CROWN: CONVERSION TO ROUTED EXPERTS WITH OUTPUT-AWARE WIDTH NARROWING FOR AUDIO-VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward networks (FFNs) account for a substantial fraction of the parameters and per-step computation in video diffusion transformers, yet their intra-expert redundancy remains underexplored in multimodal token-routing settings. We start from DaVinci-MagiHuman (DaVinci), a pretrained single-stream audio-video diffusion transformer, and sparse-upcycle the shared MLPs in its 32 middle blocks into token-level Top-1 Sparse-MoE layers, while leaving the modality-specific early and late blocks unchanged. This conversion restructures the shared FFNs into routed expert subspaces without reducing the active intermediate width. Building on TENP, we propose Elastic Expert Neuron Pruning (EENP), a modality-aware method for structured intra-expert pruning in routed multimodal DiTs. EENP estimates intermediate-neuron importance separately from video, audio, and text tokens, normalizes the aggregated statistics within each modality, and fuses them into a modality-aware score with weights 3:1:1 for video, audio, and text. By aggregating statistics within each modality before fusion, the scoring procedure prevents modalities with more tokens from dominating the pruning decision. Low-importance neurons are then structurally removed, with the retained neurons stored as a nested Fast prefix of a wider Quality width. Under a 60% nominal pruning ratio, EENP retains 92.4% of the unpruned SparseMoE VBench-2.0 score and reduces DiT denoising latency by 22.0% relative to the unpruned SparseMoE checkpoint.

Keywords Audio-video diffusion · Mixture-of-experts · Diffusion transformers · Structured pruning · Multimodal generation

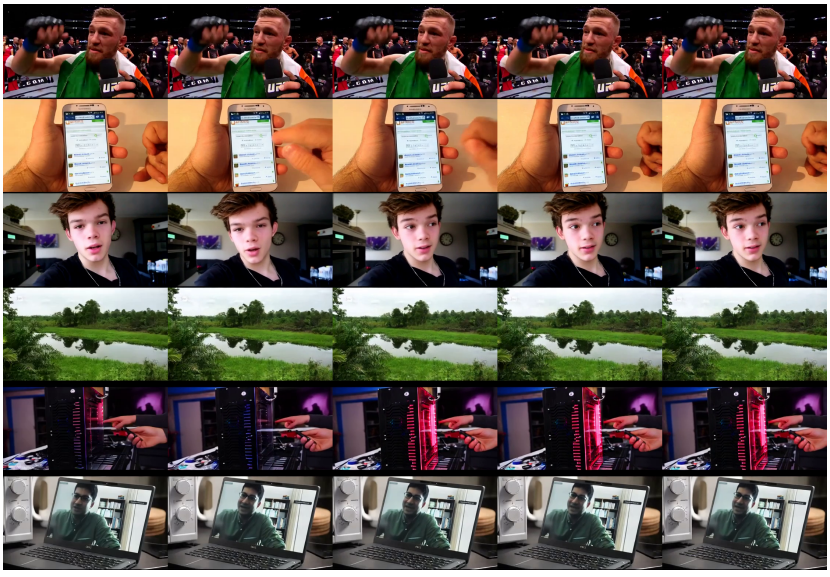

Figure 1: Qualitative video samples generated by our model.

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.

1 INTRODUCTION

FFNs account for a large share of both the parameters and the per-step compute in a video diffusion transformer (Peebles & Xie, 2023; Ma et al., 2025; Yang et al., 2025; Wan Team, 2025). In DaVinci (Chern et al., 2026), a pretrained single-stream audio-video DiT, the shared middle blocks implement these networks as dense SwiGLU layers, so every token incurs the full intermediate width, and the checkpoint contains no experts to route among.

Mixture-of-experts (MOEs) provides conditional model capacity by routing each token to a subset of expert FFNs (Fedus et al., 2022; Jiang et al., 2024; Dai et al., 2024). The same sparse routing has been applied inside diffusion transformers (Park et al., 2024). Sparse routing alone, however, does not remove redundancy within the selected expert, and constructing that expert set from a dense checkpoint (Zhang et al., 2022) is a separate step from narrowing it. Feature caching across denoising steps reduces repeated computation and leaves this expert width unchanged (Zou et al., 2026; Zheng et al., 2025b).

Weight pruning can zero individual parameters while keeping the original matrix shape (Sun et al., 2024; Frantar & Alistarh, 2023), and router-guided pruning applies that idea inside MoE experts (Xie et al., 2024). Dropping whole experts changes which experts the router can select (Lu et al., 2024). TENP (He et al., 2026) instead estimates the contribution of each intermediate neuron by projecting its output onto the complete expert output. However, TENP was developed for language MoEs, whereas an audio-video DiT routes heterogeneous video, audio, and text tokens whose activation distributions and contributions can differ substantially. Directly aggregating these tokens may cause the most frequent modality to dominate the pruning decision. We therefore aggregate neuron contributions separately within video, audio, and text tokens, standardize them within each modality, and only then fuse the modality-wise scores for structured pruning.

CROWN keeps the modality-specific head and tail blocks and restructures the 32 shared middle-block MLPs into three-expert top-1 SparseMoE layers. The router is optimized by the load-balancing term alone, yielding near-uniform expert utilization. EENP then freezes the router and removes low-scoring SwiGLU neurons, storing a wider Quality width and a nested Fast prefix in one checkpoint. Sparse upcycling constructs routed expert capacity, while subsequent width narrowing provides the actual computational reduction. On the converted model, the 60% nominal pruning ratio uses Quality retention 0.65 and Fast retention 0.40, with eight of the 32 denoising steps at the Quality width. VBench-2.0 (Zheng et al., 2025a) overall remains 92.4% of the unpruned SparseMoE score, and DiT denoise latency falls by 22.0% relative to the unpruned SparseMoE checkpoint. The same modality-aware scoring rule also transfers to MAGI-2 Preview (Sand.ai, 2026): at 20% and 40% pruning, CLIP similarity to the unpruned video is 0.900 and 0.865, and at 60% pruning, stored parameters drop to 42.8% of the 114.0B model while CLIPScore (Hessel et al., 2021) retains 99.3% of the unpruned score.

Our contributions are as follows.

(i) We are the first to convert shared middle-block FFNs into token-level SparseMoE experts and perform intra-expert pruning for a single-stream MM DiT.

(ii) We propose EENP, which aggregates and standardizes routed-token contributions within each modality, fuses them across modalities, and physically narrows SwiGLU experts into nested Fast and Quality widths.

(iii) We report the change in VBench 2.0, DiT denoise latency, and paired fidelity to adaptation clips under this pruning.

2 RELATED WORK

2.1 AUDIO-VIDEO DIFFUSION TRANSFORMERS

Latent diffusion models run the denoising process in a compressed latent space (Rombach et al., 2022), and Diffusion Transformers replace the U-Net backbone with transformer layers that scale with model width and depth (Peebles & Xie, 2023). Joint audio-video generation places a second modality on this backbone. DaVinci uses one single-stream DiT over video, audio, and text tokens (Chern et al., 2026), while Ovi fuses twin backbones across modalities (Low et al., 2025). In

the single-stream design, the shared middle blocks still apply a dense feed-forward network to every token at every denoising step, so the full intermediate width is paid on each forward pass.

2.2 From Dense FFNs to Routed Experts

Mixture-of-experts layers raise capacity by activating only a subset of expert feed-forward networks for each token. Switch Transformers established this token-choice pattern for large language models (Fedus et al., 2022), and Mixtral uses the same sparse routing inside a pretrained language-model stack (Jiang et al., 2024). DeepSeekMoE separates shared experts from routed experts so that common computation and expert specialization can coexist (Dai et al., 2024). Switch-DiT carries sparse experts into diffusion transformers and ties them to different denoising tasks (Park et al., 2024). These methods decide which expert runs. MoEfication shows that neurons inside a dense transformer FFN already behave like a mixture and can be regrouped into experts (Zhang et al., 2022), which is the route from a pretrained dense checkpoint to a routed one. After that conversion, the selected expert still executes its full intermediate width.

2.3 Pruning Inside Experts

MoE compression either removes experts or thins their computation. Expert pruning/skipping drops low-contribution experts, changing the router's available set (Lu et al., 2024); Wanda and SparseGPT prune weights one-shot while preserving tensor shapes (Sun et al., 2024; Frantar & Alistarh, 2023); MoE-Pruner sparsifies weights within an expert via the router without narrowing it (Xie et al., 2024). TENP scores intermediate neurons by alignment with the full expert output, fixes the router, and removes low-scoring SwiGLU neurons (He et al., 2026), but its score assumes single-modality language MoEs; in audio-video DiT, video/audio/text tokens differ in count and activation scale, so pooled scores may let the most frequent modality dominate. Separately, diffusion-transformer cost is reduced by reusing computation across denoising steps: dual feature caching stores token-wise features for later steps (Zou et al., 2026), and cluster-driven caching evaluates a selected token subset within a step (Zheng et al., 2025b). Both cut repeated timestep compute but leave each expert FFN's stored width unchanged, the redundancy targeted by intra-expert pruning.

3 Method

CROWN separates construction of the routed model from later compression. Stage A replaces the shared middle-block MLPs with top-1 SparseMoE layers and adapts that checkpoint. Stage B freezes the router and narrows the experts. It leaves the router parameters and the set of available experts unchanged. Token-to-expert assignments may nevertheless change as pruning modifies upstream hidden states.

3.1 Audio-Video Backbone

We build on DaVinci-MagiHuman (Chern et al., 2026), a 15B-parameter, 40-layer single-stream audio-video DiT. Layers $\{0, 1, 2, 3\}$ and $\{36, 37, 38, 39\}$ contain deterministic modality-specific projections and normalization branches for video, audio, and text tokens. The released model tokenizes audio with a 64-channel Stable Audio VAE (Evans et al., 2024). That encoder is trained for audio reconstruction and has no video alignment (Zhan et al., 2026), so we keep the native video VAE and replace only the audio branch with OmniVAE (Zhan et al., 2026), which maps 48 kHz waveforms to 128-channel semantically structured latents. The DiT audio projections are resized to this channel width.

3.2 Dense-to-SparseMoE Conversion

We replace exactly the 32 shared SwiGLU FFNs in layers 4–35 with token-choice SparseMoE layers. Each converted layer contains three routed experts and no shared expert. Figure 2 places this routed feed-forward layer in the shared middle blocks, between the modality-specific early and late layers.

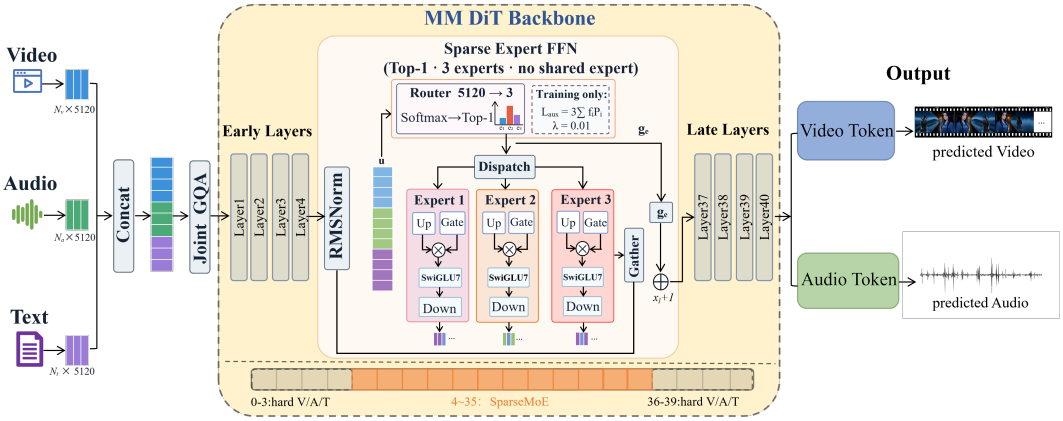

Figure 2: Converted SparseMoE middle block.

Let W_{ug} and W_{down} denote the packed up–gate and down projections of a dense middle-layer FFN. Expert 0 receives an exact copy of both matrices. Experts 1 and 2 receive independently perturbed copies:

$$W_{\text{ug}}^{(e)} = W_{\text{ug}} + \varepsilon_{\text{ug}}^{(e)}, \qquad W_{\text{down}}^{(e)} = W_{\text{down}} + \varepsilon_{\text{down}}^{(e)}, \qquad \varepsilon^{(e)} \sim \mathcal{N}(0, 10^{-6}), \quad e \in \{1, 2\}. \tag{1}$$

The router is newly initialized from $\mathcal{N}(0, 0.01^2)$ and uses no additive routing noise.

For converted layer ℓ, let $f_{\ell,e}$ be the fraction of tokens assigned to expert e and $p_{\ell,e}$ the mean router probability of that expert. The adaptation objective is

$$\mathcal{L}_{\text{conv}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{bal}} \frac{1}{L} \sum_{\ell=1}^{L} E \sum_{e=1}^{E} f_{\ell,e} p_{\ell,e}, \tag{2}$$

where $L = 32$, $E = 3$, and $\lambda_{\text{bal}} = 0.01$. Consequently, the denoising loss updates the expert parameters but provides no gradient to the router through the discrete expert selection. The router parameters are optimized solely by the load-balancing auxiliary loss.

3.3 Modality-Aware Intra-Expert Pruning

Stage B compresses the routed experts at the neuron level with EENP, freezing the router parameters and all non-FFN modules. Figure 3 summarizes this pruning and elastic-inference pipeline.

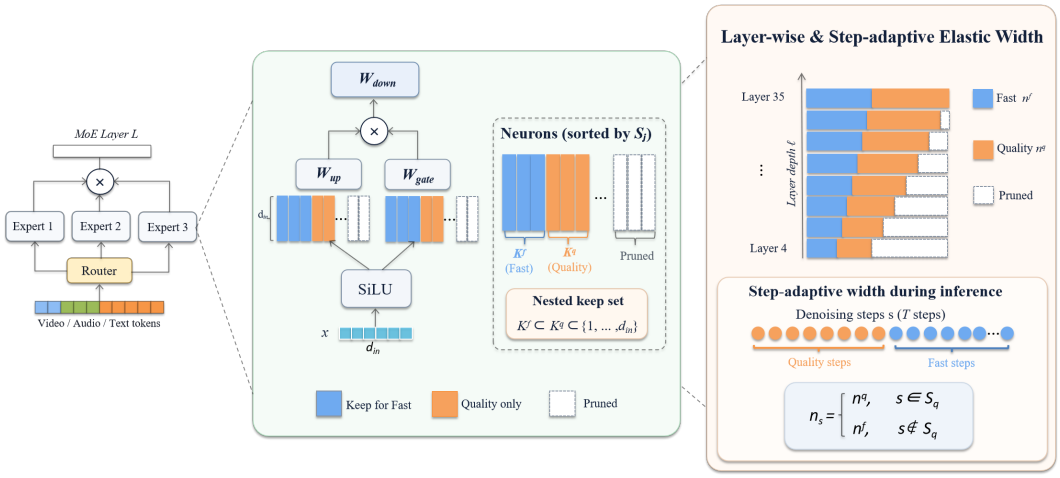

Figure 3: Overview of the EENP pruning algorithm.

At each SparseMoE layer ℓ, we observe only tokens dispatched to expert e. Padding tokens are excluded, and each valid token is assigned a modality label $m \in \mathcal{M} = \{V, A, T\}$ for video, audio, or text.

For a token h routed to expert e, let $z_{e,k}(h)$ denote the k-th SwiGLU activation and write the full expert output as $\overline{h}_e = \sum_{j=1}^{d_{\text{int}}} w^{(e)}_{\text{down},j} z_{e,j}(h)$. Following TENP, the routed projection statistic is

$$q_{e,k}(h) = \frac{\left\langle w^{(e)}_{\text{down},k} z_{e,k}(h),\ \overline{h}_e \right\rangle}{\left\|\overline{h}_e\right\|_2 + \epsilon}. \tag{3}$$

Pruning neuron k leaves the residual width d unchanged: it removes row k of $W^{(e)}_{\text{up}}$ and $W^{(e)}_{\text{gate}}$ together with column k of $W^{(e)}_{\text{down}}$.

For a routed token h, we record the intermediate activation $z_e(h)$ and accumulate the token-level statistic in Eq. (3) into a separate contribution vector for every layer, expert, and modality:

$$C_{\ell,e,m,k} = \sum_{h \in \mathcal{D}_m} \mathbb{I}[e^\star_\ell(h) = e]\, q_{\ell,e,k}(h), \qquad k = 1, \ldots, d_{\text{int}}, \tag{4}$$

where $\mathcal{D}_m$ contains calibration tokens of modality m. In addition to the sequence-level sums in Eq. (4), we accumulate the same statistics in B temporal bins, denoted by $C^{(b)}_{\ell,e,m,k}$, together with the number of routed tokens $N^{(b)}_{\ell,e}$.

We remove imbalance by standardizing each non-empty modality across the intermediate dimension and fusing the standardized scores:

$$S_{\ell,e,k} = \sum_{m \in \mathcal{M}_{\ell,e}} \lambda_m \frac{C_{\ell,e,m,k} - \mu_{\ell,e,m}}{\sigma_{\ell,e,m} + \epsilon}, \tag{5}$$

where $\mu_{\ell,e,m}$ and $\sigma_{\ell,e,m}$ are the mean and standard deviation of $\{C_{\ell,e,m,k}\}_{k=1}^{d_{\text{int}}}$, and $\mathcal{M}_{\ell,e}$ contains the modalities with at least one valid routed observation for expert e. Empty modalities are skipped rather than imputed from another modality. We use $(\lambda_V, \lambda_A, \lambda_T) = (3, 1, 1)$ by default. Appendix C shows a lower video-token expert-output error than equal fusion, while the paired fidelity difference lies inside the 95% interval. If an expert receives no valid calibration token, we use the sum of the paired up-gate row norm and the corresponding down-projection column norm as a deterministic fallback score.

We assign a common width to all experts within a layer while allowing this width to increase with depth. Let L be the number of converted SparseMoE layers ordered from shallow to deep, and let $a \in [0, 1)$ control the allocation slope. For a target average retention ratio ρ, the retained width of layer ℓ is

$$n_\ell(\rho) = \text{clip}\left(\text{round}_4\left(\frac{\rho\, x_\ell}{L^{-1} \sum_{j=0}^{L-1} x_j}\, d_{\text{int}} \right), n_{\min}, d_{\text{int}} \right), \tag{6}$$

where $\ell = 0, \ldots, L-1$, $x_\ell = 1 - a + 2a\,\ell/(L-1)$, and round_4 aligns the width for dense execution. We adjust aligned widths after clipping so that their sum matches the total width prescribed by ρ.

To support elastic inference with one checkpoint, we instantiate Eq. (6) at two average retention ratios, ρ_q and ρ_f, where $\rho_f \le \rho_q$. This produces a Quality width $n^q_\ell = n_\ell(\rho_q)$ and a Fast width $n^f_\ell = n_\ell(\rho_f)$ for each layer. For expert e, let $\pi_{\ell,e}$ be the neuron indices sorted by the fused score $S_{\ell,e,k}$ in descending order. The Quality keep-set $\mathcal{K}^q_{\ell,e}$ is the first n^q_ℓ indices of $\pi_{\ell,e}$, and the Fast keep-set $\mathcal{K}^f_{\ell,e}$ is its prefix of length n^f_ℓ, so $\mathcal{K}^f_{\ell,e} \subseteq \mathcal{K}^q_{\ell,e}$.

We estimate the importance of temporal bin b by summing, over layers and experts, the product of the mean absolute contribution and $\max(N^{(b)}_{\ell,e}, 1)$:

$$M_b = \sum_{\ell,e} \text{mean}_{m,k} \left| C^{(b)}_{\ell,e,m,k} \right| \max\left(N^{(b)}_{\ell,e}, 1 \right). \tag{7}$$

Starting from the bins with the largest M_b, we add their corresponding denoising steps to the Quality set $\mathcal{S}_q$ until that set reaches its preset size. The reported sweeps use eight Quality steps out of 32. At inference, layer ℓ uses width n^q_ℓ on steps $s \in \mathcal{S}_q$ and the Fast prefix n^f_ℓ on the remaining steps.

Finally, we physically materialize each expert at its Quality width by retaining, for the ordered index sequence $\mathcal{K}^q_{\ell,e}$, the corresponding rows of $W^{(\ell,e)}_{\text{up}}$ and $W^{(\ell,e)}_{\text{gate}}$ and the corresponding columns of $W^{(\ell,e)}_{\text{down}}$.

4 EXPERIMENTS

Stage A adapts the dense backbone and the converted SparseMoE models at a learning rate of 5×10^{-5} on 4×8 NVIDIA H200 GPUs with batch size 4. Load balancing alone updates the router. MoE32 converts layers 4–35 and MoE16 the even middle layers, on about one thousand shared clips. Dense uses its best checkpoint; converted models use the final one.

EENP freezes the router, scores neurons on 32 adaptation clips, and prunes the experts. Ratio r sets Fast retention to $1 - r$ and Quality retention to $\min(1, 1 - r + 0.25)$, with eight Quality steps and twenty-four Fast steps.

We report VBench-2.0 at 448×256, 4 seconds, 32 denoising steps, and classifier-free guidance 2. DiT latency is the mean wall-clock time of 128 clips of 5 s after one warmup, with VAE decoding excluded, on one NVIDIA H200 in FP32 with batch size 1 (PyTorch 2.9.0, CUDA 12.8). Paired evaluations use 5 s clips.

4.1 COMPRESSION QUALITY AND EFFICIENCY

Stage A: Dense-to-SparseMoE conversion. We evaluate Stage A as a conversion check rather than as an efficiency result. The dense configuration contains 15.300B trainable parameters, whereas the converted checkpoint contains 28.721B parameters, comprising 8.571B non-expert backbone parameters, 20.131B routed-expert parameters, and 0.018B text-encoder parameters. Top-1 routing activates one expert per token, so the active intermediate width remains $d_{\text{int}} = 13{,}652$ even though the three experts increase the stored capacity.

The three experts receive token fractions $[0.332, 0.335, 0.333]$, and the mean top-1 assignment entropy is 1.0927, near $\log 3 = 1.0986$. These batch-level diagnostics match the load-balancing term in Eq. (2) and do not establish task-driven routing.

The Dense row of Table 4 reports DaVinci's VBench-2.0 overall under this Enhanced Prompt protocol. Stored and active counts for the dense backbone, this checkpoint, and 60% EENP are in the same table.

Stage B: Intra-expert pruning trade-off. Following Section 3.3, a ratio r sets the Fast retention to $1 - r$ and the Quality retention to $\min(1, 1 - r + 0.25)$. Eight denoising steps use the Quality width; the remaining twenty-four steps evaluate only the Fast prefix. Appendix D reports a nominal-pruning retention comparison of this nested schedule against fixed-width ENP on paired-frame PSNR.

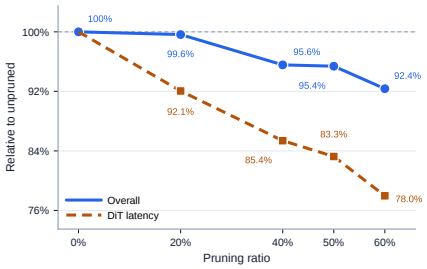

Figure 4: VBench-2.0 overall score and DiT latency of EENP.

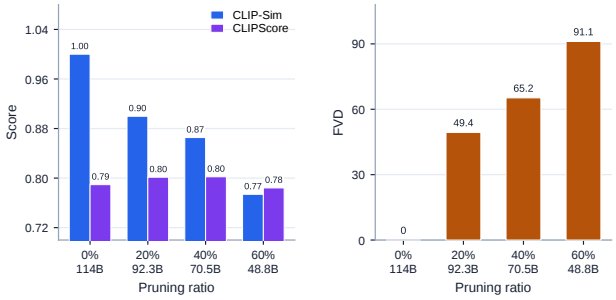

Figure 5: EENP on MAGI-2 Preview. Left: CLIP-Sim and CLIPScore. Right: FVD to the unpruned video.

Table 1: VBench-2.0 scores and DiT latency across pruning ratios.

Ratio	ρ_q/ρ_f	Creat.	Comm.	Ctrl.	HF	Phys.	Overall	DiT (s)
0%	1.00/1.00	0.232	0.577	0.220	0.977	0.588	0.519	26.83
20%	1.00/0.80	0.234	0.563	0.210	0.971	0.606	0.517	24.70
40%	0.85/0.60	0.222	0.560	0.192	0.971	0.533	0.496	22.91
50%	0.75/0.50	0.236	0.543	0.176	0.972	0.548	0.495	22.34
60%	0.65/0.40	0.210	0.534	0.148	0.972	0.531	0.479	20.92

At a 60% nominal pruning ratio, the overall score remains 0.479, or 92.4% of the unpruned SparseMoE model, while DiT denoising time falls from 26.83 s to 20.92 s, corresponding to a 22.0% re-

duction relative to the unpruned SparseMoE checkpoint. From 20% to 60%, the overall score stays between 0.517 and 0.479 while latency decreases at every ratio. Human Fidelity remains 0.972 at 60%, compared with 0.977 at 0%.

At 60%, mean denoising time falls by 5.911s, a $1.283\times$ speedup, and stored routed-expert parameters drop from 20.131B to 13.083B with eight Quality steps and twenty-four Fast steps. EENP obtains this saving by executing the retained neurons as a narrower dense GEMM. These times compare the unpruned SparseMoE checkpoint with EENP over the 20%–60% pruning ratios.

4.2 Ablation Studies

We compare the two encoders on the dense backbone. IS is the PANNs Cnn14 score on generated audio (Kong et al., 2020; Salimans et al., 2016). FD is the CLAP Fréchet distance to the source audio (Kilgour et al., 2019; Wu et al., 2023). KL is the PaSST divergence to the paired clip (Koutini et al., 2022). Dense, MoE16, and MoE32 are adapted on the same audio-video adaptation set

Table 2: Audio-encoder ablation on the dense backbone.

Audio encoder	IS↑	FD↓	KL↓
SA VAE	**1.283**	1.311	0.342
OmniVAE	1.185	**0.654**	**0.21**

containing on the order of one thousand clips, with OmniVAE under a shared training schedule at a learning rate of 5×10^{-5}. MoE16 converts the 16 even layers $\{4, 6, \ldots, 34\}$ and leaves the odd middle layers as dense SwiGLU blocks. We score Dense at its best checkpoint, and MoE16 and MoE32 at the final adaptation checkpoint. We report paired fidelity on a fixed controlled diagnostic subset drawn from this adaptation set using seed 20260914. Sampling is text-only at 448×256 for 5 s, with 32 denoising steps, classifier-free guidance 2, and seed 42 plus the clip slot.

Table 3: Paired fidelity of Dense, MoE16, and MoE32 on the 32-clip diagnostic subset.

Model	FVD↓	CLIP-Sim↑	SSIM↑	LPIPS↓	V-LPIPS↓
Dense	80.068	0.9063	0.6286	0.2852	0.0044
MoE16	66.002	0.9138	**0.6923**	0.2599	0.0032
MoE32	**64.267**	**0.9186**	0.6858	**0.2481**	**0.0029**

On this controlled diagnostic subset, MoE32 attains the lowest FVD and the best CLIP-Sim, LPIPS, and V-LPIPS, while MoE16 attains the highest SSIM. All methods are evaluated with identical captions, seeds, resolution, duration, and sampler, so these paired comparisons isolate the Stage-A conversion effect. We treat FVD as a secondary paired diagnostic, and the frame-wise metrics provide complementary evidence. These are not VBench-2.0 results; Table 6 reports that benchmark under a separate protocol and adaptation set.

Active size counts one expert under top-1 routing; at 60% it uses the schedule-averaged width of Section 4.1 (eight Quality and twenty-four Fast steps).

Table 4: Parameters and scores across stages.

Stage	Stored (B)	Active (B)	VBench↑	DiT (s)↓	FVD↓	CLIP-Sim↑
Dense	15.300	15.300	0.504	23.99	80.068	0.9063
SparseMoE	28.721	15.299	0.519	26.83	64.267	0.9186
EENP (60%)	21.673	11.692	0.479	20.92	121.012	0.8227

In Table 4, the FVD and CLIP-Sim entries are paired scores on the same 32-clip diagnostic subset used in Table 3, whereas the VBench and DiT columns follow the protocols defined in Section 4. The Dense VBench entry is DaVinci's VBench-2.0 overall score. On that subset, 60% EENP moves FVD from 64.267 to 121.012 and CLIP-Sim from 0.9186 to 0.8227.

4.3 Contextual Comparison and Paired Fidelity

Comparison with open generators. Table 5 compares the unpruned converted model and DaVinci with LTX-2.3 (HaCohen et al., 2026), Wan2.2 (Wan Team, 2025), and Ovi 1.0 (Low et al., 2025).

The row Ours is that unpruned converted model. DaVinci is the dense OmniVAE backbone after the same Koala adaptation, at the same resolution, duration, seed, and sampler. Because our model is adapted on these Koala clips whereas the external generators are not, this comparison is intended to provide contextual reference rather than a controlled held-out leaderboard. IS is the frame-wise Inception Score (Salimans et al., 2016) from an ImageNet Inception-v3 classifier (Szegedy et al., 2016), distinct from the PANNs audio IS in Section 4.2. CLAP L2 is the mean L2 distance between CLAP embeddings of the generated soundtrack and the paired source audio. Wan2.2 produces no audio. Ovi 1.0 remains slightly ahead on CLAP L2, and Wan2.2 remains ahead on CLIPScore.

Table 5: Comparison with open-source models on 32 Koala clips.

				Video Quality			Audio Quality	
Model	Architecture	Year	Params	IS↑	CLIPScore↑	CLIP-Sim↑	CLAP L2↓	KL↓
Ours	Single-stream	–	(28.7B/15.3B)	**3.64**	0.757	**0.93**	1.006	**0.618**
DaVinci	Single-stream	2026	15.3B	3.50	0.730	0.765	1.384	3.122
LTX-2.3	Dual-stream	2026	22B	3.32	0.692	0.629	1.258	1.711
Wan2.2	Video-only	2025	14B	3.57	**0.781**	0.77	–	–
Ovi 1.0	Dual-stream	2025	11B	3.50	0.76	0.765	**0.995**	1.442



Figure 6: Comparison of generation quality between our model and other models.

Paired fidelity after compression. Table 6 compares pruning methods under a controlled paired-fidelity protocol after adapting the SparseMoE DiT on an adaptation subset of approximately one thousand Koala-36M clips (Wang et al., 2025).[1] We measure pruning-induced fidelity on a fixed controlled diagnostic set using a paired evaluation protocol. All methods use identical captions and sampling seeds and generate 5 s videos at 448×256 with 32 denoising steps and classifier-free guidance 2. Neuron statistics are estimated from a calibration subset of this set. Frames are decoded at 30 fps and aligned by truncating each pair to the shorter sequence. This protocol is designed to isolate changes introduced by pruning rather than to evaluate held-out generalization.

CLIPScore (Hessel et al., 2021) measures text–video alignment between generated frames and the caption. CLIP-Sim (Radford et al., 2021) is a different quantity: semantic similarity between each generated frame and the paired adaptation frame. Table 6 uses that paired clip as the reference video. FVD (Unterthiner et al., 2018) is the distribution distance between the generated set and these adaptation clips, so it is not grouped with the frame-wise reference-video scores. SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and V-LPIPS (Chu et al., 2020) compare each generated frame with the same reference frame. The table compares pruning-induced quality under matched nominal sparsity rather than matched runtime or active width. EENP uses the nested Quality and Fast widths of Section 3.3; Wanda (Sun et al., 2024), Random, and TENP use a constant width. TENP reuses the EENP projection scores. Wanda ranks neurons by activation-weighted column

[1]Dataset and code: `https://github.com/KlingAIResearch/Koala-36M`.

magnitude, distinct from the original unstructured Wanda mask. SparseGPT (Frantar & Alistarh, 2023) zeros 60% of the weights and keeps the original GEMM shape. Runtime for the narrowed GEMMs is reported in Table 1.

Table 6: Text–video alignment, distribution quality, and reference-video fidelity on the Koala adaptation subset at 40% and 60% nominal pruning. Within each pruning ratio, the best score among pruned methods is in bold.

		Text–Video Alignment	Distribution Quality	Reference Video Fidelity			
Ratio	Method	CLIPScore↑	FVD↓	CLIP-Sim↑	SSIM↑	LPIPS↓	V-LPIPS↓
0%	Unpruned	0.7493	48.856	0.9305	0.7194	0.1750	0.0054
40%	EENP (ours)	0.7480	**69.564**	**0.9109**	**0.6746**	**0.2213**	**0.0061**
	Wanda	**0.7511**	76.104	0.8914	0.6294	0.2646	0.0070
	TENP	0.7508	76.663	0.8949	0.6429	0.2536	0.0067
	SparseGPT	0.7492	78.200	0.9007	0.6452	0.2373	0.0071
	Random	0.7449	92.487	0.8774	0.6003	0.2930	0.0094
60%	EENP (ours)	**0.7403**	**122.133**	**0.8439**	**0.5656**	**0.3382**	**0.0093**
	TENP	0.7296	132.550	0.8081	0.4950	0.4259	0.0117
	Wanda	0.7185	149.632	0.7904	0.4583	0.4569	0.0125
	Random	0.7328	155.700	0.7757	0.4289	0.4899	0.0169
	SparseGPT	0.5834	211.946	0.6469	0.4134	0.6179	0.0218

At both nominal ratios, EENP ranks first on FVD and on every reference-video column. At 40%, Wanda is slightly higher on CLIPScore (0.7511 vs. 0.7480). At 60%, EENP keeps the highest pruned CLIPScore (0.7403), while SparseGPT falls to 0.5834. CLIPScore measures agreement with the caption; FVD and the reference-video columns measure a shift from the adaptation clips. Reference-free generation quality is the VBench-2.0 evaluation in Section 4.1 and Appendix B.

4.4 Generalization to MAGI-2 Preview

We apply the same scoring rule to MAGI-2 Preview (114B, about 6B active per token), pruning only the preview transformer at 272p, 10 s, and 32 steps, without the refiner. Each ratio keeps one layer-uniform width with $(\lambda_V, \lambda_A, \lambda_T) = (3, 1, 1)$, because experts are padded to the widest keep-set in the layer. Calibration uses 32 held-out prompts, disjoint from 32 Koala-36M captions decoded at 12.5 fps. Only routed expert projections are sliced. Figure 5 reports 16-frame FVD, CLIP-Sim, and LPIPS against the unpruned video of the same prompt. Parameter counts at 20%, 40%, and 60% are 92.3B, 70.5B, and 48.8B (81.0%, 61.8%, and 42.8% of 114.0B). CLIP-Sim is 0.8998, 0.8654, and 0.7738; LPIPS is 0.3101, 0.3925, and 0.5275; FVD is 49.362, 65.217, and 91.095. CLIPScore is 0.8014, 0.8021, and 0.7841, against 0.7894 unpruned.

5 Conclusion

We present CROWN, a compressed routed audio-video diffusion transformer that restructures the shared middle-block FFNs of a pretrained single-stream DiT into top-1 SparseMoE experts and then narrows them with EENP. EENP aggregates and standardizes routed-token contributions within each modality, fuses them into a unified neuron-importance score, and physically removes low-scoring SwiGLU neurons.

This structured width narrowing retains 92.4% of the unpruned SparseMoE VBench-2.0 score and reduces DiT denoising latency by 22.0% (1.283× speedup). The same scoring rule transfers to MAGI-2 Preview: at 20% and 40% pruning, outputs remain 0.900 and 0.865 CLIP-similar to the unpruned video, and at 60% pruning CLIPScore retains 99.3% of the unpruned score while stored parameters fall to 42.8%. Sparse upcycling provides routed expert capacity, whereas the actual computational reduction comes from physically narrowing the experts. Near-uniform expert utilization indicates that the converted router remains well balanced.

REFERENCES

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics*, 39(4):75:1–75:13, 2020. doi: 10.1145/3386569.3392457.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, 2024. doi: 10.18653/v1/2024.acl-long.70.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10323–10337, 2023.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Jiangyang He, Shaolin Zhu, and Deyi Xiong. TENP: Trapezoidal expert neuron pruning for mixture-of-experts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20911–20925, 2026. doi: 10.18653/v1/2026.findings-acl.1049.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, pp. 2753–2757, 2022. doi: 10.21437/Interspeech.2022-227.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, 2024. doi: 10.18653/v1/2024.acl-long.334.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Byeongjun Park, Hyojun Go, Jin-Young Kim, Sangmin Woo, Seokil Ham, and Changick Kim. Switch diffusion transformer: Synergizing denoising tasks with sparse mixture-of-experts. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, 2024. doi: 10.1007/978-3-031-73668-1_27.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

Sand.ai. MAGI-2 preview: Scaling video generation models efficiently. `https://github.com/SandAI-org/MAGI-2-preview`, 2026. Technical blog: `https://sand.ai/blog/magi-2-preview`.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8428–8437, 2025. URL `https://github.com/KlingAIResearch/Koala-36M`.

Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. doi: 10.1109/MSP.2008.930649.

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. doi: 10.1109/TIP.2004.819861.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Yanyue Xie, Zhi Zhang, Ding Zhou, Cong Xie, Ziang Song, Xin Liu, Yanzhi Wang, Xue Lin, and An Xu. MoE-Pruner: Pruning mixture-of-experts large language model using the hints from its router. *arXiv preprint arXiv:2410.12013*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Jun Zhan, Chen Yang, Yitian Gong, Donghua Yu, Kuangwei Chen, Wenbo Zhang, Kexin Huang, Qi Luo, Zhe Xu, Ying Zhu, Jin Wang, Tengyue Zhang, Qi Chen, Zhiyu Zhang, Cheng Chang, Songlin Wang, Junqi Dai, Jiasheng Ye, Xiaogui Yang, Tianyi Liang, Xiangyu Peng, Zhaoye Fei, Shimin Li, Qinyuan Cheng, Xie Chen, Xinchi Chen, and Xipeng Qiu. OmniVAE: An audio-video VAE with cross-modal alignment for joint generation. *arXiv preprint arXiv:2607.23855*, 2026.

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*, pp. 586–595, 2018.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. MoEfication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 877–890, 2022. doi: 10.18653/v1/2022.findings-acl.71.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

Zhixin Zheng, Xinyu Wang, Chang Zou, Shaobo Wang, and Linfeng Zhang. Compute only 16 tokens in one timestep: Accelerating diffusion transformers with cluster-driven feature caching. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b. doi: 10.1145/3746027.3755479.

Chang Zou, Shikang Zheng, Evelyn Zhang, Runlin Guo, Haohang Xu, Zhengyi Shi, Conghui He, Xuming Hu, and Linfeng Zhang. Rethinking token-wise feature caching: Accelerating diffusion transformers with dual feature caching. *IEEE Transactions on Image Processing*, 2026. doi: 10.1109/TIP.2026.3698363.