# VIRTUAL TOKENS, REAL SUPERVISION: TASK-AWARE EXTRAPOLATION FOR VISION ROPE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

To improve generalization beyond the training scale, recent methods introduce Vision RoPE into vision models. However, at substantially larger spatial or spatiotemporal resolutions, existing extrapolation methods typically optimize from only one of two signals: source-scale task supervision or target-grid geometry. Source-supervised methods do not observe the target scale, whereas target-aware post-hoc corrections usually rely on geometric or statistical heuristics without target-conditioned task supervision. Their inability to use both signals simultaneously limits the effectiveness of extrapolation optimization, especially at large resolutions. We present **Virtual Target Calibration** (VTC), a source-only post-hoc method that constructs a target-conditioned surrogate from source-scale inputs. VTC stretches RoPE coordinates and assigns analytic multiplicities to visual keys to model target-scale phase changes and enlarged softmax competition. Source-task Fisher weights then guide the joint calibration of a block-unitary frequency mixer and a positive attention inverse-temperature factor. Calibration is performed at a small set of target anchors and interpolated to unseen resolutions while the backbone remains frozen; classification calibration takes about five minutes on RTX 4090 hardware, with at most 200 updates per anchor. Across image classification, high-resolution image generation, and long-video generation, VTC improves extrapolation over native inference and is competitive with or better than heuristic post-hoc corrections, without target-scale inputs or labels.

## 1 INTRODUCTION

Vision models are routinely evaluated on token grids larger than the grid used during training. This occurs when an image generator is asked for a higher resolution, when a video model must produce a longer sequence, or when a classifier is applied to a larger crop. Directly evaluating a pretrained Vision RoPE model on the enlarged grid changes its relative phases and its attention normalization, and can cause structural degradation, repetitive patterns, or loss of fine detail (Su et al., 2024; Heo et al., 2024; Lu et al., 2024b; Yang et al., 2025).

Existing extrapolation approaches expose only one side of the problem, as illustrated in Figure 1(a). RoPE design methods learn spatial or spectral priors from source-resolution task supervision, but do not observe the deployment grid (Heo et al., 2024; Li et al., 2025; Min et al.; Liu et al., 2026a; Li et al., 2026). Post-hoc methods know the requested target geometry, but their corrections are usually derived from coordinate, frequency, entropy, or stage statistics rather than a target-conditioned task signal (Chen et al., 2024; Peng et al., 2024; Lu et al., 2024b; Zhao et al., 2025a; Issachar et al., 2026; Zhao et al., 2025b; 2026; Liu et al., 2026b; Yan et al., 2026; Rajabi et al., 2026). The key question is whether target-scale effects can be estimated using only source-scale inputs while keeping calibration cost low.

Multi-resolution training can provide direct target supervision, but its cost scales with the number of tokens and is particularly high for large images and long videos (Beyer et al., 2023; Dehghani et al., 2023). VTC instead calibrates a small set of target resolutions as anchors and interpolates the calibration parameters between them. This turns a potentially resolution-by-resolution procedure into a small number of short source-only runs.

For an anchor resolution, VTC forms a virtual target state from a source-scale input, as illustrated in Figure 1(b). It stretches the source RoPE coordinates to the target positional span and assigns



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.

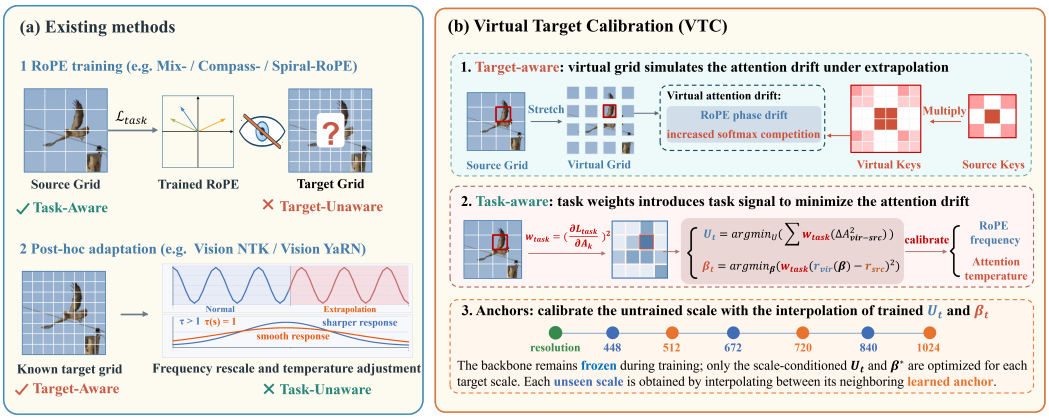


Figure 1: Overview of VTC. (a) Existing approaches typically use either source-task supervision or target-grid heuristics, so they cannot optimize the model with both target-grid information and task supervision at a large resolution. (b) VTC combines a virtual target state with source-task Fisher calibration: stretched coordinates and analytic key multiplicities model the target grid, while a frequency mixer and inverse-temperature factor are calibrated at a few anchors and interpolated to unseen resolutions with the backbone frozen.

each visual key an analytic multiplicity that represents the larger number of target micro-keys competing in the softmax. The virtual state changes the target geometry and key population without synthesizing target-scale features or materializing target tokens. Source-task gradients then define Fisher weights, so calibration focuses on attention entries that affect the frozen task objective. We jointly learn a frequency mixer and an inverse-temperature factor; the latter compensates for the change in task-signal transmission after softmax when the represented key population grows.

The backbone and task head remain frozen throughout calibration. Each anchor uses at most 200 source-scale updates; for classification, one backbone can be calibrated in about five minutes on RTX 4090 hardware, and interpolated resolutions require no additional optimization. We evaluate the same protocol on image classification, high-resolution image generation, and video-length extrapolation.

Our contributions are:

- **Source-only target calibration.** We introduce a virtual target state that combines endpoint-preserving RoPE coordinate stretching with analytic key multiplicities, and use source-task Fisher weights to calibrate frequency geometry and post-softmax response without target-scale inputs or labels.
- **Anchor interpolation for unseen resolutions.** We learn scale-conditioned calibration parameters at a small set of target anchors and interpolate them in log-resolution space, providing a continuous calibration rule within the anchor range while keeping the backbone frozen.
- **Low-cost cross-task extrapolation.** Each anchor requires at most 200 updates; classification calibration takes about five minutes on RTX 4090 hardware. Experiments across multiple RoPE designs and spatial and temporal generation settings show consistent improvements over native extrapolation and competitive performance relative to post-hoc baselines.

## 2 Related Work

We restrict this review to RoPE on regular image and video token grids and to methods directly relevant to spatial or temporal extrapolation. We do not aim to survey positional encodings for camera-aware, panoramic, point-cloud, or cross-modal alignment settings.

**Vision RoPE design.** RoPE represents relative displacement through coordinated rotations of query and key channels (Su et al., 2024). RoPE-ViT studies axial and mixed two-dimensional frequency layouts, while later designs explore isotropic directions, multi-directional projections, balanced higher-dimensional spectra, and head-specific transformations (Heo et al., 2024; Min et al.; Liu et al., 2026a; Li et al., 2026; 2025). These methods determine the positional basis learned during source-scale training, but source-scale supervision does not directly expose the model to all larger grids encountered at deployment.

**Post-hoc extrapolation.** A complementary line adapts a trained model after the target grid is specified. Position Interpolation rescales coordinates, while YaRN combines frequency interpolation with attention-temperature scaling (Chen et al., 2024; Peng et al., 2024). For image and diffusion models, FiT supports flexible resolutions, while DyPE, UltraImage, and ExtraVAR adapt positional or attention corrections according to denoising spectra, entropy, or generation stages (Lu et al., 2024b; Issachar et al., 2026; Zhao et al., 2025b; Yan et al., 2026). For temporal extrapolation, RIFLEx, UltraViCo, and TIDE use intrinsic frequency, attention-decay, or timestep-dependent temperature corrections; SEGA conditions RoPE scaling on the latent spatial spectrum (Zhao et al., 2025a; 2026; Liu et al., 2026b; Rajabi et al., 2026). These methods provide useful target-grid corrections, but their correction signals are generally derived from geometric, frequency, temporal, entropy, stage, or spectral statistics rather than task sensitivity under a virtual target state.

**Multi-resolution training.** Another approach exposes a model to multiple resolutions during training. FlexiViT randomizes patch sizes to support different computational scales, while NaViT trains with variable-resolution sequences and native aspect ratios (Beyer et al., 2023; Dehghani et al., 2023). Such methods provide direct exposure to diverse grids, but repeatedly processing large token populations can make the accessible training range expensive. VTC instead combines target-aware virtual coordinates and analytic key multiplicities with source-task supervision, and calibrates the resulting parameters at a small set of anchors before interpolating them to unseen resolutions.

**Transformer backbones for generation.** Transformers have become standard backbones for diffusion and image/video generation, including U-ViT, DiT, PixArt-$\alpha$, and DiffiT (Bao et al., 2023; Peebles & Xie, 2023; Chen et al., 2024; Hatamizadeh et al., 2024). Recent systems further scale these architectures to video and flexible resolution generation, including VDT, SANA, and Lumina-T2X (Lu et al., 2024a; Xie et al., 2025; Peng et al., 2025). Their token grids grow directly with image resolution or video length, making post-hoc positional extrapolation especially important for these models.

## 3 Preliminaries for Virtual Target Calibration (VTC)

### 3.1 Rotary Position Embedding on Visual Token Grids

For a visual-token lattice $\mathcal{P}$, let $p_i \in \mathbb{R}^r$ be the coordinate of token $i$, and let $q_i, k_j \in \mathbb{R}^{d_h}$ be the unrotated query and key of one attention head. RoPE rotates each two-channel feature block according to the coordinate,

$$R(p) = \mathrm{blkdiag}_{m=1}^{d_h/2} \left[\mathrm{Rot}(\omega_m^\top p)\right]. \tag{1}$$

The relative rotation gives the logit

$$a_{ij}(\mathcal{P}) = \frac{q_i^\top R(p_j - p_i) k_j}{\sqrt{d_h}} = \frac{1}{\sqrt{d_h}} \mathrm{Re} \sum_m c_{ijm} e^{\mathrm{i}\omega_m^\top (p_j - p_i)}, \tag{2}$$

where $c_{ijm}$ is the complex query-key interaction in channel $m$. Thus, changing the coordinate span changes relative phases even when the unrotated token features are held fixed.

### 3.2 Coordinate Change and Key-Population Change

Let $\Delta p_{ij}^0$ and $\Delta p_{ij}^s$ denote source and target relative coordinates. Holding the content interaction fixed, the positional change is

$$\delta a_{ij}^{\mathrm{pos}} = \frac{1}{\sqrt{d_h}} \mathrm{Re} \sum_m c_{ijm} \left( e^{\mathrm{i}\omega_m^\top \Delta p_{ij}^s} - e^{\mathrm{i}\omega_m^\top \Delta p_{ij}^0} \right). \tag{3}$$

Higher-frequency channels are more sensitive to the same coordinate change. This expression isolates the positional effect; an actual target forward can also change the unrotated features.

For a fixed query, adding comparable target keys changes the softmax normalizer. If $Z_0$ is the source partition function and $Z_{\mathrm{new}}$ is the contribution of additional keys, then

$$\pi_j^+ = \frac{\pi_j^0}{1+\rho}, \qquad \rho = \frac{Z_{\mathrm{new}}}{Z_0}. \tag{4}$$

More generally, if a source key represents $\mu_j$ exchangeable micro-keys, its group mass is proportional to $\mu_j e^{a_j}$. This key-population effect changes the attention response and the downstream weighted value sum. The two effects can interact in a real forward; VTC models them with separate but jointly optimized calibration terms.

# 4 Virtual Target Calibration

## 4.1 Method Overview

Let $\mathcal{P}_0$ denote the visual-token grid used to train a frozen transformer $f_\theta$, and let $\mathcal{P}_s$ be a larger target grid. Extrapolation changes both the RoPE phase range and the number of keys that compete in each attention softmax. Virtual Target Calibration (VTC) builds a source-only surrogate of this target state: it stretches source coordinates, represents the enlarged key population with analytic multiplicities, and calibrates a frequency correction together with a positive inverse-temperature factor. Only these calibration parameters are optimized; the backbone and task head remain frozen.

VTC has two coupled components. The virtual state is target-aware because it encodes the target positional span and key population on source features. The source task derivatives are task-aware because they identify which attention responses affect the frozen downstream objective. Response matching is part of this task-aware calibration, rather than a separate third module.

## 4.2 Target-aware Virtual State

For a source patch coordinate $p_d \in \{0, \dots, n_{0,d} - 1\}$, VTC maps it to the endpoint-preserving virtual target coordinate

$$\widetilde{p}_d^{(s)} = p_d \frac{n_{s,d} - 1}{n_{0,d} - 1}. \tag{5}$$

We apply this mapping to both spatial axes for images and to the temporal axis for video-length extrapolation. The virtual pass reuses the source feature tensor and changes only the coordinates supplied to RoPE; it does not synthesize or materialize target-scale features. Let $a^0_{rj}$ be a native source logit and $a^{\mathrm{V}}_{rj}(U_s)$ the corresponding virtual logit after applying the frequency correction $U_s$.

A visual source key represents a group of target micro-keys. For an exact content-preserving duplication, the multiplicity is $|G_j|$. In the implementation, we use the analytic visual multiplicity

$$\mu_j = \frac{N_s^{\mathrm{vis}}}{N_0^{\mathrm{vis}}} \quad \text{for visual keys,} \qquad \mu_j = 1 \quad \text{for class, register, text, or local keys,} \tag{6}$$

Register tokens are treated as non-expanding keys, consistent with their role as auxiliary tokens in vision transformers (Darcet et al., 2024). which may be non-integer and represents the expected number of target micro-keys represented by one source key. The virtual softmax therefore uses $\mu_j$ in its partition function without constructing a target-scale feature tensor. Under exchangeable micro-keys this grouped construction is exact; with within-group phase, content, or downstream-score drift it is an approximation. The associated bounds are given in Appendix A.

A controlled content-preserving experiment visualizes this approximation in Figure 2. The virtual response follows the spatial pattern of the explicit target more closely than the source-native response in the shown setting. This experiment isolates coordinate stretching and key multiplicity by copying target micro-key content from the source; it does not test feature drift in a real target forward. The real-model diagnostic is reported in Section 5.2.

## 4.3 Task-aware Fisher Calibration

At each target anchor, VTC first evaluates the identity virtual state ($U_s = I$, $\beta_s = 1$) on source-scale inputs. Let $r$ index an attention row and let $\mathcal{V}$ denote its visual keys. The virtual task gradient and its detached empirical-Fisher weight are

$$g^{\mathrm{V}}_{rj} = \frac{\partial \mathcal{L}_{\mathrm{task}}}{\partial a^{\mathrm{V}}_{rj}}, \qquad w^{\mathrm{V}}_{rj} = \mathrm{sg}\left( \min\left\{ \frac{(g^{\mathrm{V}}_{rj})^2}{\mathbb{E}_{u \in \mathcal{V}}[(g^{\mathrm{V}}_{ru})^2] + \epsilon}, c \right\} \right). \tag{7}$$

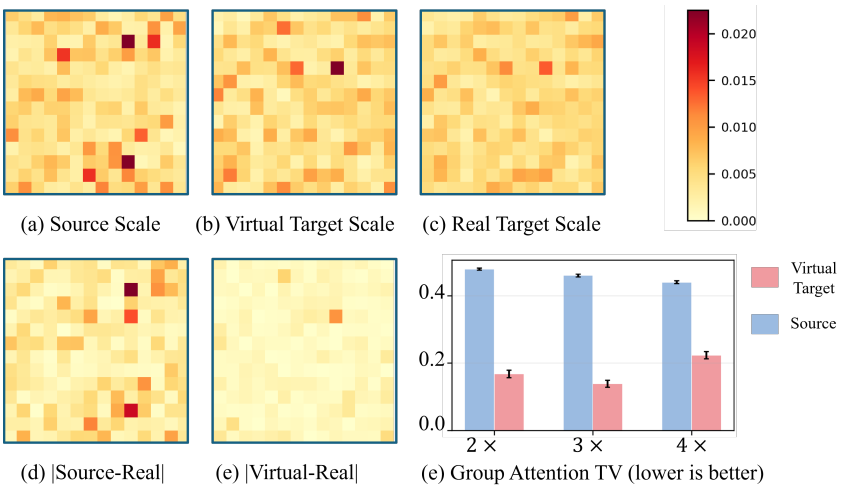


Figure 2: Controlled virtual-target validation under content-preserving key duplication. The virtual state uses stretched RoPE coordinates and analytic key multiplicities while retaining source features. Grouped deviations from an explicit target construction are lower than for source-native attention in the shown extrapolation factors; the experiment does not claim to reproduce target-side feature changes.

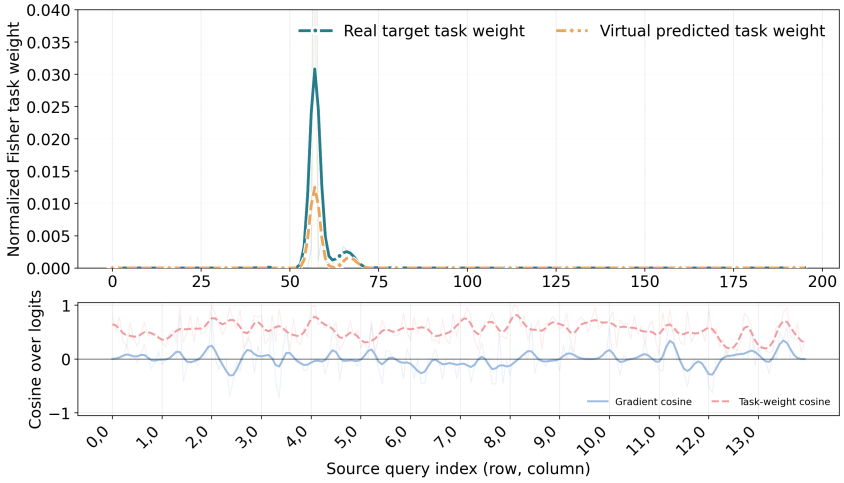


Figure 3: Real-image diagnostic of task-Fisher transfer at a representative layer and head. The virtual and explicit target weights are compared on common grouped support. The Fisher-weight agreement is more stable than signed-gradient agreement in this example; it remains an empirical local proxy rather than target supervision.

The square preserves local sensitivity while avoiding cancellation between opposite signed directions. Detachment, the cap $c$, and $\epsilon$ are numerical safeguards; the weights are computed from visual attention entries. The local interpretation and transfer bound are given in Appendix A.

Figure 3 shows a real-image check that the virtual task-Fisher pattern tracks the explicit target on grouped support.

We use these weights to calibrate the virtual RoPE logits toward the native source logits:

$$\mathcal{L}_{\text{freq}}(s, U_s) = \mathbb{E}\left[\sum_r \sum_{j \in \mathcal{V}} w_{rj}^{\text{V}} \left(a_{rj}^{\text{V}}(U_s) - \text{sg}(a_{rj}^{0})\right)^2\right]. \tag{8}$$

This objective keeps source-task-sensitive relative-logit structure stable while $U_s$ compensates for the target positional span. It is a calibration signal, not a claim that source and target logits are identical.

**Response calibration.** Let $p_{rj} = \mathrm{softmax}(\beta a_r)_j$ and $c_{rj} = \partial \mathcal{L}_{\mathrm{task}} / \partial p_{rj}$. The softmax Jacobian gives

$$g_{rj} = \frac{\partial \mathcal{L}_{\mathrm{task}}}{\partial a_{rj}} = \beta p_{rj} \left( c_{rj} - \sum_k p_{rk} c_{rk} \right). \tag{9}$$

Only the centered downstream score is observable from the logit gradient. At the native source state, with $p^0_{rj} = \mathrm{softmax}(a^0_r)_j$ and $g^0_{rj} = \partial \mathcal{L}_{\mathrm{task}} / \partial a^0_{rj}$, we recover a centered score by

$$\widehat{c}_{rj} = \frac{g^0_{rj}}{p^0_{rj} + \epsilon} - \sum_k p^0_{rk} \frac{g^0_{rk}}{p^0_{rk} + \epsilon}. \tag{10}$$

This representation is invariant to adding a row-wise constant to all logits.

With multiplicity, one represented micro-key has probability

$$\widetilde{p}_{rj} = \frac{\exp(\beta_s a^{\mathrm{V}}_{rj}(U_s))}{\sum_k \mu_k \exp(\beta_s a^{\mathrm{V}}_{rk}(U_s))}, \qquad m_{rj} = \mu_j \widetilde{p}_{rj}. \tag{11}$$

The predicted target Fisher transmission of group $j$ is the sum of squared micro-key gradients,

$$\widetilde{\mathcal{I}}^T_{rj} = \mu_j \beta_s^2 \widetilde{p}^2_{rj} \left( \widehat{c}_{rj} - \sum_k m_{rk} \widehat{c}_{rk} \right)^2, \qquad \widetilde{\mathcal{I}}^T_r = \sum_j \widetilde{\mathcal{I}}^T_{rj}. \tag{12}$$

The multiplicity changes the score center through the target-like mass $m$; $\beta_s$ then rescales how strongly logit differences are expressed after softmax. This is why the inverse-temperature factor is optimized jointly with $U_s$ rather than applied as a fixed global heuristic.

Let $\mathcal{I}^0_r = \sum_j (g^0_{rj})^2$. We match task-sensitive response energy using a clipped log-ratio,

$$\mathcal{L}_{\mathrm{resp}} = \frac{\sum_r \overline{w}_r \left[ \mathrm{clip}\left( \log \frac{\widetilde{\mathcal{I}}^T_r + \epsilon}{\mathcal{I}^0_r + \epsilon}, -\tau, \tau \right) \right]^2}{\sum_r \overline{w}_r}, \qquad \overline{w}_r = \mathrm{sg}\left( \min\left\{ \frac{\mathcal{I}^0_r}{\mathbb{E}_u[\mathcal{I}^0_u] + \epsilon}, c \right\} \right). \tag{13}$$

At the source scale, $U = I$, $\beta = 1$, and $\mu = 1$ recover $\widetilde{\mathcal{I}}^T_r = \mathcal{I}^0_r$ up to the probability floor. Away from the source scale, this objective is a source-anchored, target-conditioned surrogate rather than a target-label loss.

### 4.4 Joint Calibration and Anchor Interpolation

For each layer, head, and target anchor, VTC parameterizes the frequency correction and inverse-temperature factor as

$$U_{\ell h,s} = \exp(S_{\ell h,s}), \qquad S^H_{\ell h,s} = -S_{\ell h,s}, \qquad \beta_{\ell h,s} = \exp(b_{\ell h,s}). \tag{14}$$

The skew-Hermitian generator makes $U$ block-unitary, while the exponential parameterization guarantees $\beta > 0$. Only $S_s$ and $b_s$ are optimized:

$$\mathcal{L}_{\mathrm{VTC}} = \mathcal{L}_{\mathrm{freq}} + \lambda_R \mathcal{L}_{\mathrm{resp}} + \lambda_I \|S_s\|_2^2 + \lambda_\beta \|b_s\|_2^2. \tag{15}$$

The backbone and task head are frozen. At the source scale, $U = I$, $\beta = 1$, and $\mu_j = 1$, so the virtual construction is anchored to the native computation. For an unseen resolution, $S$ and $b$ are linearly interpolated in log-resolution space between neighboring anchors and then exponentiated; this preserves unitary $U$ and positive $\beta$ without further optimization. Numerical safeguards and the calibration schedule are listed in Appendix B.

## 5 Experiments

We evaluate VTC on image classification, image generation, and video generation. Unless otherwise stated, the backbone and task head are frozen during calibration, source-scale inputs provide all calibration supervision, and each target anchor uses at most 200 updates. Target-scale inputs are used only for final evaluation.

Table 1: ImageNet-1K resolution extrapolation. Values are Top-1 accuracy (%). Each target size reports the frozen baseline, Vision-YaRN, and VTC in that order. VTC and Vision-YaRN use the same checkpoint and evaluation protocol as the baseline.

| | 512×512 | | | 768×768 | | | 1024×1024 | | |
|---|---|---|---|---|---|---|---|---|---|
| Backbone | Baseline | Vision-YaRN | VTC | Baseline | Vision-YaRN | VTC | Baseline | Vision-YaRN | VTC |
| ViT-S / Compass-RoPE | 80.34 | 80.73 | **82.19** | 65.88 | 73.86 | **77.45** | 34.22 | 60.12 | **65.17** |
| ViT-S / Mix-RoPE | 79.07 | 80.38 | **81.44** | 54.71 | 72.53 | **74.59** | 16.55 | 56.45 | **58.30** |
| ViT-S / Axial-RoPE | 76.08 | 76.77 | **78.70** | 53.92 | 61.07 | **64.24** | 20.62 | 40.56 | **41.39** |
| ViT-S / Spiral-RoPE | 78.84 | 78.64 | **80.23** | 63.98 | 69.01 | 73.30 | 35.30 | 52.31 | **59.54** |
| ViT-5 | 80.63 | 79.71 | **81.43** | 69.79 | 69.35 | **74.99** | 41.84 | 52.60 | **55.02** |

Table 2: ImageNet-1K component ablation. Values are Top-1 accuracy (%). The natural source-scale extrapolation is reported in Table 1; this table isolates the two VTC components and their combination. "Response-only" includes analytic key multiplicities and the learned response scale $\beta_s$.

| RoPE | Target size | Frequency-only | Response-only | Full VTC |
|---|---|---|---|---|
| Mix-RoPE | 512 | 79.96 | 80.74 | **81.44** |
| | 768 | 62.04 | 69.21 | **74.59** |
| | 1024 | 23.67 | 42.34 | **58.30** |
| Compass-RoPE | 512 | 81.08 | 81.65 | **82.19** |
| | 768 | 69.65 | 75.42 | **77.45** |
| | 1024 | 38.60 | 62.85 | **65.18** |

### 5.1 IMAGENET CLASSIFICATION

We evaluate ViT-S models trained at $224 \times 224$ on ImageNet-1K (Dosovitskiy et al., 2021; Touvron et al., 2021). For every classification backbone, calibration anchors are $512$, $672$, $800$, and $1024$ pixels; parameters at intermediate resolutions are obtained by interpolation. The main comparison reports $512$, $768$, and $1024$ pixels, while the other evaluated resolutions are given in Appendix C.

Across all 15 backbone–resolution entries in Table 1, VTC improves over native extrapolation and also exceeds Vision-YaRN. The mean VTC gain over native extrapolation is 1.81, 11.26, and 26.18 percentage points at 512, 768, and 1024 pixels, respectively; the mean gain over Vision-YaRN is 1.55, 3.75, and 3.48 points at the same resolutions. The growing advantage at larger grids indicates that the response correction is most useful when the unadjusted key-population and phase changes are severe. These aggregate gains cover five RoPE backbones and three target sizes.

**Component ablation.** We isolate the frequency and response components on the Mix-RoPE and Compass-RoPE runs. "Frequency-only" optimizes $U_s$ with the task-weighted logit objective, while "response-only" uses the multiplicity-aware response calibration with $\beta_s$ and fixed analytic multiplicities. Full VTC combines both components.

Both isolated components improve over the corresponding native baseline, and the joint model is the strongest variant in every displayed ablation row. The response-only branch is especially helpful at 1024 pixels, where enlarged key competition is strongest, while frequency-only calibration provides a smaller but consistent gain. Their combination therefore addresses two coupled failure modes.

### 5.2 ATTENTION RESPONSE ANALYSIS

To test the mechanism behind the accuracy gains, we compare the source-native state, the virtual state expanded onto the target micro-key support, and the explicit target state. The source-native attention remains on its native support; only the virtual state is expanded before comparison. Figure 4 shows a qualitative Compass-RoPE example, and Table 3 reports aggregate real-model diagnostics.

The qualitative maps show how VTC changes the spatial allocation of response at an extrapolated resolution; in the shown example, it places more weight on the foreground than native extrapolation. The aggregate provides a stronger mechanistic check. Full VTC reduces grouped TV from $0.6254$ for source-native attention to $0.5068$, an absolute reduction of $0.1186$ (19.0%), and reduces relative response $L_2$ from $0.6681$ to $0.5683$, an absolute reduction of $0.0997$ (14.9%). Coordinate-only calibration accounts for most of the TV improvement ($0.5445$; 12.9% relative reduction), whereas

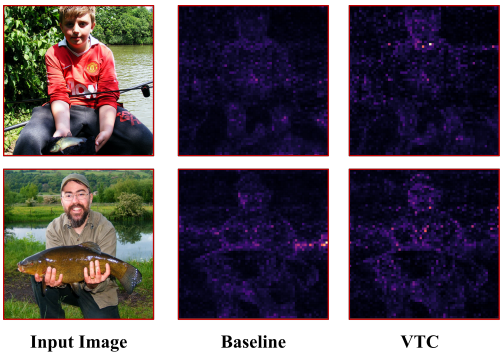


Figure 4: Qualitative attention maps from Compass-RoPE ViT at an extrapolated resolution. Each row shows the input image (left), the frozen baseline attention (middle), and the full VTC attention (right). Relative to the baseline, VTC concentrates more response on task-relevant foreground regions, such as the person and the fish, while reducing diffuse background activation.

Table 3: Aggregate real-model attention-response diagnostic over the reported Compass-RoPE validation on 100 images from ImageNet-1K (five layers and target sizes 512, 768, and 1024). Lower is better. Source-native attention is measured on its native support; virtual attention is expanded to target micro-keys before comparison with the direct target forward.

| Variant | Metric | Value |
|---|---|---|
| Source native | Grouped TV ↓ | 0.6254 |
| | Response rel. $L_2$ ↓ | 0.6681 |
| Coordinate-only | Grouped TV ↓ | 0.5445 |
| | Response rel. $L_2$ ↓ | 0.6209 |
| Multiplicity-only | Grouped TV ↓ | 0.5775 |
| | Response rel. $L_2$ ↓ | 0.6003 |
| Full VTC | Grouped TV ↓ | **0.5068** |
| | Response rel. $L_2$ ↓ | **0.5683** |

Table 4: FiTv2-XL image-generation extrapolation on ImageNet-1K. Inception Score (IS) and precision/recall are better when higher; FID and sFID are better when lower.

| Resolution | Method | IS ↑ | FID ↓ | sFID ↓ | Precision ↑ | Recall ↑ |
|---|---|---|---|---|---|---|
| 384×384 | FiTv2 baseline | 102.09 | 42.12 | 59.67 | 0.44 | **0.69** |
| | Vision-YaRN | 157.93 | 23.23 | 35.13 | 0.55 | 0.48 |
| | VTC | **221.72** | **14.20** | **33.25** | **0.69** | 0.65 |
| 512×512 | FiTv2 baseline | 6.18 | 264.44 | 181.22 | 0.01 | 0.12 |
| | Vision-YaRN | 22.44 | 152.13 | 122.11 | 0.12 | 0.47 |
| | VTC | **33.32** | **117.85** | **100.32** | **0.18** | **0.56** |

multiplicity-only calibration contributes more to response matching (0.6003; 10.1% relative reduction). The full model is best on both metrics, indicating that coordinate alignment and key-population correction provide complementary benefits across layers, target sizes, and validation images.

### 5.3 High-Resolution Image Generation

We evaluate FiTv2-XL, a diffusion transformer pretrained at $256 \times 256$, on ImageNet-1K image generation (Lu et al., 2024b; Peebles & Xie, 2023; Chen et al., 2024). For each method and resolution, we generate 10,000 images with ODE sampling for 250 steps and classifier-free guidance (CFG) 1.5 (Song et al., 2021a;b; Salimans & Ho, 2022; Lipman et al., 2023; Ma et al., 2024). Calibration anchors are $320 \times 320$, $448 \times 448$, and $520 \times 520$; the reported $384 \times 384$ and $512 \times 512$ settings are obtained by interpolation. We report Inception Score (IS), Fréchet Inception Distance (FID), spatial FID (sFID), precision, and recall; higher values are better for IS, precision, and recall, whereas lower values are better for FID and sFID.

At $384 \times 384$, VTC leads on IS, FID, sFID, and precision, and recovers more recall than Vision-YaRN. At $512 \times 512$, it is best on all five metrics. Relative to native extrapolation, VTC reduces FID by 66.3% at $384 \times 384$ and by 55.4% at $512 \times 512$; relative to Vision-YaRN, the reductions are 38.9% and 22.5%, respectively. Figure 5 gives a qualitative view: artifacts remain at the larger resolutions, but VTC better preserves the main subject and local details. These gains should be interpreted together with the reported sample count and sampling protocol.

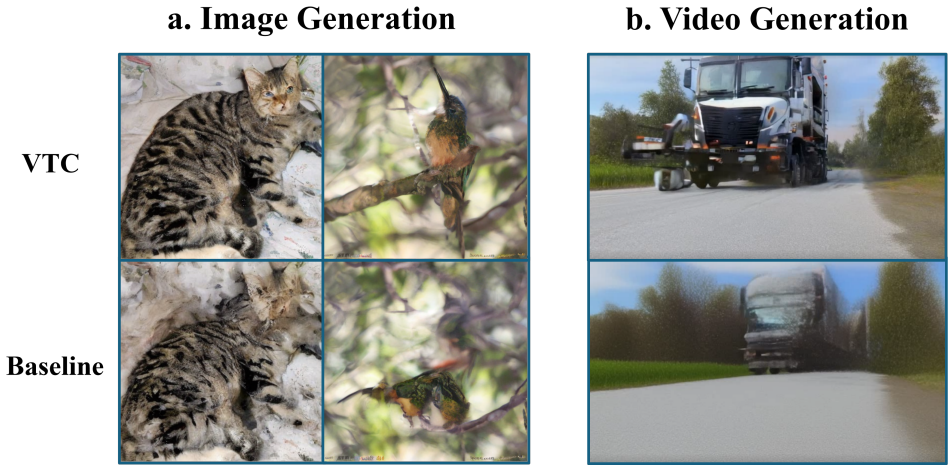


Figure 5: Qualitative samples for image and video extrapolation. Each pair compares the baseline and VTC under an extrapolated setting. Although extrapolation artifacts remain, VTC better preserves the main subject and local details.

### 5.4 Video Length Extrapolation

We evaluate CogVideoX-5B when extrapolating from its native 49 frames to a 97-frame target. We sample 100 prompts from the VBench prompt set and generate three seeds per prompt, yielding 300 videos per method (Yang et al., 2025; Huang et al., 2024; Ho et al., 2022). We report VBench Image Quality, Dynamic Degree, and Overall Consistency; higher values are better for all three metrics. Calibration uses native-length clips and at most 200 updates at the 97-frame anchor. Figure 5 shows a representative qualitative comparison. Although temporal extrapolation still introduces artifacts, VTC preserves the main object structure and local appearance more reliably than the native baseline, consistent with the metric improvements in Table 5. The table reports aggregate scores over all 300 generated videos.

Table 5: CogVideoX-5B extrapolation from 49 to 97 frames on 300 generated videos. We report three VBench metrics; higher values are better.

| Method | Image Quality | Dynamic Degree | Overall Consistency |
|---|---|---|---|
| CogVideoX baseline | 0.532 | 0.557 | 0.180 |
| VTC | **0.584** | **0.593** | **0.188** |

### 5.5 Computation Cost

All calibration runs use four RTX 4090 GPUs with a per-GPU batch size of 16. The four GPUs run in parallel. Image-classification calibration takes about 5 minutes per backbone, image-generation calibration about 10 minutes, and video calibration about 7 hours. The video run is longer because CogVideoX-5B is substantially more expensive per update. These measurements describe the post-hoc calibration budget; they are not a hardware-normalized comparison with pretraining, for which the published setup does not provide a directly comparable GPU-hour accounting.

## 6 Conclusion

VTC constructs a target-conditioned surrogate from source-scale inputs by stretching RoPE coordinates and modeling the enlarged key population with analytic multiplicities. Source-task Fisher responses then calibrate frequency geometry and post-softmax signal transmission while the backbone remains frozen. The controlled experiment supports this construction under its exchangeable-content assumption, and experiments on classification, image generation, and video generation show that source-only calibration can improve large-grid extrapolation at a small post-hoc cost.

## Ethics Statement

This work uses publicly available datasets and pretrained models and does not collect new data from human participants or use personally identifiable information as an experimental variable. We follow the intended research use and licensing requirements of the datasets and evaluation resources.

## Reproducibility Statement

We report the model settings, source and target grids, calibration objectives, anchor resolutions, update budget, and evaluation protocols needed to reproduce the experiments. Implementation code, configuration files, and evaluation scripts will be released upon publication.

## AI Use Statement

AI-based tools were used for language editing, LaTeX cleanup, and reference organization. The authors reviewed and verified all technical claims, derivations, experiments, and final text and take full responsibility for the content of this paper.

## References

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14496–14506. IEEE, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2024.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Noam Issachar, Guy Yariv, Sagie Benaim, Yossi Adi, Dani Lischinski, and Raanan Fattal. Dype: Dynamic position extrapolation for ultra high resolution diffusion. In *Forty-third International Conference on Machine Learning*, 2026.

Boyang Li, Yulin Wu, Sizhe Xu, Nuoxian Huang, Zhonghang Yuan, Shangyi Guo, Shu Yang, and Takahiro Yabe. nd-roPE: A generalized roPE for n-dimensional position embedding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YPXDQkU7XW.

Jiaye Li, Baoyou Chen, Hui Li, Zilong Dong, Jingdong Wang, and Siyu Zhu. Head-wise adaptive rotary positional encoding for fine-grained image generation. *arXiv preprint arXiv:2510.10489*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haoyu Liu, Sucheng Ren, Tingyu Zhu, Peng Wang, Cihang Xie, Alan Yuille, Zeyu Zheng, and Feng Wang. Spiral roPE: Rotate your rotary positional embeddings in the 2d plane. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=ORwyqAzROx.

Yihua Liu, Fanjiang Ye, Bowen Lin, Rongyu Fang, and Chengming Zhang. Tide: Text-informed dynamic extrapolation with step-aware temperature control for diffusion transformers. *arXiv preprint arXiv:2603.08928*, 2026b.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *International Conference on Learning Representations*, volume 2024, pp. 19259–19286, 2024a.

Zeyu Lu, ZiDong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and LEI BAI. Fit: Flexible vision transformer for diffusion model. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=jZVen2JguY.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Chengxi Min, Wei Wang, and Yao Zhao. Compass-rope: Isotropic rotary position embeddings for vision transformers. In *Forty-third International Conference on Machine Learning*.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Gao Peng, Le Zhuo, Dongyang Liu, DU DU, Xu Luo, Longtian Qiu, Yuhang Zhang, Rongjie Huang, Shijie Geng, Renrui Zhang, et al. Lumina-t2x: Scalable flow-based large diffusion transformer for flexible resolution generation. In *International Conference on Learning Representations*, volume 2025, pp. 37254–37292, 2025.

Javad Rajabi, Kimia Shaban, Koorosh Roohi, David B Lindell, and Babak Taati. Sega: Spectral-energy guided attention for resolution extrapolation in diffusion transformers. *arXiv preprint arXiv:2605.22668*, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

Feihong Yan, Shaoyu Liu, Haixuan Wang, Shuai Lu, Linfeng Zhang, Huiqi Li, and Xiangyang Ji. Extravar: Stage-aware rope remapping for resolution extrapolation in visual autoregressive models. *arXiv preprint arXiv:2605.10045*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=v3B79m7t8Z.

Min Zhao, Bokai Yan, Xue Yang, Hongzhou Zhu, Jintao Zhang, Shilong Liu, Chongxuan Li, and Jun Zhu. Ultraimage: Rethinking resolution extrapolation in image diffusion transformers. *arXiv preprint arXiv:2512.04504*, 2025b.

Min Zhao, Hongzhou Zhu, Yingze Wang, Bokai Yan, Jintao Zhang, Guande He, Ling Yang, Chongxuan Li, and Jun Zhu. Ultravico: Breaking extrapolation limits in video diffusion transformers. In *International Conference on Learning Representations*, volume 2026, pp. 122283–122309, 2026.

# A Detailed Theoretical Bounds

## A.1 Virtual-logit approximation

The virtual state preserves source content while changing the RoPE coordinates. For a target micro-key $q \in G_j$, write the target phase and content interaction as

$$\phi^T_{rq,m} = \phi^{\mathrm{V}}_{rj,m} + \delta\phi_{rq,m}, \qquad c^T_{rqm} = c^{\mathrm{V}}_{rjm} + \delta c_{rqm}.$$

The target and virtual logits differ only through these two residuals. Applying $|e^{\mathrm{i}\delta} - 1| \le \min\{2, |\delta|\}$ to each complex RoPE channel gives

$$\left|a^T_{rq} - a^{\mathrm{V}}_{rj}\right| \le \frac{1}{\sqrt{d_h}} \sum_m \left[|\delta c_{rqm}| + |c^{\mathrm{V}}_{rjm}| \min\{2, |\delta\phi_{rq,m}|\}\right]. \tag{16}$$

The first term measures content drift inside a source cell; the second measures phase drift caused by representing several target positions with one virtual source position. The bound is therefore diagnostic rather than a statement that the virtual logit is always accurate: high-frequency channels, large within-cell spatial variation, or a large target/source coordinate mismatch can make the second term dominant. Conversely, when target micro-keys remain locally similar and their phase offsets are small, the virtual logit is a good representative for the group. This explains why the method uses grouped responses and why the controlled experiment isolates coordinate and multiplicity effects before testing real feature drift.

A logit bound matters only through the softmax response. For a row-wise perturbation $\Delta a$, the first-order softmax change is

$$\Delta p \approx \beta(\mathrm{diag}(p) - pp^\top)\Delta a,$$

so a bounded logit error is attenuated for saturated probabilities but has a larger effect when probability mass is distributed across many comparable keys. This is the regime created by resolution extrapolation. The empirical checks in the paper therefore compare attention TV/JS and grouped task responses rather than treating logit error alone as the end metric.

## A.2 Multiplicity approximation

Let the true target group mass and its virtual multiplicity approximation be

$$M^T_{rj} = \frac{\sum_{q\in G_j} e^{\beta a^T_{rq}}}{Z_T}, \qquad m_{rj} = \frac{\mu_j e^{\beta a^{\mathrm{V}}_{rj}}}{Z_V},$$

where $\mu_j = |G_j|$ for visual groups and $\mu_j = 1$ for non-visual keys. Suppose $a^T_{rq} = a^{\mathrm{V}}_{rj} + \varepsilon_{rq}$ and $|\varepsilon_{rq}| \le \epsilon_a$. Every term in the group numerator then lies between $e^{-\beta\epsilon_a}$ and $e^{\beta\epsilon_a}$ times the virtual term. The same argument applies to all groups in the partition function, which leads to

$$\left|\log \frac{M^T_{rj}}{m_{rj}}\right| \le 2\beta\epsilon_a. \tag{17}$$

The factor of two has a concrete interpretation: one factor comes from the error in the numerator of group $j$, and one comes from the error in the global normalizer. Under exact exchangeability, $\varepsilon_{rq} = 0$ for every micro-key and the multiplicity representation is exact. In real images or videos, exchangeability is only approximate because neighboring target keys can see different content and have different phases. Multiplicity still captures the dominant counting effect, but it cannot represent the residual within-group structure. In the implementation, the analytic ratio $N^{\mathrm{vis}}_s/N^{\mathrm{vis}}_0$ may be non-integer; it is a continuous relaxation of the same log-partition expression and should be interpreted as an expected micro-key count. Thus, the exact bound above applies literally to integer exchangeable groups, while the implementation uses the corresponding relaxed approximation.

A useful refinement is obtained by writing $\varepsilon_{rq} = \bar{\varepsilon}_j + \eta_{rq}$ with $\sum_{q\in G_j} \eta_{rq} = 0$. The group numerator contains the factor

$$\frac{1}{\mu_j} \sum_{q\in G_j} e^{\beta\eta_{rq}} \approx 1 + \frac{\beta^2}{2} \mathrm{Var}_{q\in G_j}(\eta_{rq})$$

when the residuals are small. Thus, zero-mean heterogeneity produces a second-order partition error, while a nonzero group mean produces a first-order logit shift. Both effects are amplified by $\beta$; large target groups can also accumulate more heterogeneous content. This is why the relevant validation compares grouped target mass and micro-key TV/JS, rather than assuming that a single virtual logit reproduces every target micro-key.

### A.3 Task-Fisher response transfer

The task-gradient construction is a local sensitivity approximation. Let $g^T_{rq} = \partial \mathcal{L}/\partial a^T_{rq}$ be the target logit gradient and let $\widetilde{g}_{rj}$ be the virtual representative gradient for group $G_j$. The correct comparison is on common grouped support:

$$I^T_{rj} = \sum_{q \in G_j} (g^T_{rq})^2, \qquad \widetilde{I}_{rj} = \mu_j (\widetilde{g}_{rj})^2,$$

not between unequal-length source and target vectors. If $|g^T_{rq} - \widetilde{g}_{rj}| \le \epsilon_g$, then

$$\left| I^T_{rj} - \widetilde{I}_{rj} \right| \le \mu_j \epsilon_g \left( 2|\widetilde{g}_{rj}| + \epsilon_g \right). \tag{18}$$

The first-order term is proportional to the representative gradient and the second-order term remains when the representative is small. Consequently, a small logit/probability approximation error does not automatically imply an accurate task weight: downstream score drift can still change $g$.

For a small logit perturbation,

$$\delta \mathcal{L}_{\text{task}} \approx \sum_j g_{rj}\, \delta a_{rj}, \qquad \mathbb{E}[(\delta \mathcal{L}_{\text{task}})^2] \approx \sum_j \sigma^2_{rj} g^2_{rj}, \tag{19}$$

under locally zero-mean, weakly correlated perturbations. Squared gradients are therefore a local Fisher sensitivity proxy, while signed gradients can cancel across directions. This is a second-order surrogate for task sensitivity, not a claim that virtual Fisher weights equal the target loss or target gradients. The source and target share the frozen downstream head, which makes the source sensitivity informative near the source boundary; feature drift and later-layer propagation are the reasons the paper also reports real-model attention agreement and treats global weight-map equality as an empirical question.

### A.4 Softmax-chain derivation and response calibration

Let $p_{rj}$ be an attention probability and $c_{rj} = \partial \mathcal{L}_{\text{task}}/\partial p_{rj}$ its downstream score. With response scale $\beta$, differentiating $p = \mathrm{softmax}(\beta a)$ gives

$$\frac{\partial \mathcal{L}_{\text{task}}}{\partial a_{rj}} = \beta p_{rj} \left( c_{rj} - \sum_k p_{rk} c_{rk} \right). \tag{20}$$

The centered score is invariant to adding a constant to every logit, which is why the response objective focuses on transmission through the softmax rather than absolute logit offsets. At the source state, we recover a centered score from the native probability and logit gradient as

$$\widehat{c}_{rj} = \frac{g^0_{rj}}{p^0_{rj} + \epsilon} - \sum_k p^0_{rk} \frac{g^0_{rk}}{p^0_{rk} + \epsilon}. \tag{21}$$

With multiplicity, the virtual group mass is $m_{rj} = \mu_j \widetilde{p}_{rj}$, where $\widetilde{p}_{rj} \propto e^{\beta a_{rj}(U_s)}$. The predicted grouped Fisher transmission is consequently

$$\widetilde{\mathcal{I}}_{rj} = \mu_j \beta^2 \widetilde{p}^2_{rj} \left( \widehat{c}_{rj} - \sum_k m_{rk} \widehat{c}_{rk} \right)^2,$$

and the row transmission is the sum over groups. The multiplicity changes the center of the downstream score through the target-like mass distribution; $\beta$ then adjusts how sharply the available logit differences are expressed after that enlarged competition. This is the mechanism-level reason for calibrating $\beta$ jointly with the frequency mixer rather than applying a post-hoc global temperature alone.

### A.5 Source anchoring and scope of the approximation

At the source scale, $U = I$, $\beta = 1$, and $\mu_j = 1$. The virtual support, probabilities, recovered downstream scores, and grouped Fisher transmission then reduce to their native counterparts up to the probability floor $\epsilon$. This boundary condition prevents calibration from introducing an arbitrary task response at the training scale. Away from the source boundary, the preceding bounds show exactly what can fail: phase/content drift perturbs the representative logits, within-group non-exchangeability perturbs the multiplicity mass, and downstream score drift perturbs Fisher transmission. The method is therefore best understood as a source-anchored approximation whose quality must be checked on target-aligned attention responses; it is not a proof that a source-only virtual forward reconstructs all target features.

## B Calibration Algorithm and Implementation Details

For each target anchor, VTC caches source logits, probabilities, and gradients, recovers $\widehat{c}$ and $\mathcal{I}^0$, computes detached virtual Fisher weights, and updates only $S_s$ and $b_s$. The backbone remains frozen. The probability floor $\epsilon$, Fisher cap $c$, log-ratio threshold $\tau$, and gradient clipping are numerical safeguards and do not change the stated derivations. Each anchor uses at most 200 updates; in the ImageNet classification setting, calibration takes about five minutes on RTX 4090 hardware.

All phase construction, probability normalization, downstream-score recovery, and Fisher accumulation use FP32. The frozen backbone may use bfloat16. Temperature is applied to the scaled QK logits immediately before softmax, and the multiplicity log-count is added to the same softmax input. Target resolution is used only to construct virtual coordinates and multiplicities during calibration.

## C Additional Classification Results

The main table reports the baseline, Vision-YaRN, and VTC at $512$, $768$, and $1024$ pixels. This appendix reports every classification resolution present in the recorded evaluation results, including the main-table resolutions, using only the natural source extrapolation and full VTC. Rows are grouped by backbone. Bold labels mark calibration anchors $512$, $672$, $800$, and $1024$; regular labels are interpolated evaluations. The exported logs contain no score row at the $800$ anchor, so the corresponding entries are shown as unavailable rather than interpolated. Compass logs do not contain a natural-baseline result for the intermediate rows, so those cells are left unavailable.

Across the complete logs, full VTC consistently improves the natural source-only extrapolation wherever both values are available, including the intermediate resolutions between anchors. The table also makes clear which entries are measured anchor/interpolation evaluations and which anchor rows are not present in the exported result files.

Table 6: Complete ImageNet-1K classification results from the recorded resolution logs. Values are Top-1 accuracy (%). Natural is the frozen source extrapolation and Full VTC is the complete calibration method. Bold resolution labels are calibration anchors; regular labels are interpolated evaluations.

| Backbone | Resolution | Natural | Full VTC |
|---|---|---|---|
| Mix-RoPE | **512** | 79.07 | 81.44 |
| | 640 | 71.60 | 79.47 |
| | **672** | 68.27 | 78.49 |
| | 720 | 62.19 | 76.98 |
| | 768 | 54.71 | 74.59 |
| | 864 | 37.83 | 69.36 |
| | **1024** | 16.55 | 58.30 |
| Compass-RoPE | **512** | 80.34 | 82.19 |
| | 640 | 75.43 | 80.67 |
| | **672** | 73.55 | 80.21 |
| | 720 | 70.14 | 78.97 |
| | 768 | 65.88 | 77.45 |
| | 864 | 55.23 | 74.14 |
| | **1024** | 34.22 | 65.18 |
| Axial-RoPE | **512** | 76.08 | 78.70 |
| | 640 | 67.85 | 72.91 |
| | **672** | 64.91 | 70.61 |
| | 720 | 59.85 | 68.81 |
| | 768 | 53.92 | 64.24 |
| | 864 | 40.58 | 55.24 |
| | **1024** | 20.62 | 41.39 |
| Spiral-RoPE | **512** | 78.84 | 80.24 |
| | 640 | 73.16 | 77.79 |
| | **672** | 71.12 | 76.88 |
| | 720 | 67.82 | 75.50 |
| | 768 | 63.98 | 73.31 |
| | 864 | 54.31 | 69.19 |
| | **1024** | 35.30 | 59.54 |
| ViT-5 | **512** | 80.63 | 81.44 |
| | 640 | 76.61 | 79.04 |
| | **672** | 75.03 | 78.06 |
| | 720 | 72.23 | 76.62 |
| | 768 | 69.79 | 74.99 |
| | 864 | 60.20 | 69.84 |
| | **1024** | 41.84 | 55.02 |