{VLM} via a two layer cross-attention projector Ψ with residual connection:

$$\tilde{\mathbf{Z}} = \Psi(\mathbf{Z}_{\text{spatial}}) = \text{CA}\big(\mathbf{Z}_{\text{spatial}}, \mathbf{F}_{\text{sem}}\big)\,\mathbf{W}_{\text{proj}}, \tag{4}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_g \times d_{\text{VLM}}}$. The cross-attention in Eq. (4) regrounds each spatial token in the full semantic feature space before VLM ingestion: a spatial token representing a cup handle re-attends to the semantic patch features that encode 'cup' appearance, fusing geometric localization with object identity. The VLM input sequence is $\mathscr{X} = [\tilde{\mathbf{Z}}; \mathbf{L}_{\text{tokens}}; \mathbf{e}_s]$ (spatial tokens, language embeddings, proprioceptive embedding). The VLM is adapted via LoRA (r=16, α=32) on all self-attention projection matrices.

3.3. Flow-Matching Action Expert

The action expert is a 300M-parameter transformer with 6 layers at hidden dimension $d_e = 512$. Each layer contains self-attention, two cross-attention blocks attending to VLM hidden states $\mathbf{H}_{\text{vlm}}$ and DA-CoT action tokens $\mathbf{L}_{\text{action}}$, and an MoE feedforward block with 8 experts, top-2 routing, and hidden dimension 2048. The dual cross-attention conditioning is:

$$\mathbf{H}_{\text{act}}^{(\ell)} = \mathbf{H}_{\text{act}}^{(\ell)} + \text{CA}\left(\mathbf{H}_{\text{act}}^{(\ell)}, \mathbf{H}_{\text{vlm}}\right) + \text{CA}\left(\mathbf{H}_{\text{act}}^{(\ell)}, \mathbf{L}_{\text{action}}\right), \quad (5)$$

where $\mathbf{H}_{\text{act}}^{(0)} = \mathbf{e}_s$ (proprioceptive embedding). The two conditioning streams are functionally distinct: $\mathbf{H}_{\text{vlm}}$ carries the VLM's fused semantic-visual representation, while $\mathbf{L}_{\text{action}}$ carries the 3D geometric reasoning distilled by DA-CoT. This design encourages DA-CoT outputs to provide geometric information that complements the VLM hidden states.

The action distribution is modeled via conditional flow matching (Lipman et al., 2023). Given straight-line interpolation $\mathbf{a}_t = (1-t)\mathbf{a}_0 + t\mathbf{a}_1$ with $\mathbf{a}_0 \sim \mathscr{N}(\mathbf{0}, \mathbf{I})$, the velocity field $v_\theta : \mathbb{R}^{7L_{\text{act}}} \times [0,1] \to \mathbb{R}^{7L_{\text{act}}}$ is trained by:

$$\mathscr{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{a}_0,\mathbf{a}_1}\left[\|v_\theta(\mathbf{a}_t, t \mid \mathbf{H}_{\text{act}}) - (\mathbf{a}_1 - \mathbf{a}_0)\|^2\right]. \quad (6)$$

At inference, the ODE $d\mathbf{a}_t/dt = v_\theta(\mathbf{a}_t, t \mid \mathbf{H}_{\text{act}})$ is integrated from t=0 to t=1 with 10 Euler steps. The model predicts action chunks of $L_{\text{act}} = 10$ future 7-DoF delta poses (6-DoF end-effector plus 1-DoF gripper), with temporal ensemble weighting $\delta_t = 0.01$ s across overlapping chunk predictions.

4. Experiments

4.1. Baseline MSE Comparison

Table 1. First-step offline MSE on 500 validation samples. Lower is better. Baseline evaluations are preliminary and are included as ongoing work.

Method	Action MSE ↓	Traj Pos MSE ↓	N
Spatial-VLA	1.221296	1.247288	500
OpenVLA (baseline)	0.3784	0.3945	500
GST-VLA (Stage 1)	**0.261720**	**0.265201**	500

Protocol note. All values use the same 500 validation samples and first-step prediction protocol. Action MSE is computed over 7 action dimensions. Trajectory Position MSE is computed over the first-step *xyz* displacement. Baseline evaluations are still being finalized and may be updated in the next revision.

4.2. Trajectory MSE Curve

We follow the same 500-sample first-step protocol for the GST-VLA and Spatial-VLA comparison and report the trajectory position MSE with a running-mean curve.

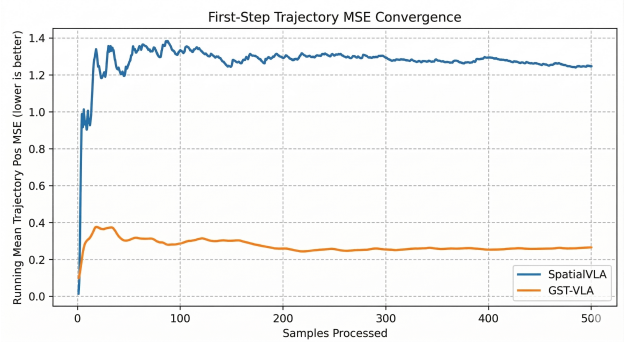

Figure 2. Running-mean trajectory position MSE curve over 500 validation samples (current preliminary curves: GST-VLA and Spatial-VLA).

5. Conclusion

GST-VLA introduces two contributions for 3D-grounded VLA models. The Gaussian Spatial Tokenizer converts frozen depth and visual features into anisotropic 3D Gaussian tokens whose covariance encodes surface orientation and whose opacity encodes geometric confidence, with spatial attention pooling concentrating capacity on task-relevant geometry. Depth-Aware Chain-of-Thought reasoning supervises explicit 3D spatial verbalization as an intermediate generation target, with cross-attention to the raw Gaussian field during each thought. Current offline MSE results provide initial validation; ongoing simulation, real-world, and ablation studies will further quantify component-level contributions on precision-demanding manipulation tasks.

Acknowledgment

This research was supported by the Regional Innovation System & Education(RISE) program through the Gyeongbuk RISE CENTER, funded by the Ministry of Education(MOE) and the Gyeongsangbuk-do, Republic of Korea.(2026-RISE-15-115)

References

Jiang, H. et al. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, pp. 2679–2713. PMLR, 2024.

Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., Wei, F., Deng, Y., Xu, S., Zhang, Y., et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2023.

Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Yuan, T., Liu, Y., Lu, C., Chen, Z., Jiang, T., and Zhao, H. Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. *arXiv preprint arXiv:2510.13375*, 2025.

Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., and Levine, S. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Zhang, Z., Wu, R., Sun, L., and Zhang, L. Gpstoken: Gaussian parameterized spatially-adaptive tokenization for image representation and generation. *arXiv preprint arXiv:2509.01109*, 2025.

Zhao, Q., Lu, Y., Kim, M. J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025.

Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., and Gan, C. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning (ICML)*, pp. 61229–61245. PMLR, 2024.