Introduction (§1) → Preliminary (§2) → Vision-Action Models (§3) → Vision-Language-Action Models (§4) → Datasets and Benchmarks (§5) → Challenges and Future Directions (§6) → Conclusion (§7)

Preliminary (Section 2)

Input Modalities: Sensors, Latent Feature, Language, Vehicle State

Generative Model → Action / Guidance → Action Prediction Model → Trajectory / Control / Language

Vision-Action Models (Section 3)

End-to-End VA
1. Action-Only Models
2. Perception-Action Models

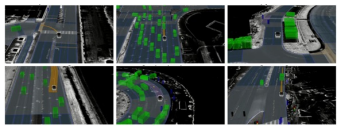
World Models
1. Image-Based World Models
2. Occupancy-Based World Models
3. Latent-Based World Models

Vision-Language-Action Models (Section 4)

End-to-End VLA
1. Textual Action Models
2. Numerical Action Models

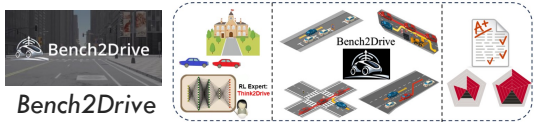
Dual System VLA
1. Explicit Action Guidance
2. Implicit Representations Transfer

Datasets and Benchmarks (Section 5)



Open-Loop

Closed-Loop



Open-Loop & Closed-Loop

Challenges and Future Directions (Section 6)

Challenges:
- Computation Efficiency
- Generalization
- High-Quality Training Data
- Model Interpretability
- Long-Horizon Coherence
- Foundation Model Construction

Future Directions:
- Combined with World Model
- Injecting World Knowledge
- Continual Learning
- Standardized Evaluations
- Human-AI Interaction