Vision-Action Coupling in World-Model Policies

Kun Chu

Knowledge Technology, University of Hamburg

KNOWLEDGE TECHNOLOGY

WTM WM SEMINAR

THE REFRAME

One joint distribution, four queries

$$p(o_{t+1:t+k}, a_{t+1:t+k} \mid o_t, l)$$

the joint predictive-control distribution (Eq. 4)

Policy

marginalize out future observations

$p(a \mid o_t, l)$

Passive world model

marginalize out actions

$p(o' \mid o_t, l)$

Controllable world model

condition on actions

$p(o' \mid o_t, a)$

Inverse dynamics (IDM)

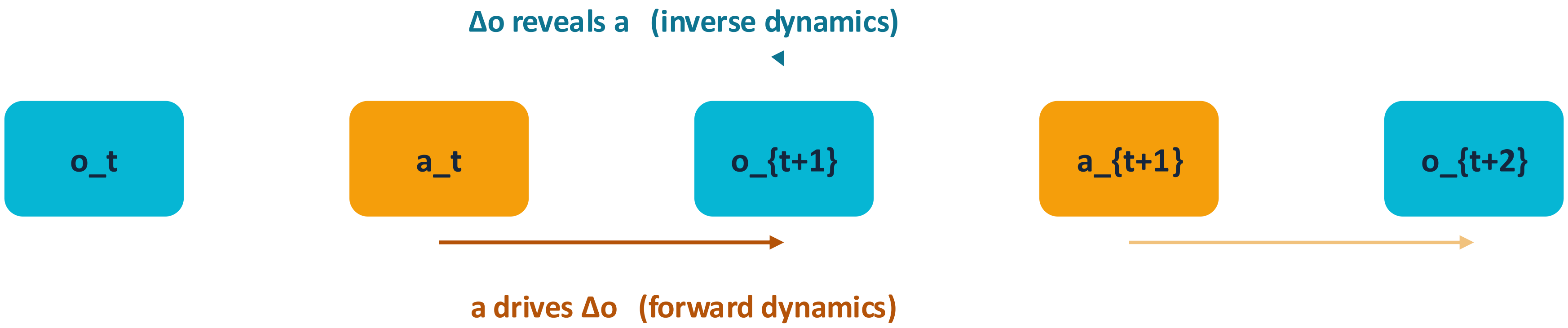
condition on an observation sequence

$p(a \mid o_{t:t+k})$

"World model" and "policy" are not separate — they are factorizations of one density. (Conceptually)

WHY THEY'RE COUPLED

Vision and action are coupled by the data itself



Temporal coupling

Forward: $p(o' \mid o_t, a)$ — controllable WM

Inverse: $p(a \mid o_t, o')$ — IDM

Frequency mismatch: control rate ≠ camera rate → the later "mismatch."

Spatial coupling

Action changes only a **local region** of the scene.

Masked IDM (VidMan, Vidar) targets action-relevant pixels.

3D-flow methods (NovaFlow) read it geometrically.

Three axes for every paradigm

A

Coupled, and where?

Is the future observation coupled with the action inside the model — and where is that link realized?

B

Implicit or explicit?

Is the coupling a built-in training objective, or just an external data flow between modules?

C

How is it built?

Pixel- vs latent-level. Via a conditioning head, shared weights, or cross-attention between streams?

Applied identically to all five paradigms — the axes are the lens, the paradigms are the spectrum.

PARADIGM 1 / 5

IDM-style: decoupled predict-then-act

COUPLING

External — the predicted future is the interface; separate, often-frozen modules.

IMPLICIT / EXPLICIT

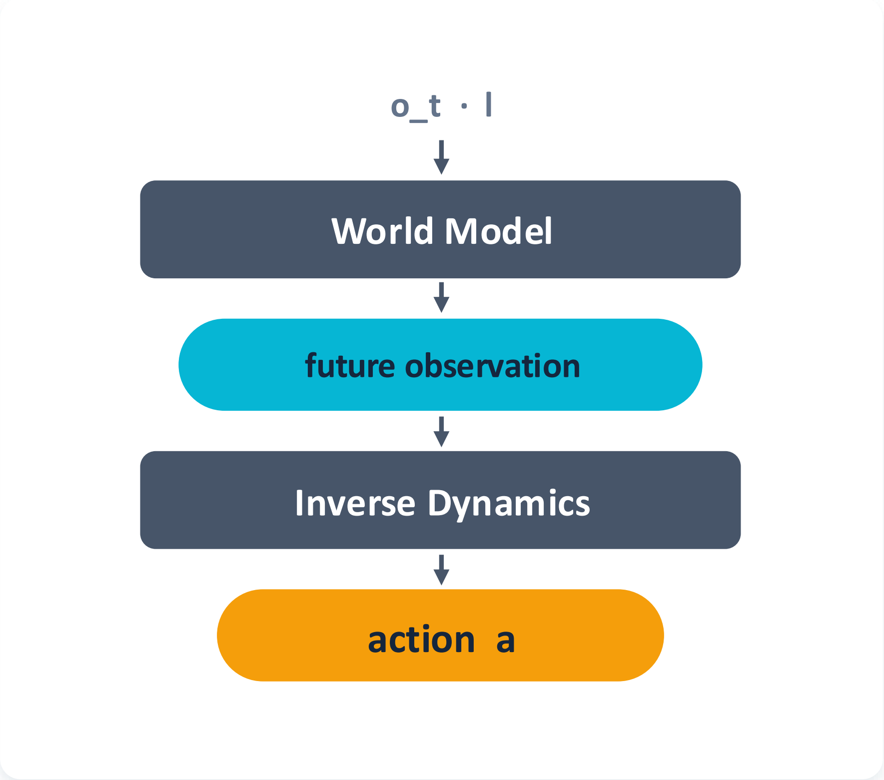
Explicit, but decoupled — no shared weights.

CONSTRUCTION

Condition the action head on Φ(future). Pixel → latent → 3D-structured.

FACTORIZATION · passive WM → IDM

`p(o' | o_t, l) · p(a | o_t, o', l)`



Representative: UniPi · VidMan/Vidar · VPP/Video2Act · TC-IDM · NovaFlow · Say-Dream-ACT

PARADIGM 2 / 5

Single-backbone: joint, full sharing

COUPLING

Shared parameter space — o' and a are co-generated in one pass.

IMPLICIT / EXPLICIT

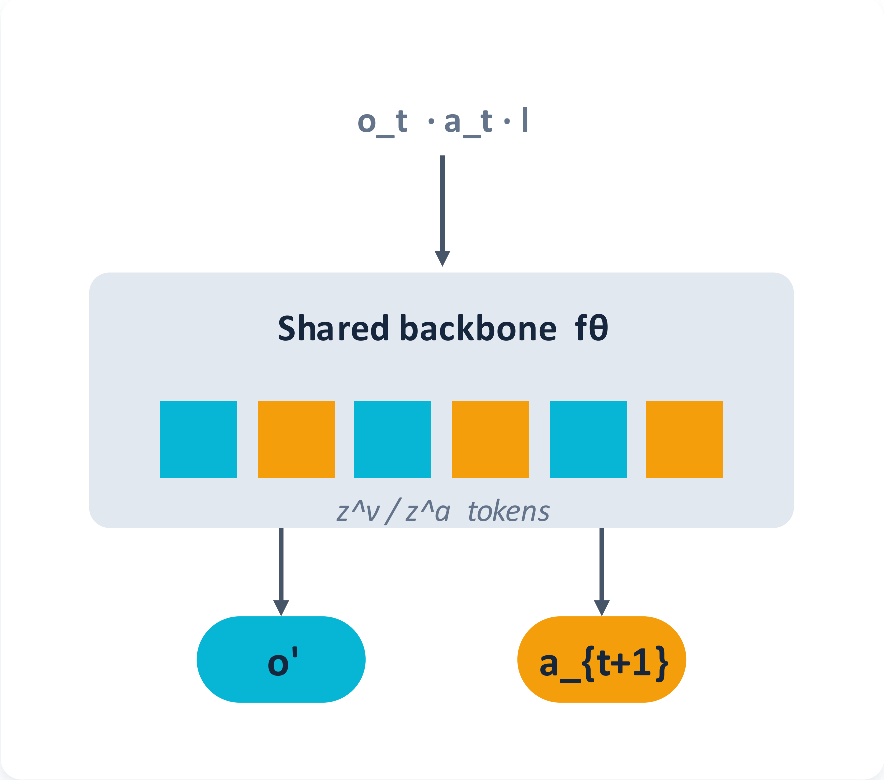
Explicit and joint — a single unified objective.

CONSTRUCTION

Concatenate [z^v ; z^a] → shared denoiser + attention. Mostly pixel/video.

THE MISMATCH

frequency · scale · optimization — managed, not resolved (modality-specific timesteps/heads; visual branch optional at inference).



Representative: UVA · UWA · VideoVLA · Cosmos Policy · DreamZero · GigaWorld-Policy

PARADIGM 3 / 5

MoE / MoT: joint but specialized

COUPLING

Cross-expert — separate video and action experts, coupled by attention.

IMPLICIT / EXPLICIT

Explicit and joint, but specialized — parameters not shared.

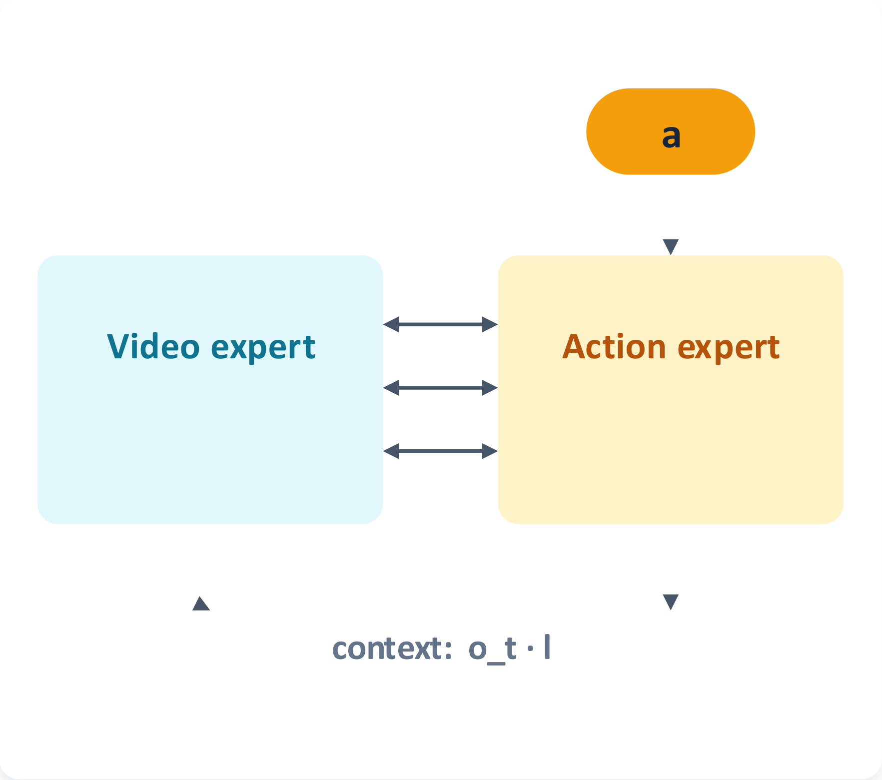
CONSTRUCTION

Cross / shared attention between expert streams. Latent-level.

PROBLEMS → TECHNIQUES

stop video from dominating action; cut rollout cost → single-step denoising, latent forecasting, train-time-only video.

$$(\mathbf{h}^{v}_{\ell+1},\, \mathbf{h}^{a}_{\ell+1}) = \mathcal{F}^{\mathrm{mix}}_{\ell}(\mathbf{h}^{v}_{\ell},\, \mathbf{h}^{a}_{\ell};\, o_t,\, l)$$



Representative: GE-Act · Motus · LingBot-VA · BagelVLA · Fast-WAM · LDA-1B · FRAPPE

PARADIGM 4 / 5

Unified VLA: implicit, internalized

COUPLING

Internalized in a semantic VLM backbone; often only at train time.

IMPLICIT / EXPLICIT

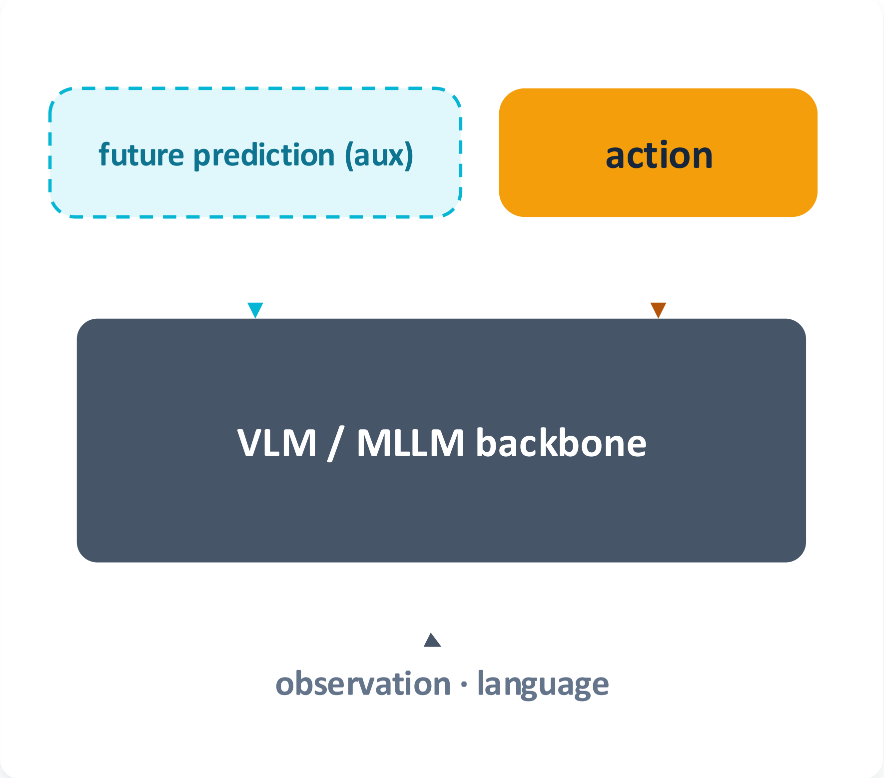
Mixed — explicit future-image OR implicit latent or semantic prediction.

CONSTRUCTION

Auxiliary future-prediction head, co-trained with the action head.

KEY IDEA

the world model acts as a regularizer on the action representation; the future head is usually dropped at inference.



Representative: GR-1 · UP-VLA · WorldVLA · DreamVLA · UniVLA · F1 · HALO · TriVLA

PARADIGM 5 / 5

Latent-space world modeling

COUPLING

Entirely in representation space — the action net's own latent is made predictive.

IMPLICIT / EXPLICIT

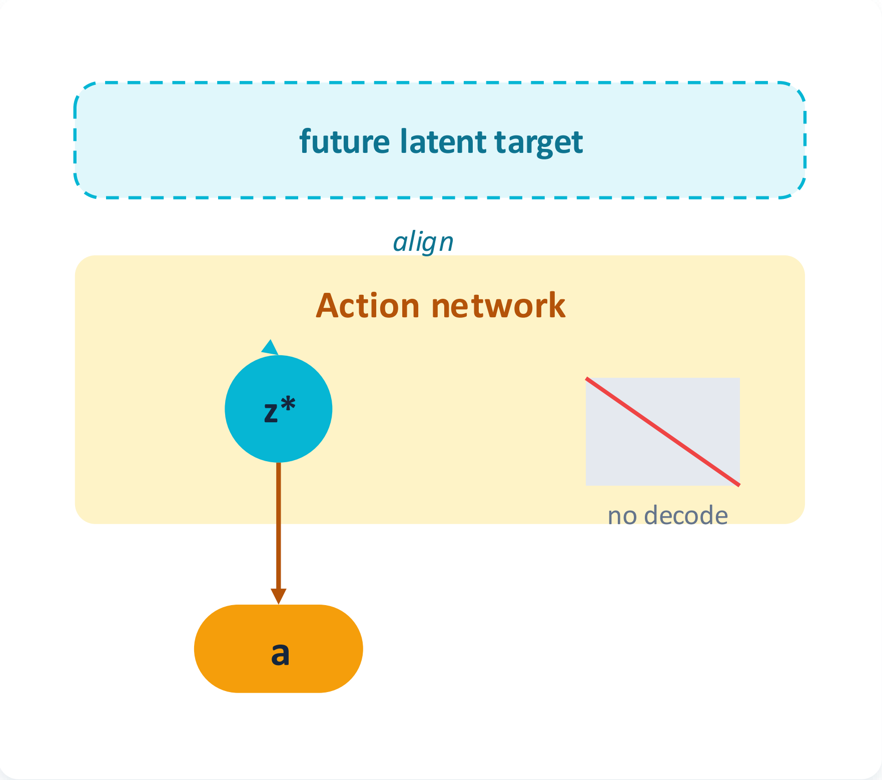
Implicit output (no pixels), but explicitly trained alignment.

CONSTRUCTION

Latent alignment / JEPA target / conditioning head. No decoding.

FACTORIZATION

`p(a | o_t, l, z*)  —  z* predicted, never decoded`



Representative: FLARE · VLA-JEPA · JEPA-VLA · WoG · DIAL · (symbolic: VisualPredicator)

SYNTHESIS

Locus of coupling

external interface → shared weights → cross-expert attention → auxiliary objective → latent alignment

1. IDM-style
2. Single-backbone
3. MoE / MoT
4. Unified VLA
5. Latent-space WM

Representational level

pixel / video ——→ latent features ——→ abstract / symbolic

vision / observation　　action

AT A GLANCE

Five paradigms, five ways to couple

Paradigm	Where coupled	Impl. / Expl.	How (level + mechanism)	Visual @ inference
IDM-style	External — predicted future as interface	Explicit, decoupled	pixel → latent → 3D; condition action head on Φ(future)	Active (unless latent-feature)
Single-backbone	Shared parameter space	Explicit, joint	pixel/video; token concat + shared attention	Often optional / marginalized
MoE / MoT	Cross-expert	Explicit, specialized	latent; cross/shared attention between experts	Reducible (single-step / latent / train-time)
Unified VLA	Internalized (often train-time)	Mixed (image or latent)	auxiliary future head on a VLM trunk	Usually dropped
Latent-space WM	Representation space	Implicit out / trained	latent; alignment / JEPA / conditioning head	None (no generation)

Φ = feature extractor over the predicted future. Strong LIBERO results appear in every row — see next slide.

EVIDENCE

The value isn’t tied to one paradigm

Table 5 Representative results on the LIBERO standard 4-suite benchmark, grouped by how world modeling is integrated with policy learning. Methods with no directly reported number under the standard Spatial/Object/Goal/Long protocol are omitted.

Group	Method	Spatial	Object	Goal	Long	Avg
Decoupled	UniPi (Du et al., 2023)	–	–	–	0.0	–
	MimicVideo (Pai et al., 2025)	94.2	96.8	90.6	94.0	93.9
	Say-Dream-ACT (Gu et al., 2026)	99.4	99.2	98.6	95.4	98.1
Single-backbone	UVA (Li et al., 2025c)	–	–	–	90.0	–
	VideoPolicy (Liang et al., 2025a)	–	–	–	94.0	–
	Cosmos Policy (Kim et al., 2026)	98.1	100.0	98.2	97.6	98.5
	UD-VLA (Chen et al., 2026b)	94.1	95.7	91.2	89.6	92.7
MoE / MoT	Motus (Bi et al., 2025)	96.8	99.8	96.6	97.6	97.7
	LingBot-VA (Li et al., 2026b)	98.5	99.6	97.2	98.5	98.5
Unified VLA	RynnVLA-002 (Cen et al., 2025)	99.0	99.8	96.4	94.4	97.4
	DreamVLA (Zhang et al., 2025e)	97.5	94.0	89.5	89.5	92.6
	UniVLA (Bu et al., 2025b)	96.5	96.8	95.6	92.0	95.2
	Unified VLA (Wang et al., 2025)	95.4	98.8	93.6	94.0	95.5
	CoWVLA (Yang et al., 2026a)	97.2	97.8	94.6	92.8	95.6
	F1 (Lv et al., 2025)	98.2	97.8	95.4	91.3	95.7
	TriVLA (Liu et al., 2025d)	91.2	93.8	89.8	73.2	87.0
Latent-space WM	VLA-JEPA (Sun et al., 2026)	96.2	99.6	97.2	95.8	97.2
	JEPA-VLA (Miao et al., 2026)	97.2	98.0	95.6	94.8	96.4

Methods are grouped by the manner in which world modeling is incorporated into policy learning, rather than by publication year alone.
Avg denotes the average success rate over the four LIBERO suites when directly reported by the original paper. “–” indicates that the corresponding suite-level result was not directly reported under the standard Spatial/Object/Goal/Long protocol.

What it means

Competitive results appear across decoupled, shared, mixture, unified, and latent designs — so photorealistic video generation is not required for effective control.

THE DISCRIMINATOR

Long-horizon, action-faithful consistency — not visual realism. (§8.1, causal conditioning gaps.)

What matters is a future that stays **causally faithful to the pending action** and **stable over long horizons**.

the coupling, kept honest

Symbolic / planner-facing world models (§3.6, §8.5) are the abstract sibling of latent prediction — discrete, rule-based dynamics instead of learned embeddings.

Synthesis of §3 & §8.1 · Hou et al., 2026 — World Model for Robot Learning: A Comprehensive Survey