CROSSPORTRAIT: CROSS-SPECIES PORTRAIT ANIMATION VIA ADVERSARIAL MOTION ALIGNMENT

Anonymous authors
Paper under double-blind review

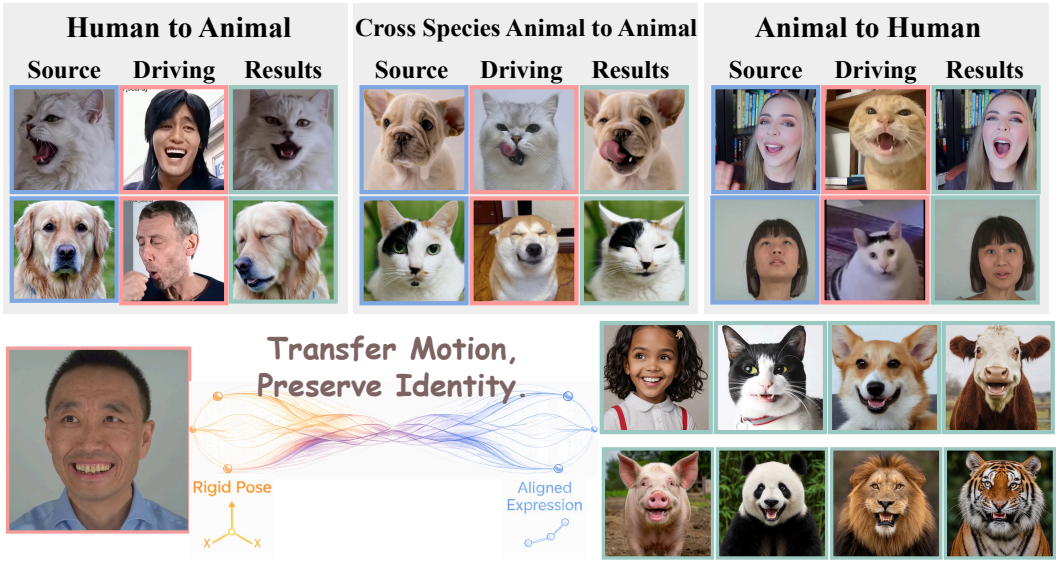

Figure 1: Cross-species animation by CrossPortrait. Our method transfers motion while preserving source identity in human-to-animal, animal-to-animal, and animal-to-human scenarios. **Blue**, **red**, and **green** borders denote **source** portraits, **driving** images, and **CrossPortrait outputs**, respectively.

ABSTRACT

Cross-species portrait animation, which aims to animate a source portrait of *species A* using a driving video of *species B*, is particularly challenging due to substantial anatomical differences in facial structures across species. Previous works typically fine-tune human portrait animation models on animal video datasets but often suffer from driving-species leakage. To address this, we propose **CROSSPORTRAIT**, a novel framework for expressive cross-species portrait animation, supporting diverse scenarios including human-to-animal, animal-to-human and cross-species animal-to-animal (e.g. cat-to-dog) animation. Specifically, we decompose the portrait motion into rigid head poses and expression keypoints. To accommodate species-specific facial structures while facilitating cross-species expression alignment, we introduce a species-routed expression encoder that incorporates species-specific LoRAs into a unified expression encoder. Furthermore, we introduce expression-level adversarial domain adaptation (Ex-ADA) to suppress driving-species-specific information in the expression latents, thereby mitigating species leakage and promoting a shared expression space across species. Extensive experiments show that our method enables expressive and identity-preserving portrait animation across diverse species and outperforms other state-of-the-art approaches.

1 INTRODUCTION

Portrait animation synthesizes a video of a source portrait that follows a driving video. Although AIGC-based methods (Zhao et al., 2025; Wang et al., 2025; Cheng et al., 2025; Shi et al., 2026) have achieved expressive human portrait animation, most of them suffer from generalization to other

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

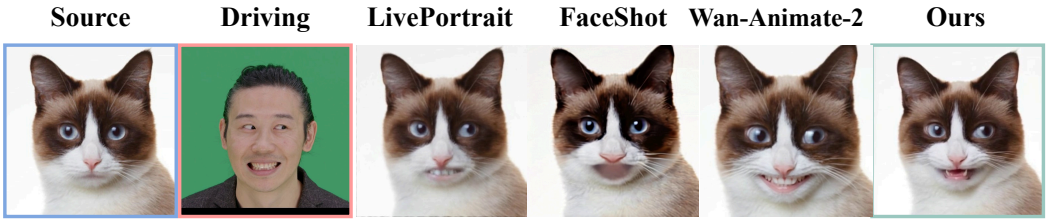

Figure 2: Comparison of species leakage with LivePortrait (Guo et al., 2024), FaceShot (Gao et al., 2025), and Wan-Animate-2 (Wang et al., 2026). These approaches tend to transfer human dental characteristics to the animated cat, whereas our method faithfully follows the driving mouth motion while preserving the source cat's oral morphology. For a fair comparison, we further post-train all applicable baselines on the same animal training data.

species, let alone cross-species animation. In this paper, we focus on *cross-species portrait animation*, where the source and driving input may depict different species, such as humans, cats, or dogs, while the output preserves the source identity, appearance, and species-specific morphology. We use $X \rightarrow Y$ to denote a driving clip from species X applied to a source portrait of species Y; for example, H→A denotes a human driving clip applied to an animal source portrait. Specifically, our cross-species animation setting mainly includes human-to-animal (H→A), animal-to-human (A→H), and cross-species animal-to-animal (A1→A2) animation, as shown in Fig. 1.

Portrait motion can be decomposed into head pose and facial expression. Although head poses across different species can be represented by a shared rigid transformation comprising rotation, translation, and scale (RTS), facial expressions vary substantially across species due to their distinct anatomical structures. Therefore, a key challenge lies in learning a shared expression space that aligns semantically corresponding expressions across species, such that the expression of a driving subject can be effectively transferred to a source subject of a different species. However, directly encoding the driving image into an expression feature may entangle expression-related information with species-specific cues, leading to species-information leakage. As illustrated in Fig. 2, such leakage may undesirably transfer characteristics of the driving species to the animated source, thereby degenerating source-species fidelity. Consequently, another key challenge is to effectively suppress species-specific identity cues.

To address these two challenges, we propose an adversarial expression learning framework for cross-species portrait animation. For the first challenge, we employ a shared expression encoder to map facial expressions of humans, cats, and dogs into a common representation space, and leverage LivePortrait (Guo et al., 2024) as a training framework to align expression representations across species. To accommodate species-specific anatomical variations without compromising the shared representation space, we further introduce a species-routed expression encoder, which incorporates species-specific LoRA (Hu et al., 2022) adapters into a unified driving-side expression encoder. The shared backbone facilitates cross-species expression alignment, while the species-specific adapters enhance the representation capacity for capturing fine-grained expression patterns unique to each species. For the second challenge, we introduce expression-level adversarial domain adaptation (Ex-ADA) (Ganin et al., 2016), where a domain classifier is jointly trained to predict the species category from the encoded expression features. Through adversarial training with a gradient reversal layer (GRL), the expression encoder is encouraged to suppress species-specific identity cues, yielding more species-agnostic expression representations and thereby mitigating driving-species leakage. Finally, we reshape the driving expression features into expression tokens and inject them into the video DiT (Team Wan et al., 2025) via cross-attention, enabling expressive cross-species portrait animation while preserving source-species fidelity.

Overall, our contributions are summarized as follows:

- We propose **CrossPortrait**, a unified dual-path framework for cross-species portrait-motion alignment: rigid head motion is represented with dense pose-flow, while facial deformation is represented with shared implicit-keypoint tokens, covering human-to-animal, animal-to-human, and cross-species animal-to-animal animation.
- We introduce **species-routed expression encoder**, which inserts species-specific LoRA adapters into a unified driving-side expression encoder, preserving capacity for anatomical differences while providing a shared expression-conditioning interface.

- We introduce **expression-level adversarial domain adaptation** (Ex-ADA), applied to expression features to suppress driving-species-specific information, mitigate species leakage, and avoid adversarially aligning rigid pose or source morphology.

2 Related Work

Motion representations for portrait animation. Portrait animation methods condition generation on motion representations with varying levels of structure. Parametric models provide interpretable pose and expression coefficients: PIRenderer (Ren et al., 2021), for example, uses coefficients from a 3D morphable model. Self-supervised approaches instead discover sparse keypoints or regions and convert their motion into dense warps (Siarohin et al., 2019; 2021; Zhao & Zhang, 2022). FaceVid2Vid (Wang et al., 2021) decomposes learned 3D keypoints into canonical geometry, rigid pose, and non-rigid deformation, while LivePortrait (Guo et al., 2024) extends this interface with stitching and retargeting. Dense motion can also condition video generators: MOFA-Video (Niu et al., 2024) converts sparse controls into motion fields, while FG-Portrait (Xu et al., 2026) derives 3D motion flow from a parametric head model. Other methods learn motion conditions directly from images or latent features. X-Portrait (Xie et al., 2024) extracts hierarchical features from driving frames, whereas HunyuanPortrait and X-NeMo (Xu et al., 2025; Zhao et al., 2025) inject compact implicit motion representations through attention. DeX-Portrait (Shi et al., 2026) combines explicit global pose with latent expression, and DPE (Pang et al., 2023) studies pose–expression disentanglement. Wan-Animate (Cheng et al., 2025) uses causal convolutions to compress an implicit face representation to the video-latent frame rate. Overall, explicit geometry offers a stable, interpretable interface but is tied to predefined anatomy, whereas learned representations are more flexible but not inherently comparable across domains. We therefore model rigid head motion geometrically and facial deformation with learned implicit-keypoint tokens. Each keypoint trajectory is refined at the original frame rate before the retained tokens are packed to match the VAE timeline.

Portrait animation beyond human faces. Most portrait-animation methods study cross-identity transfer within the human-face domain, including X-Portrait (Xie et al., 2024), EMOPortraits (Drobyshev et al., 2024), Follow-Your-Emoji (Ma et al., 2024; 2026), HunyuanPortrait (Xu et al., 2025), and X-NeMo (Zhao et al., 2025). A smaller body of work broadens the source domain. Animate-X (Tan et al., 2025) animates anthropomorphic characters from pose guidance, FaceShot (Gao et al., 2025) retargets facial landmarks to diverse characters, FantasyPortrait (Wang et al., 2025) reports animal-source examples, and LivePortrait provides an animal model. These works mainly show that human-derived motion can animate broader source appearances, rather than extracting motion from non-human drivers. Animal-driven portrait animation remains less explored. FaceShot includes several non-human driving examples, although its CharacBench evaluation uses human drivers. AnimateZoo (Xu et al., 2024) studies cross-species motion transfer, but focuses on full-body actions and mismatched skeletons rather than portrait-scale head pose and facial deformation. Our work instead considers human, cat, and dog sources and drivers, covering human-to-animal, animal-to-human, same-species animal, and cross-species animal-to-animal animation within one shared generator.

Adversarial domain alignment and domain-specific routing. Domain-adversarial learning reduces domain-predictive information by training a domain classifier through a gradient reversal layer (Ganin et al., 2016). Related approaches separate shared and domain-private features (Bousmalis et al., 2016), condition alignment on task predictions (Long et al., 2018), or use multiple discriminators for finer-grained alignment (Pei et al., 2018). Residual adapters provide domain-specific capacity around a shared model (Rebuffi et al., 2017), while low-rank adaptation offers parameter-efficient routed updates (Hu et al., 2022). We adopt this dual-path interface for cross-species motion conditioning; related video-generation work explores similar alignment and decoupling choices (Zhang et al., 2025; Zheng et al., 2026). At inference, the driving-species label selects a species-routed LoRA adapter in the expression encoder to capture species-specific appearance and anatomy. During Stage 1 training, Ex-ADA is applied to both source- and driving-frame expression features from same-species reconstruction pairs; at inference, only the driving-side expression features are used for motion conditioning. The rigid-pose pathway, source-reference pathway, and shared video generator remain species-agnostic and are not adversarially aligned.

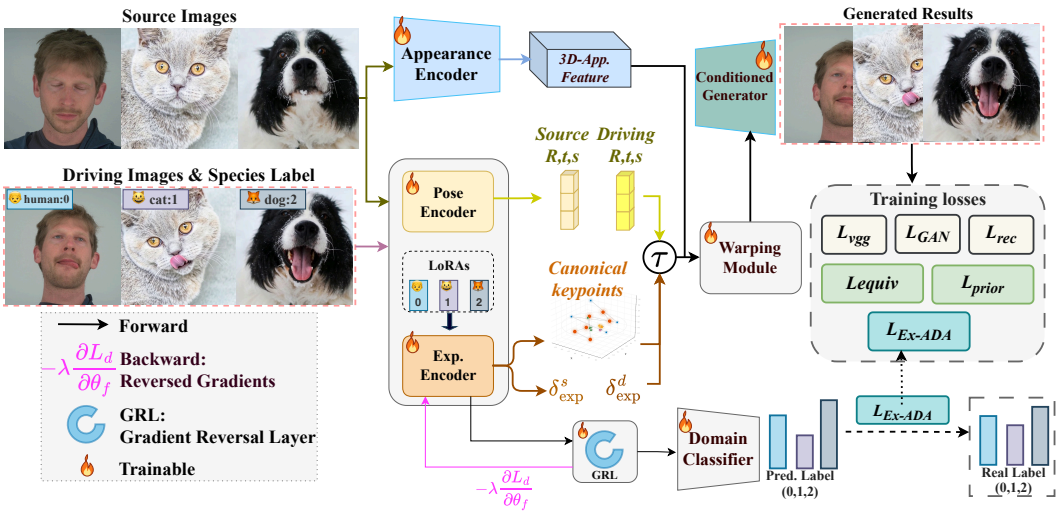

Figure 3: **Stage 1: Learning cross-species motion alignment.** We align expression representations across species while suppressing species-specific identity cues through the species-routed expression encoder and expression-level adversarial domain adaptation (Ex-ADA).

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Given a source portrait I_s and a driving video $V_d = (I_d^t)_{t=1}^T$, possibly from different species, we synthesize

$$\hat{V} = G\left(I_s, M(V_d)\right), \tag{1}$$

where $\hat{V}$ preserves the source identity, appearance, and **species-specific morphology** while following the driving motion, $M(\cdot)$ denotes the motion extractor, and $G(\cdot)$ is the generator. Unlike previous human portrait animation methods, our goal is to enable **cross-species** animation, although the training datasets contain only same-species pairs including human, cat, and dog videos. As illustrated in Fig. 3 and Fig. 4, our method contains two stages:

1. **Learning cross-species motion alignment**. We employ LivePortrait (Guo et al., 2024) for motion alignment. Specifically, rigid head motion is represented by an affine transformation, while facial expression is encoded as shared implicit keypoints. First, we augment the shared expression backbone with LoRA components used by the species-routed expression encoder, enhancing the encoder's capacity to model anatomically distinct expression patterns while preserving a common representation across species. We then introduce expression-level adversarial domain adaptation (Ex-ADA) to suppress driving-species-specific information in the expression latents, thereby mitigating species leakage and facilitating cross-species expression alignment.
2. **DiT-based cross-species portrait animation.** Given the pretrained pose and expression encoders from Stage 1, we inject head pose as a motion flow and expression features as cross-attention tokens into the DiT-based video generation backbone. Benefiting from the well-aligned expression encoders with suppressed species leakage, the model achieves faithful cross-species animation despite being trained solely on same-species data pairs.

3.2 LEARNING CROSS-SPECIES MOTION ALIGNMENT

Cross-species motion transfer requires a common expression representation that accommodates anatomical differences while suppressing driving-species cues. We learn this representation within the LivePortrait (Guo et al., 2024) reconstruction framework using same-species source–driving pairs from human, cat, and dog videos.

Motion representations. We decompose portrait motion into rigid head pose and non-rigid facial expression. For an input portrait I with species label $c \in \{\text{human}, \text{cat}, \text{dog}\}$, we estimate

$$(R, t, s) = E_{\text{pose}}(I), \quad \delta_{\text{exp}} = H_{\text{exp}}(E_{\text{exp}}(I; c)) . \tag{2}$$

Here, R, $\mathbf{t}$, and s denote rotation, translation, and scale, respectively. The expression encoder produces an intermediate feature $F_{\text{exp}} = E_{\text{exp}}(I; c) \in \mathbb{R}^{768}$, which is mapped by the shared prediction head H_{exp} to $K = 21$ implicit 3D keypoint offsets $\boldsymbol{\delta}_{\text{exp}} \in \mathbb{R}^{K\times 3}$. During Stage 1, these motion components are combined with canonical keypoints to drive the shared warping and reconstruction modules, while appearance is provided by the source portrait.

Species-routed expression encoder. A shared expression backbone provides a common representation space, but must accommodate anatomically distinct expression patterns. We therefore augment selected backbone layers with LoRA components used by the species-routed expression encoder (Hu et al., 2022). The backbone and expression prediction head are shared across species, while these components provide additional capacity for interpreting subtle species-dependent facial expressions. As illustrated on the right of Fig. 6, the LoRA components used by the species-routed expression encoder enable more accurate expression reenactment.

Expression-level adversarial domain adaptation (Ex-ADA). Species-specific adaptation alone does not prevent expression features from retaining anatomical or appearance cues that can leak into the animated source. We therefore introduce expression-level adversarial domain adaptation (Ex-ADA). A three-class (human, cat, and dog) species classifier D takes the expression feature as input through a gradient reversal layer (GRL). Given a same-species source–driving pair (I_s, I_d) with species label c, we define the ADA loss as

$$\mathcal{L}_{\text{ADA}} = \frac{1}{2} \sum_{i\in\{s,d\}} \text{CE}\left(D\left(\text{GRL}(F^i_{\text{exp}})\right), c\right), \tag{3}$$

where $F^i_{\text{exp}} = E_{\text{exp}}(I_i; c)$ and CE denotes cross-entropy. The GRL acts as an identity mapping during the forward pass, while reversing and scaling the gradient during backpropagation. The ADA loss trains the classifier to predict species labels, while the reversed gradients encourage the expression encoder to suppress species-specific information in its features.

Ex-ADA complements the reconstruction objective: reconstruction and motion priors encourage the representation to retain expression-related information, while adversarial training discourages reliance on species cues. Together, they promote a shared expression interface for cross-species transfer without paired cross-species supervision. As shown on the left of Fig. 6, Ex-ADA enables expression transfer while better preserving the source animal's species-specific oral morphology.

3.3 DiT-Based Cross-Species Portrait Animation

We freeze the motion encoders learned in Stage 1 and condition a shared video DiT on driving pose and expression, while the reference pathway provides source appearance and morphology.

Pose condition. Following DeX-Portrait (Shi et al., 2026), we convert the predicted rotation, translation, and scale into dense pose-flow fields. These fields are temporally packed and projected to match the VAE latents, then concatenated with the noisy latents to provide spatial motion guidance.

Expression condition. For each driving frame t, we convert the predicted expression offsets δ^t_{exp} into keypoint-wise expression tokens. Specifically, for the k-th keypoint, we first obtain an initial token $z_{t,k}$ and refine its temporal trajectory:

$$z_{t,k} = \phi(\delta^t_{\text{exp},k}) + p_k, \qquad \widetilde{z}_{t,k} = z_{t,k} + \mathcal{G}_{\text{causal}}(z_{\le t,k}), \quad k = 1, \ldots, K. \tag{4}$$

where ϕ is a shared MLP that projects each 3D keypoint offset into a token embedding, p_k is a learned keypoint embedding, and $\mathcal{G}_{\text{causal}}$ applies left-padded causal convolutions independently to each keypoint trajectory. Refinement operates at the original frame rate, and its final projection is zero-initialized to preserve the initial frame-wise representation.

To match the VAE timeline, we prepend three copies of the first token frame and pack every four consecutive frames into $4K$ expression tokens per latent time step. After projection to the DiT hidden dimension, these tokens serve as keys and values in expression cross-attention adapters, with video tokens as queries. The adapters are inserted every two transformer blocks and use zero-initialized output projections. This supplies temporally refined expression guidance through the shared cross-species expression representations learned in Stage 1.

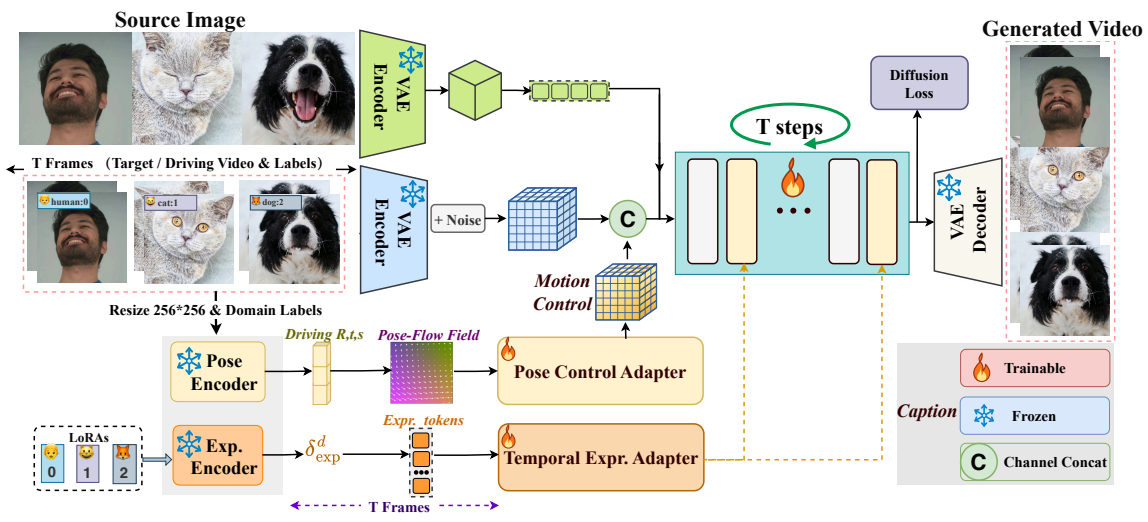

Figure 4: **Stage 2: DiT-based cross-species portrait animation.** We inject the pose and expression features extracted from pretrained pose and expression encoders into the video DiT for expressive cross-species animation.

3.4 Training

In Stage 1, pose and expression encoders (Liu et al., 2022) are initialized from pretrained human weights of LivePortrait (Guo et al., 2024) and jointly optimized with reconstruction, motion prior, and ADA losses. The warping and reconstruction modules are shared across species. In Stage 2, pose and expression encoders are frozen. We train the DiT denoising backbone, motion projections, causal expression refiner, and shared expression cross-attention adapters on human, cat, and dog videos. Neither stage introduces generator-side species routing to encourage learning facial structures from the source image. Complete objectives and optimization settings are provided in Appendix A.

4 Experiments

We evaluate motion fidelity and identity preservation across three transfer settings: human-to-animal (H→A), animal-to-human (A→H), and animal-to-animal (A→A). For A→A, we use balanced cat–dog transfer as the primary trained-species setting and additionally assess generalization to unseen source species qualitatively. We further conduct controlled studies with a joint No-Alignment baseline and ablations of Ex-ADA, the species-routed expression encoder, and temporal stabilization.

4.1 Experimental Setup

Data and splits. Both stages use the same training data, comprising public human portrait datasets and internally collected cat and dog videos. Stage 1 uses two sampled frames per video for reconstruction training, while Stage 2 uses continuous 81-frame clips. All evaluation identities and driving clips are held out from training. The animal self-reenactment benchmark contains 50 cat and 50 dog videos; the remaining evaluation covers H→A, A→H, cross- and same-species A→A, and unseen-source transfer with trained-species drivers. Dataset details, preprocessing, resolutions, and split construction are provided in Appendix G.

Compared methods. We group the compared methods by their roles and supported settings. LivePortrait-Animal (Guo et al., 2024) serves as our principal external animal-trained baseline for H→A, A→H, and the applicable A→A settings. FaceShot (Gao et al., 2025), Follow-Your-Emoji (Ma et al., 2024), FantasyPortrait (Wang et al., 2025), and Wan-Animate-2 (Wang et al., 2026) provide additional comparisons in the cross-species settings they support. FantasyPortrait's released pipeline relies on human-face detection/alignment and pretrained facial-motion features, so it does not officially support animal videos as drivers. Its animal examples use human drivers for animal portraits; we therefore retain it only for H→A and mark animal self-reenactment as N/A. Human-only methods are evaluated only in the H→H sanity check because they do not support the cross-species

Table 1: Animal self-reenactment reconstruction and H→A pairwise preference evaluation. Self-reenactment uses 50 cat and 50 dog videos. Each preference value denotes the rate at which users prefer the corresponding method over CROSSPORTRAIT.

Method	Animal Self-Reenactment				H→A User Preference: Method over CROSSPORTRAIT			
	PSNR ↑	SSIM ↑	LPIPS ↓	L_1 ↓	Expr. (%) ↓	Pose (%) ↓	ID (%) ↓	Overall (%) ↓
Follow-Your-Emoji	21.38	0.7639	0.1730	0.05030	0.0	0.0	0.0	0.0
FaceShot	16.17	0.6019	0.3822	0.1012	3.2	1.6	12.9	3.2
LivePortrait-Animal	24.37	0.7720	0.1438	0.04010	21.2	21.2	34.6	23.1
FantasyPortrait	N/A	N/A	N/A	N/A	0.0	0.0	12.9	0.0
Wan-Animate-2	23.54	0.7834	0.1955	0.0590	30.0	22.5	34.2	19.8
No-Alignment Baseline	24.59	0.7743	0.1400	0.04001	23.6	26.9	14.9	26.2
CROSSPORTRAIT	**25.06**	**0.7893**	**0.1360**	**0.03970**	/	/	/	/

settings. Appendix C shows that CROSSPORTRAIT achieves the best or tied-best H→H scores on all metrics except APD, on which it ranks second. The **No-Alignment Baseline** uses the same human–cat–dog data and generator backbone, providing a matched-data joint baseline for the cross-species motion interface; component ablations are reported separately below. For each supported setting, all methods are evaluated using identical source–driving pairs and their official preprocessing. Each applicable baseline is post-trained on the same animal training data using its official pipeline. Methods with partial applicability are evaluated only on the settings they support, and unsupported method–direction pairs are marked N/A. Protocol-specific applicability is summarized in Appendix Table 5.

Evaluation metrics. For paired animal self-reenactment, we use the first frame of each target video as the source image and the remaining frames as ground-truth targets. We report peak signal-to-noise ratio (PSNR), structural similarity (SSIM) (Wang et al., 2004), learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), and L_1 error. For A→H, we additionally report CSIM between the generated human face and the source portrait to measure source identity preservation; CSIM is not used to measure motion fidelity. For same-species animal cross-identity transfer, we report pairwise preferences in Appendix Table 6. AED and APD are reported only for human frames. Detailed metric definitions are provided in Appendix B.

Pairwise user study. Because the evaluated cross-species and animal cross-identity settings lack paired ground truth and a validated animal-motion evaluator, H→A, A→H, and cross- and same-species A→A trials use anonymized pairwise preferences. Participants view the source portrait, a synchronized driving clip, and two anonymized outputs, with the output order randomized and counterbalanced. They independently judge four criteria: Expr. measures expression fidelity and temporal alignment; Pose measures head-pose agreement; the ID criterion measures preservation of source identity and morphology, including driving-species leakage; and Overall measures motion, identity and morphology preservation, temporal coherence, and visual quality. The study enrolled 50 participants, each completing 10 comparison trials and five ablation trials, yielding 500 comparison and 250 ablation trials in total. A→H pairs are balanced between cat and dog drivers. For cross-species A→A, we use equal numbers of cat→dog and dog→cat pairs, and every regime is balanced across the motion strata defined in Appendix G. Every question includes a tie option, with ties assigned a preference score of 0.5. We report Method over CROSSPORTRAIT preference rates. Lower values indicate better performance by CROSSPORTRAIT. Failures of applicable official pipelines on the fixed evaluation pairs are counted as losses, while unsupported method–direction pairs are marked N/A and excluded from the analysis.

4.2 CROSS-SPECIES RESULTS

Quantitative comparison. Tables 1 and 2 report reconstruction, CSIM, and pairwise preference results across animal self-reenactment and cross-species transfer. Figure 5 shows fixed-pair examples, and Appendix Table 4 reports the H→H sanity check. CROSSPORTRAIT achieves the best self-reenactment metrics and highest A→H CSIM, while every valid baseline-over-CROSSPORTRAIT preference rate remains below 50% across cross-species settings.

Qualitative comparison. Figures 1 and 5 present bidirectional cross-species transfer, including unseen source portraits driven by trained species. Baselines show motion errors, geometric distortion, and species-specific feature leakage, while CROSSPORTRAIT better preserves source morphology and follows the driving motion.

Table 2: A→H and cross-species A→A (A1→A2) evaluation. A→H uses balanced cat and dog driving clips. Each value denotes the preference rate for the corresponding method over CROSS-PORTRAIT; lower values indicate better performance by CROSSPORTRAIT.

Method	A→H (Balanced cat/dog drivers)					Cross-species A→A (A1→A2)			
	CSIM ↑	Expr. (%) ↓	Pose (%) ↓	ID (%) ↓	Overall (%) ↓	Expr. (%) ↓	Pose (%) ↓	ID (%) ↓	Overall (%) ↓
FaceShot	0.8657	0.0	0.0	6.2	0.0	0.0	0.0	0.0	0.0
Follow-Your-Emoji	0.8445	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LivePortrait-Animal	0.8625	0.0	0.0	0.0	0.0	0.0	0.0	12.5	0.0
Wan-Animate-2	0.8629	0.4	8.8	4.2	7.9	6.2	8.2	0.0	3.8
No-Alignment Baseline	0.8451	18.2	8.2	0.0	12.2	23.9	17.1	10.9	20.6
CROSSPORTRAIT	**0.8825**	/	/	/	/	/	/	/	/



Figure 5: Qualitative comparison of CROSSPORTRAIT with competing methods. The top, middle, and bottom rows show human-to-animal, animal-to-human, and cross-species cat↔dog animation, respectively. **Blue**, **red**, and **green** borders indicate **source** portraits, **driving** images, and **Cross-Portrait outputs**, respectively. Red boxes mark visible artifacts, and gray cells indicate unsupported cases.

4.3 ABLATION STUDIES

All ablation variants use the same data, source–driving pairs, and generator backbone, with the corresponding user-study results summarized in Table 3.

No-Alignment Baseline. The No-Alignment variant uses a single shared ConvNeXt encoder to process both pose-flow and expression-token motion. It removes the LoRA components used by the species-routed expression encoder and does not apply expression-level adversarial domain adaptation (Ex-ADA). This configuration substantially weakens cross-species motion transfer and can yield unstable mouth geometry (Figure 6, left).

w/o Ex-ADA. Removing only expression-level ADA leaves more driving-species information in the expression features, increasing species-specific feature leakage in H→A transfer (Figure 6, left). ADA suppresses species-predictive information at the expression level without applying adversarial alignment to the pose-flow or source-reference pathways.

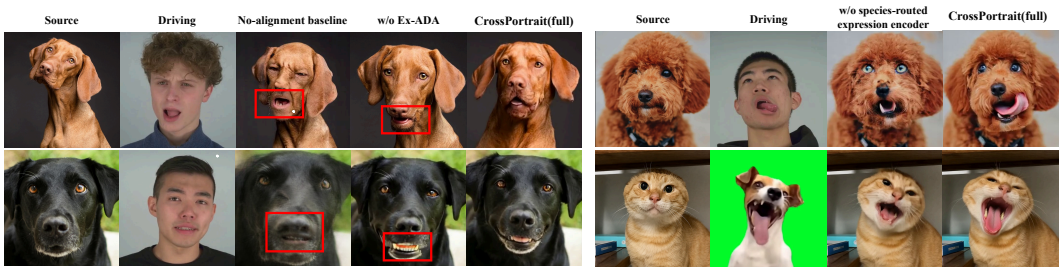

Figure 6: Qualitative H→A ablations. Left: No-Alignment and w/o expression-level ADA. Right: w/o species-routed expression encoder.

Table 3: Ablation user study on bidirectional human–animal animation. Each value denotes the preference rate for the corresponding method over CROSSPORTRAIT; lower values indicate better performance by CROSSPORTRAIT.

Variant	Expr. (%) ↓	Pose (%) ↓	ID (%) ↓	Overall (%) ↓
(A) No-alignment baseline	6.4	15.4	11.5	14.1
(B) w/o Ex-ADA	23.1	26.9	23.1	19.2
(C) w/o species-routed expression encoder	21.2	30.3	33.3	27.9
(D) w/o temporal stabilization.	30.0	48.0	44.0	32.0

w/o species-routed expression encoder. Removing the species-routed expression encoder leaves a shared expression encoder to model species-dependent anatomical variations, weakening fine-grained motions such as tongue protrusion and mouth opening. The species-routed expression encoder restores the ability to model variations while retaining a shared expression backbone (Figure 6, right).

w/o temporal stabilization. Replacing causal temporal refinement with non-causal processing allows future driving states to influence earlier expression tokens, leading to less temporally consistent mouth and tongue motion. This results in lower Expression and Overall preference rates, and can also slightly degrade perceived pose quality during joint generation.

5 LIMITATIONS

Our training data cover human, cat, and dog drivers. Generalization to unseen species is therefore evaluated only qualitatively on the source side, using drivers from the trained species, while unseen-species driving and arbitrary facial topologies remain unexplored. Moreover, cross-species animation lacks paired ground truth and validated cross-species motion metrics. As a result, preference-based evaluation and identity metrics cannot fully characterize motion accuracy or biological plausibility.

6 CONCLUSION

We presented CROSSPORTRAIT, a unified framework for cross-species portrait animation across human, cat, and dog domains. Rather than enforcing domain invariance over all motion information, CROSSPORTRAIT uses complementary pathways for explicit geometric motion and learned expression conditioning, retains species-specific modeling capacity through the species-routed expression encoder, and applies alignment only to expression features to suppress species-predictive information. A lightweight causal temporal module further stabilizes expression-motion tokens before video generation. Our results show that this design enables a single shared generator to support diverse source–driving directions while improving motion transfer and source preservation over applicable baselines. Ablation studies further indicate the importance of balancing shared motion representations with species-specific modeling capacity for cross-species animation. While the model generalizes qualitatively to unseen source species when driven by trained species, unseen driver species remain unexplored. Extending the framework to a broader range of species and less structured motion domains is a promising direction for future work.

Reproducibility Statement

We describe the model architecture, training objectives, optimization settings, data construction, evaluation splits, metric definitions, and user-study protocol in the main paper and the supplementary appendices. All evaluation source identities and driving clips are held out from training, and the implementation details needed to reproduce the reported experiments are provided in the supplementary material.

AI Assistance

We used large language models to assist with language editing and manuscript organization, including improving grammar, clarity, and presentation. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript, claims, and references.

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 343–351, 2016.

Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Towards real-world blind face restoration with generative diffusion prior. *arXiv preprint arXiv:2312.15736*, 2023. URL https://arxiv.org/abs/2312.15736.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-Animate: Unified character animation and replacement with holistic replication, 2025. URL https://arxiv.org/abs/2509.14055.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. EMOPortraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8498–8507, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://www.jmlr.org/papers/v17/15-239.html.

Junyao Gao, Yanan Sun, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, and Cairong Zhao. FaceShot: Bring any character into life. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oJA1GUqRww.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. LivePortrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. URL https://arxiv.org/abs/2407.03168.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 42(4):161:1–161:14, 2023. doi: 10.1145/3592455. URL https://doi.org/10.1145/3592455.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 1640–1650, 2018.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-Your-Emoji: Fine-controllable and expressive freestyle portrait animation. In *ACM SIGGRAPH Asia 2024 Conference Papers*, pp. 110:1–110:12, 2024. doi: 10.1145/3680528.3687587.

Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Zhifeng Li, Wei Liu, Linfeng Zhang, and Qifeng Chen. Follow-Your-Emoji-Faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *International Journal of Computer Vision*, 134(3):130, 2026. doi: 10.1007/s11263-025-02685-z.

Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired human captures for complete, driveable, and generalizable avatars. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83008–83023. Curran Associates, Inc., 2024. doi: 10.52202/079017-2640. URL https://papers.nips.cc/paper_files/paper/2024/hash/9712b78386cebdc3db7f1a48c2d20edb-Abstract-Datasets_and_Benchmarks_Track.html.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. MOFA-Video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pp. 111–128, 2024. doi: 10.1007/978-3-031-72655-2_7.

Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-Ming Yan. DPE: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2023.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3934–3941, 2018. doi: 10.1609/aaai.v32i1.11767.

Pexels. Pexels. https://www.pexels.com/. Accessed: 2026-07-08.

Pixabay. Pixabay. https://pixabay.com/. Accessed: 2026-07-08.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30, pp. 506–516, 2017.

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. PIRenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13759–13768, 2021.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2490–2501, 2024. doi: 10.1109/CVPR52733.2024.00241.

Yuxiang Shi, Zhe Li, Yanwen Wang, Hao Zhu, Xun Cao, and Ligang Liu. DeX-Portrait: Disentangled and expressive portrait animation via explicit and latent motion representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40009–40019, 2026.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13653–13662, 2021.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-X: Universal character image animation with enhanced motion representation. In *International Conference on Learning Representations*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Xindi Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, Gang Cheng, Hai Xu, and Bang Zhang. Wan-Animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026. URL https://arxiv.org/abs/2608.06009.

Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. FantasyPortrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025. URL https://arxiv.org/abs/2507.12956.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10039–10049, 2021.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 657–666, 2022.

You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-Portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657459.

Yating Xu, Yunqi Miao, Evangelos Ververas, Jiankang Deng, and Jifei Song. FG-Portrait: 3d flow guided editable portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32947–32956, 2026.

Yuanfeng Xu, Yuhao Chen, Zhongzhan Huang, Zijian He, Guangrun Wang, Philip H. S. Torr, and Liang Lin. AnimateZoo: Zero-shot video generation of cross-species animation via subject alignment. *arXiv preprint arXiv:2404.04946*, 2024. URL https://arxiv.org/abs/2404.04946.

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. HunyuanPortrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15909–15919, 2025. doi: 10.1109/CVPR52734.2025.01483.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 122647–122676. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b1d4973f5d21708abef3cd6f17d842c8-Paper-Conference.pdf.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3657–3666, 2022.

Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-NeMo: Expressive neural motion reenactment via disentangled latent attention. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bfb6a69c0d9e2bc596e1cd31f16fcdde-Abstract-Conference.html.

Guangting Zheng, Haojing Chen, Hao Li, Jingtao Zhang, Zhen Yang, Xiaosong Jia, Xue Yang, Shaofeng Zhang, and Yanyong Zhang. Enhancing video physical consistency via role-aware joint training and modality-decoupled denoising. *arXiv preprint arXiv:2607.04653*, 2026. URL https://arxiv.org/abs/2607.04653.

Supplementary Material

A Method and Training Details . . . 15
A.1 Pose Representation and Pose-Flow Conditioning . . . 15
A.2 Adversarial Schedule and Dataset Sampling . . . 15
A.3 Stage-1 Reconstruction Objective . . . 15
B Evaluation Metric Details . . . 16
B.1 Notation and Temporal Aggregation . . . 16
B.2 Average Expression Distance . . . 16
B.3 Average Pose Distance . . . 16
B.4 Cosine Similarity . . . 16
C Human-to-Human Sanity Check . . . 16
C.1 Evaluation Setup . . . 16
D Baseline Applicability . . . 17
D.1 Applicability Matrix . . . 17
E Implementation Details . . . 17
E.1 Stage-1 Network Architecture . . . 17
E.2 Stage-1 Optimization . . . 18
E.3 Stage-2 Motion Conditioning . . . 18
E.4 Stage-2 Optimization . . . 19
F Same-Species Animal Evaluation . . . 19
F.1 Evaluation Setting . . . 19
G Data and Evaluation Protocol . . . 19
G.1 Training Data . . . 19
G.2 Evaluation Splits . . . 20
G.3 Pairwise User-Study Protocol . . . 20

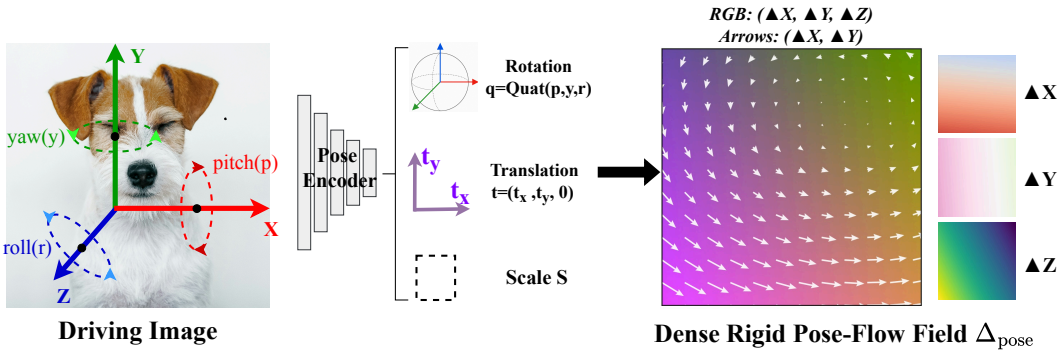

Figure 7: Pose branch for dense rigid pose-flow. PoseNet predicts pitch, yaw, roll, translation, and scale; these parameters are converted into the dense three-channel displacement field Δ_{pose}.

A Method and Training Details

A.1 Pose Representation and Pose-Flow Conditioning

Pose parameterization. PoseNet predicts pitch, yaw, and roll, together with translation and scale:

$$p, y, r = H_{\text{rot}}(F_{\text{pose}}), \qquad t = H_t(F_{\text{pose}}), \qquad s = H_s(F_{\text{pose}}). \tag{5}$$

Each rotation angle is obtained by taking a soft expectation over B_{bin} bins, $\theta = \sum_i \text{softmax}(l)_i b_i$, where l denotes the angle logits and b_i the corresponding bin centers. The resulting Euler angles are converted to a quaternion $q = \text{Quat}(p, y, r)$. For translation, we use $t = (t_x, t_y, 0)$ in normalized crop coordinates. These pose parameters are further converted into a dense rigid pose-flow field, as illustrated in Figure 7. For video conditioning, four consecutive pose-flow fields are channel-packed, resulting in 12 channels, which are projected by a $1 \times 1 \times 1$ Conv3D to the 48-channel VAE latent-control dimension.

A.2 Adversarial Schedule and Dataset Sampling

After a 3,000-step warm-up, the GRL multiplier used by Ex-ADA follows

$$\lambda(p) = \frac{2}{1 + \exp(-\gamma p)} - 1, \tag{6}$$

where p denotes the normalized training progress and $\gamma = 10$. The sampling weights of the human, cat, and dog datasets have total masses of 6.7, 7.5, and 7.5, respectively, corresponding to approximately 30.9%, 34.6%, and 34.6% of the training mixture. Training batches are shuffled from this mixture rather than explicitly constructed to be class-balanced.

A.3 Stage-1 Reconstruction Objective

The full reconstruction-stage loss is

$$\begin{aligned}\mathcal{L}_{\text{stage1}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{equiv}}\mathcal{L}_{\text{equiv}} + \lambda_{\text{kp}}\mathcal{L}_{\text{kp}}^{\text{prior}} \\ + \lambda_{\text{def}}\mathcal{L}_{\text{def}}^{\text{prior}} + \lambda_{\text{gan}}\mathcal{L}_{\text{gan}} + \lambda_{\text{Ex-ADA}}\mathcal{L}_{\text{Ex-ADA}}.\end{aligned} \tag{7}$$

All terms operate on Stage-1 source–driving reconstruction pairs. Stage 1 contains no temporal operation, and causal processing is used only in the Stage-2 expression refiner. The reconstruction and perceptual terms preserve image fidelity, the equivariance term regularizes motion consistency, and the keypoint and deformation priors stabilize the predicted geometry. The global GAN term improves local realism, while Ex-ADA penalizes species-predictive information in the expression representation through the gradient-reversal pathway.

B Evaluation Metric Details

B.1 Notation and Temporal Aggregation

Here, $\hat{I}^t$ and I_d^t denote the generated and driving frames at time t, respectively, and T denotes the number of evaluated frames.

B.2 Average Expression Distance (AED)

AED measures the accuracy of facial expression transfer using parameters extracted by SMIRK (Retsinas et al., 2024). Following the evaluation protocol of prior work, we construct the expression vector by concatenating the facial-expression, eyelid, and jaw-pose parameters predicted by SMIRK:

$$E_{\mathrm{SMIRK}}(I) = [\mathbf{e}(I), \mathbf{l}(I), \mathbf{j}(I)], \tag{8}$$

where $\mathbf{e}$, $\mathbf{l}$, and $\mathbf{j}$ denote the facial-expression, eyelid, and jaw-pose parameters, respectively. AED is computed as the mean absolute difference between the expression vectors of the generated and driving frames:

$$\mathrm{AED} = \frac{1}{TD_E} \sum_{t=1}^{T} \left\| E_{\mathrm{SMIRK}}(\hat{I}^t) - E_{\mathrm{SMIRK}}(I_d^t) \right\|_1, \tag{9}$$

where D_E is the dimensionality of the expression vector. Lower AED indicates better agreement with the driving expression.

B.3 Average Pose Distance (APD)

APD measures head-pose accuracy using the pose parameters predicted by SMIRK. Let $P_{\mathrm{SMIRK}}(I)$ denote the SMIRK head-pose parameter vector. APD is computed as

$$\mathrm{APD} = \frac{1}{TD_P} \sum_{t=1}^{T} \left\| P_{\mathrm{SMIRK}}(\hat{I}^t) - P_{\mathrm{SMIRK}}(I_d^t) \right\|_1, \tag{10}$$

where D_P is the dimensionality of the pose vector. Lower APD indicates better agreement with the driving head pose.

B.4 Cosine Similarity (CSIM)

Identity preservation is evaluated using cosine similarity between ArcFace (Deng et al., 2019) embeddings. For cross-identity reenactment, CSIM is computed between each generated frame and the source portrait:

$$\mathrm{CSIM} = \frac{1}{T} \sum_{t=1}^{T} \cos\left(A_{\mathrm{ArcFace}}(\hat{I}^t), A_{\mathrm{ArcFace}}(I_s) \right), \tag{11}$$

where $A_{\mathrm{ArcFace}}(\cdot)$ denotes the pretrained ArcFace embedding extractor, and $\cos(\cdot,\cdot)$ denotes cosine similarity. Higher CSIM indicates better preservation of the source identity.

C Human-to-Human Sanity Check

C.1 Evaluation Setup

Human-to-human animation is not the primary focus of this work. We nevertheless include an H→H evaluation as a sanity check to examine whether multi-species training and the proposed motion-conditioning design maintain standard human portrait-animation performance. We evaluate on 100 held-out human videos from Pixabay (Pixabay) and Pexels (Pexels), using the same preprocessing protocol as in the animal evaluation. For self-reenactment, the first frame serves as the source portrait and the remaining frames provide pixel-level ground truth. For cross-identity reenactment, the source and driving videos contain different identities. We compare with Follow-Your-Emoji, LivePortrait-Human, FaceShot, HunyuanPortrait (Xu et al., 2025), X-NeMo (Zhao et al., 2025), and

Table 4: Human-to-human sanity check. Self-reenactment is evaluated with pixel-level ground truth, while cross-identity reenactment uses unpaired human motion and identity metrics.

Method	Self-Reenactment					Cross-Identity Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	L_1 ↓	CSIM ↑	AED ↓	APD ↓	CSIM ↑
Follow-Your-Emoji	22.20	0.7650	0.1204	0.0428	0.8130	0.3039	0.0540	0.4282
LivePortrait-Human	20.83	0.7419	0.1443	0.0518	0.8750	0.3328	0.2615	0.5528
FaceShot	17.99	0.6836	0.2382	0.0696	0.7078	0.3868	0.3104	0.2971
HunyuanPortrait	21.46	0.7584	0.1328	0.0459	0.8589	0.2270	0.1182	0.7208
X-NeMo	21.12	0.7501	0.1397	0.0486	0.8463	0.2250	0.0200	0.6650
Wan-Animate	24.05	0.8065	0.1055	0.0316	0.8407	0.2402	**0.0190**	0.6056
CROSSPORTRAIT	**24.20**	**0.8156**	**0.1016**	**0.0305**	**0.8765**	**0.2241**	0.0191	**0.7308**

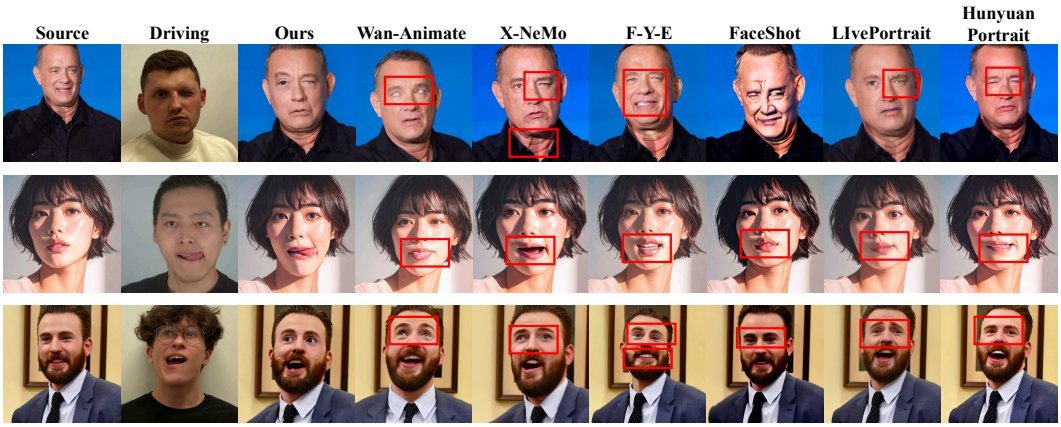

Figure 8: Qualitative H→H cross-identity comparison on fixed source–driving pairs. Columns show the source portrait, driving frame, and outputs from CROSSPORTRAIT, Wan-Animate, X-NeMo, Follow-Your-Emoji, FaceShot, LivePortrait-Human, and HunyuanPortrait. Red boxes highlight local artifacts in the eye or mouth regions.

Wan-Animate (Cheng et al., 2025). For self-reenactment, we report PSNR, SSIM, LPIPS, L_1, AED, APD, and CSIM. For cross-identity reenactment, where pixel-level ground truth is unavailable, we report the unpaired human-face metrics AED, APD, and CSIM. Table 4 and Figure 8 report the quantitative and qualitative comparisons, respectively.

D BASELINE APPLICABILITY

D.1 APPLICABILITY MATRIX

Table 5 distinguishes source-side animal generalization from support for animal driving. Cross-species A→A is evaluated quantitatively with balanced cat↔dog pairs; additional qualitative examples use unseen species only as sources.

E IMPLEMENTATION DETAILS

E.1 STAGE-1 NETWORK ARCHITECTURE

PoseNet and ExpNet use separate ConvNeXt backbones initialized from the corresponding weights of the human LivePortrait checkpoint and are jointly trained under the Stage-1 objectives. Both networks take 256×256 portrait crops as input. PoseNet uses 66-bin heads for pitch, yaw, and roll, with soft expectation converting the predicted distributions to angles at a 3° interval before quaternion conversion. The expression head predicts 21 shared 3D keypoint offsets. The species-routed expression encoder is implemented using LoRA residuals inserted only into the pwconv2 linear layers of the last three blocks of ConvNeXt stage 3 and all three blocks of stage 4, corresponding to zero-based stage indices 2 and 3. Each human, cat, and dog route uses rank 8, $\alpha = 8$, and dropout

Table 5: Applicability of compared methods under our evaluation protocol.

	Same-species				Cross-species
Method	H→H	A→A	H→A	A→H	A→A
Cross-domain and animal-capable baselines					
Follow-Your-Emoji	Yes	Partial	Yes	Partial	Partial
LivePortrait-Animal	N/A	Yes	Yes	Yes	Partial
FaceShot	Yes	Yes	Yes	Yes	Yes
FantasyPortrait	Yes	N/A	Yes	N/A	N/A
Wan-Animate-2	Yes	Yes	Yes	Yes	Yes
Human-only baselines					
LivePortrait-Human	Yes	N/A	N/A	N/A	N/A
HunyuanPortrait	Yes	N/A	N/A	N/A	N/A
X-NeMo	Yes	N/A	N/A	N/A	N/A
Wan-Animate	Yes	N/A	N/A	N/A	N/A
Matched-data comparison					
No-Alignment Baseline	Yes	Yes	Yes	Yes	Yes
CROSSPORTRAIT	Yes	Yes	Yes	Yes	Yes

Protocol notes. "Yes" denotes an applicable setting covered by the evaluation protocol, while "Partial" denotes cases in which valid outputs are available for only a subset of source–driving pairs. N/A denotes an officially unsupported setting and is excluded from the corresponding evaluation. "Failed" denotes an applicable setting that was attempted but did not produce a valid output; such failures are retained in the evaluation rather than discarded. LivePortrait-Human and LivePortrait-Animal are separate released pipelines; the latter supports animation with animal data and is evaluated on animal-driven settings where valid outputs can be obtained. FantasyPortrait is used as a zero-shot H→A baseline in our cross-species evaluation and does not officially support animal driving. Wan-Animate and Wan-Animate-2 are treated as separate baselines: the former is used in the H→H sanity check, while the latter is evaluated in cross-species and animal-driven settings.

0.05. The LoRA up-projection is zero-initialized, such that all routed branches are initially identical to the shared backbone.

The three-class species discriminator receives the 768-dimensional ExpNet backbone feature and has linear dimensions $768 \to 256 \to 64 \to 3$. The two hidden layers use LeakyReLU with slope 0.2 and dropout 0.3; Gaussian noise with standard deviation 0.1 is added to discriminator inputs during training. The GRL is disabled for the first 3,000 optimization steps and then follows the schedule in Eq. (6) with $\gamma = 10$. Its expression-gradient scale is 0.30. Ex-ADA is applied only to the expression branch and not to the pose branch.

E.2 STAGE-1 OPTIMIZATION

Stage 1 uses mean squared error (MSE) reconstruction, VGG perceptual, equivariance, keypoint-prior, deformation-prior, global image generative adversarial network (GAN), and Ex-ADA losses. Their respective weights are 0.1, 0.1, 1.0, 1.0, 1.0, 0.5, and 0.3. We optimize with Adam using $\beta = (0, 0.999)$. The learning rates are 5×10^{-5} for PoseNet and the remaining generator modules, 10^{-5} for the shared ExpNet backbone, 2×10^{-5} for the expression keypoint heads, 5×10^{-5} for the species-routed LoRA parameters, and 10^{-4} for the image discriminator. Training runs for approximately three days on four GPUs using bfloat16 mixed precision, a per-GPU batch size of 20 without gradient accumulation, gradient clipping at 1.0, and seed 42.

E.3 STAGE-2 MOTION CONDITIONING

Stage 2 uses a Wan2.2-Fun-5B-Control backbone and a frozen Wan VAE (Team Wan et al., 2025). Training clips contain 81 frames at 512×512 with temporal stride 1; shorter clips are padded by repeating sampled frames. Driving frames are resized to 256×256 for MotionNet. The frozen pose branch converts absolute pose into a three-channel dense pose-flow field at the VAE latent spatial resolution. Four consecutive fields are channel-packed, with the first field repeated to align the first latent time step, and a $1 \times 1 \times 1$ Conv3D projects the resulting 12 channels to the 48-channel VAE latent control dimension.

Table 6: Same-species animal cross-identity user study. Entries are pairwise preference rates for CROSSPORTRAIT over each evaluated baseline, with ties counted as half a vote. The preference columns are percentages (%): Expr. measures expression fidelity and temporal alignment; Pose measures head-pose agreement; ID measures source identity and morphology preservation, including driving-species leakage; and Overall measures motion, source preservation, temporal coherence, and visual quality.

Method	Valid n	Expr. (%) ↑	Pose (%) ↑	ID (%) ↑	Overall (%) ↑
FaceShot	50	100	100	100	100
Follow-Your-Emoji	50	100	100	100	100
LivePortrait-Animal	50	100	100	91	100
Wan-Animate-2	50	100	98	100	100

Each of the 21 three-dimensional expression offsets is mapped by a $3 \to 64 \to 512$ multilayer perceptron with Gaussian error linear unit (GELU) activations and added to a learned keypoint embedding. The token-wise causal residual refiner applies non-affine LayerNorm followed by left-padded temporal convolutions $512 \to 1024 \to 512$, both with kernel size 3, and a sigmoid linear unit (SiLU) activation between them. The final convolution is zero-initialized, and the encoder output is added residually to preserve temporal length and provide an identity initialization. After prepending three copies of the first frame, every four frames are packed into 84 tokens and linearly projected to the DiT hidden dimension. Expression cross-attention adapters are inserted every two transformer blocks, and their output projections are zero-initialized. Expression conditioning is independently dropped for each training sample with probability 0.2.

E.4 STAGE-2 OPTIMIZATION

The VAE, PoseNet, ExpNet, and Stage-1 species-routed LoRA parameters are frozen in Stage 2. We train the full DiT denoising backbone, the $3 \to 64 \to 512$ expression projection and positional embedding, the pose Conv3D projection, the causal expression refiner, and the expression cross-attention adapters. The flow-matching target is the difference between sampled Gaussian noise and the clean video latent, optimized with the configured sigma-weighted MSE objective. We use AdamW with $\beta = (0.9, 0.999)$, weight decay 0.03, and $\epsilon = 10^{-10}$. The main DiT learning rate is 2×10^{-5}, the patch-embedding rate is 10^{-5}, and the learning rate for motion projections and expression cross-attention adapters is 10^{-4}. The learning-rate scheduler is constant; the configuration specifies 500 warmup steps per process and the distributed launcher passes $500 \times 4 = 2{,}000$ warmup steps to the scheduler. Training uses four GPUs, bfloat16 precision, gradient checkpointing, a per-GPU batch size of 1 without accumulation, gradient clipping at 0.05, and seed 42. Checkpoints and validation samples are produced every 1,500 and 750 optimization steps, respectively.

F SAME-SPECIES ANIMAL EVALUATION

F.1 EVALUATION SETTING

Table 6 reports the supplementary same-species animal cross-identity user study. Because no broadly accepted species-aware motion evaluator or paired cross-identity ground truth is available, animal-to-animal cross-identity evaluation uses pairwise human preference rather than a human-face metric.

G DATA AND EVALUATION PROTOCOL

G.1 TRAINING DATA

Both stages are trained on the same underlying human–cat–dog collection, which combines the human portrait datasets NeRSemble (Kirschstein et al., 2023), AVA-256 (Martinez et al., 2024), PFHQ (Chen et al., 2023), and VFHQ (Xie et al., 2022) with internally collected cat and dog portrait videos. Each animal domain contributes approximately 15 hours of data. For Stage 2, the same collection is organized into a merged metadata index, where the known driving-species label s_{drv} selects the corresponding ExpNet route. Both stages follow the same portrait detection and cropping pipeline. Stage 1 feeds 256×256 crops to PoseNet and ExpNet and supervises reconstruction

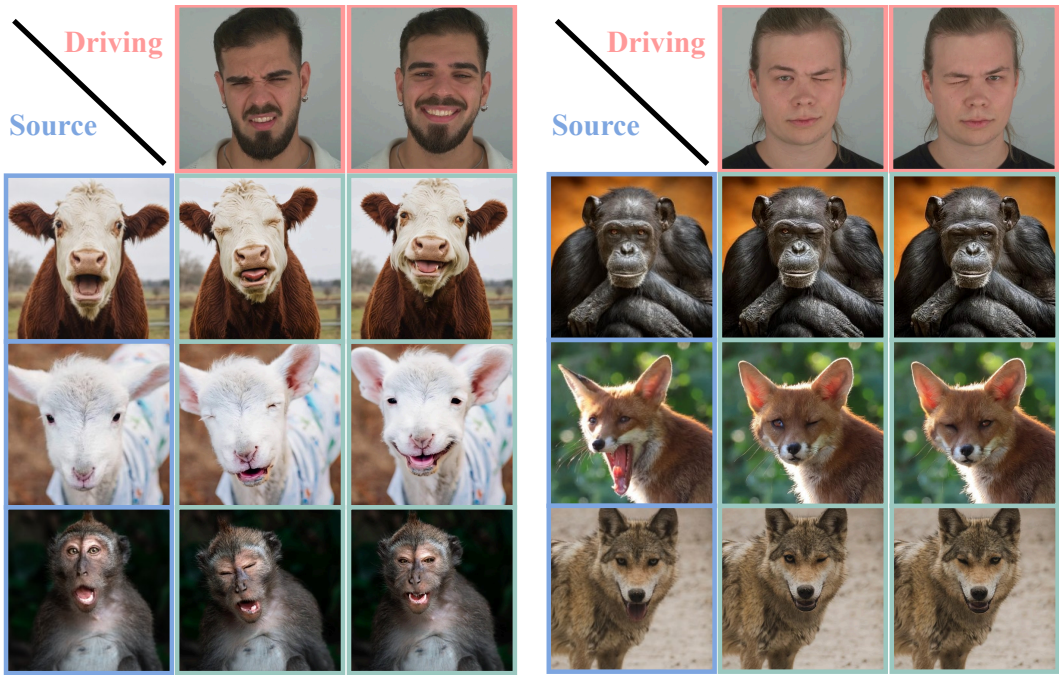

Figure 9: Qualitative unseen-source generalization by CROSSPORTRAIT. Our method transfers human driving motion while preserving the identity and species-specific morphology of source portraits from unseen species. **Blue**, **red**, and **green** borders denote **source** portraits, **driving** images, and **CrossPortrait outputs**, respectively. The driving species is human and is covered by training; this figure therefore evaluates source-side generalization only and does not evaluate unseen-species driving.

at 512×512, whereas Stage 2 trains on 512×512 video clips with MotionNet inputs resized to 256×256.

G.2 EVALUATION SPLITS

All source identities and driving clips used for evaluation are held out from training. The animal self-reenactment subset contains 50 cat and 50 dog videos collected from Pixabay (Pixabay) and Pexels (Pexels), each cropped to 512×512 and trimmed to 3–5 seconds. Beyond self-reenactment, the evaluation includes H→A, A→H, cross-species A→A, same-species animal cross-identity transfer, and a qualitative unseen-source study with trained-species drivers (Fig. 9).

G.3 PAIRWISE USER-STUDY PROTOCOL

For each pairwise user-study setting, driving clips are balanced across five motion types: head rotation, head nodding, eye closure, mouth opening, and mixed motion. This selection emphasizes motions with meaningful correspondence across species and avoids over-weighting a single expression or pose pattern. For H→A, cat and dog source portraits are equally represented so that the preference rates are not dominated by one animal species.

Each comparison uses the same source–driving pair for the methods being compared. Participants judge expression fidelity, head-pose agreement, source identity and morphology preservation, and overall visual quality in separate preference questions. A tie contributes half a vote to each method, so a reported preference rate is computed as the wins plus half of the ties divided by the number of valid comparisons. This protocol makes the preference columns comparable across settings while keeping the motion distribution balanced.