MuSViT: A Foundation Vision Model for Sheet Music Representation

Carlos Penarrubia, Antonio Rios-Vila, Eliseo Fuentes-Martinez, Juan C. Martinez-Sevilla, Francisco J. Castellanos, María Alfaro-Contreras, and Jorge Calvo-Zaragoza

Pattern Recognition and Artificial Intelligence Group, University of Alicante, Spain

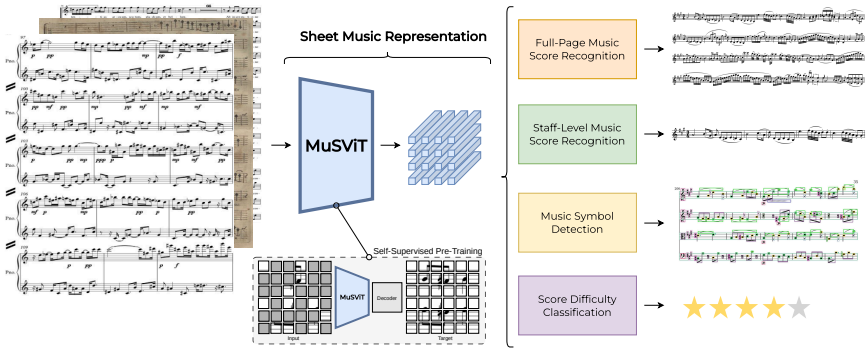

Fig. 1: Overview of MuSViT. MuSViT is pre-trained on diverse sheet music pages using Masked Autoencoders: patches are randomly masked and the model learns to reconstruct the missing content from the remaining visible context. We evaluate the generality of the learned representations by probing the encoder across four diverse downstream tasks: full-page and staff-level music score recognition, music symbol detection, and score difficulty classification. Pre-trained models, code, and reproducibility resources are available through the project page: `https://grfia.dlsi.ua.es/musvit`.

Abstract. Foundation models have transformed vision and language processing by providing rich, reusable representations that transfer across diverse tasks. Sheet music, as a visual encoding of musical language, lacks such a strong domain-specific backbone. We introduce MuSViT (**Mu**sic **S**core **Vi**sion **T**ransformer): the first foundation vision model for sheet music representation—a ViT encoder pre-trained via Masked Autoencoders on 9.7 million pages from the International Music Score Library Project (IMSLP). To handle the complexity of real-world scores, we adopt a two-stage curriculum: a synthetic warm-up on typeset scores followed by large-scale training on the full IMSLP corpus. We evaluate MuSViT on four downstream tasks—full-page and staff-level music score recognition, music symbol detection, and score difficulty classification—under two scenarios: linear probing (frozen encoder) and fine-tuning. Under linear probing, MuSViT consistently outperforms modern vision encoders, revealing that general-purpose representations, regardless of scale, fall systematically short on the structured symbolic properties

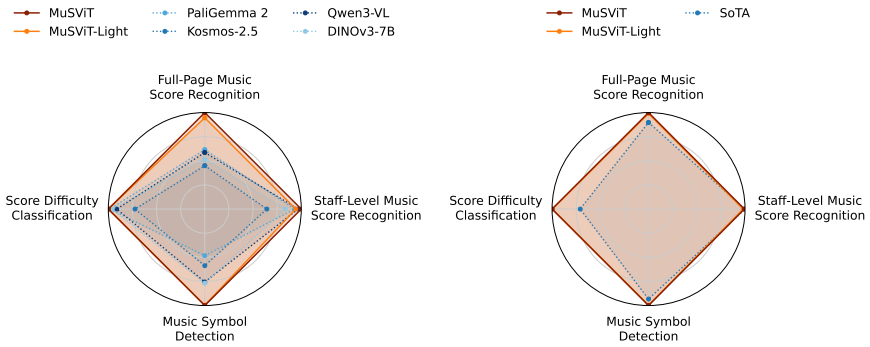

Fig. 2: MuSViT performance across four downstream tasks. **Left:** *Linear probing* (frozen encoder)—MuSViT (solid) consistently outperforms general-purpose vision encoders (dashed), demonstrating superior representation quality. **Right:** *Fine-tuning*—MuSViT generally outperforms state-of-the-art methods (SoTA). Axes represent normalized performance on each task (higher is better); see Section 3 for detailed results.

of musical notation. Under fine-tuning, MuSViT generally improves upon task-specific state-of-the-art methods. An additional embedding-transcription consistency analysis reveals that MuSViT encodes symbolic musical structure directly in its representation space—unlike other encoders, whose embeddings do not correlate with music notation content. These results establish MuSViT as a foundation backbone for sheet music understanding.

Keywords: Representation Learning · Foundation Vision Model · Sheet Music · MuSViT

1 Introduction

Music represents an essential part of human tradition and a valuable element of cultural heritage. Over the centuries, musical compositions have been transmitted both orally and in written form, with music encoded in documents known as scores (or sheet music). Today, vast collections of these scores are preserved in libraries and archives, and large-scale digitization initiatives have made millions of pages available online.

However, the majority of this material exists only as raw page images. Due to the high cost of manual transcription, most scores have not been converted into structured digital formats that enable indexing, retrieval, or automated analysis. As a result, despite the growing availability of digitized music documents, much of this cultural treasure remains effectively inaccessible and underutilized.

Traditional approaches to automated music document analysis—most notably in the field of Optical Music Recognition (OMR)—have focused on task-specific systems, employing end-to-end models or multi-stage pipelines trained on limited annotated datasets [7]. These systems often remain brittle, with performance degrading sharply when confronted with unfamiliar notation styles,

engraving conventions, or visual artifacts that differ from the training data [36]. The extreme heterogeneity of real-world music scores continues to limit the generalization capabilities of these specialized models.

Recent advances in deep learning have demonstrated the value of foundation models for *representation learning* from large-scale unlabeled data. In computer vision, Vision Transformers (ViTs) pre-trained via self-supervised learning—such as DINO [9, 25, 33]—learn transferable visual representations that can serve as general-purpose backbones for diverse downstream tasks. Language-image pre-training approaches such as CLIP [27] and SigLIP [40], together with vision-language models such as PaliGemma [6,34] and Qwen-VL [3,4,38], extend these principles to multimodal understanding.

A similar trend is observed in the audio domain: general-purpose models such as Qwen-Audio [10, 11] and Audio Flamingo [1, 14, 19] address a wide range of audio tasks, while music-specific foundation models [17, 20] demonstrate that domain-specific pre-training consistently outperforms general-purpose alternatives on music audio tasks. Despite this progress across text, vision, and audio, to the best of our knowledge, *no foundation model exists for sheet music*—a highly structured visual domain with unique symbolic properties.

In this paper, we introduce MuSViT (**Mu**sic **S**core **Vi**sion **T**ransformer), a ViT encoder pre-trained on 9.7 million sheet music pages from the IMSLP[1] using Masked Image Modeling (MIM) [16]. MuSViT learns rich visual representations by reconstructing masked regions of full-page sheet music images, thus capturing the structural and symbolic properties of music notation without requiring any annotated data. An overview of MuSViT is shown in Fig. 1.

To assess the quality of the learned representations, we evaluate MuSViT on four diverse downstream tasks from the related literature: *full-page music score recognition* (page-level transcription) [32], *staff-level music score recognition* (staff-level transcription) [24], *music symbol detection* (dense object localization) [21], and *score difficulty classification* (document-level estimation) [29]. Following standard practice for evaluating foundation models [2, 6, 11, 19, 20, 27], we adopt two complementary protocols: *linear probing*, where the encoder is frozen and only lightweight task-specific heads are trained, which directly measures the quality of the extracted representations against other pre-trained vision encoders; and *fine-tuning*, where the encoder is trained jointly with a downstream head, which measures transfer-learning capacity and enables a fair comparison against state-of-the-art task-specific methods. Our results show that MuSViT achieves strong performance in both scenarios (see Fig. 2), highlighting the importance of music-specific representations for sheet music tasks.

Beyond task performance, an embedding-transcription consistency analysis demonstrates that MuSViT representations align more closely with symbolic musical content than those of general-purpose vision encoders. This provides direct evidence that domain-specific pre-training captures musically meaningful structures and semantics.

The contributions of this work are summarized as follows:

[1] https://imslp.org/

1. We introduce and publicly release MuSViT, the first vision foundation model for sheet music representation. MuSViT is a ViT encoder pre-trained on 9.7 million pages of sheet music collected from the IMSLP via MIM.[2]
2. We evaluate MuSViT on four representative downstream tasks—full-page music score recognition, staff-level music score recognition, music symbol detection, and score difficulty classification—under two scenarios: *linear probing* (frozen encoder) and *fine-tuning*.
   (a) Under *linear probing*, MuSViT consistently outperforms general-purpose pre-trained vision encoders, suggesting that sheet music analysis requires representations that general computer vision models—regardless of scale or pre-training diversity—do not capture well.
   (b) Under *fine-tuning*, MuSViT surpasses task-specific state-of-the-art methods on three of the four tasks and matches the best results on the fourth, confirming that music-specific pre-training provides a strong initialization for task-specific adaptation.
3. Our embedding-transcription consistency analysis provides direct evidence that MuSViT representations correlate with symbolic musical content, in contrast to those of general-purpose vision models.

By publicly releasing MuSViT, we aim to provide the research community with a foundation representation layer that accelerates progress across a wide range of sheet music applications.

2 MuSViT: Music Score Vision Transformer

MuSViT (**Mu**sic **S**core **Vi**sion **T**ransformer) is a self-supervised Vision Transformer (ViT) designed to learn meaningful visual representations for sheet music images. These serve as a general backbone that can be reused across diverse downstream tasks. This section describes MuSViT in detail, covering the pre-training strategy, the encoder architecture, and the training data.

2.1 Pre-Training Strategy

Sheet music differs fundamentally from natural images: it is a visual encoding of a symbolic language, where every element—such as noteheads, stems, accidentals, or rests—has a precise semantic meaning in terms of duration and pitch, following strict music notation rules. This structured nature has two important implications for the design of a ViT-based pre-training strategy. First, musical symbols are small and dense, requiring fine-grained patches such that each patch captures only elementary notational content—i.e. a single notehead, stem, or accidental—rather than spanning entire notes or measures, which would cause distinct symbols to blend within a single token. Second, the rigid spatial grammar of music notation makes local context highly informative: the identity and

[2] The model, pre-training code, and evaluation scripts are available through the project page: `https://grfia.dlsi.ua.es/musvit`.

position of neighboring symbols strongly constrain what can appear in a masked region.

Among MIM approaches [16], we adopt Masked Autoencoders (MAE) [15] for pre-training. MAE learns visual representations by reconstructing randomly masked portions of the input image. With a high masking ratio, entire measures and long sequences of symbols are occluded at once. To reconstruct these missing regions, the encoder cannot simply interpolate textures as it might for natural images; instead, it must infer which symbols should appear (duration), where they should be placed vertically on the staff (pitch), and how they fit into the surrounding sequence (context). In other words, MAE forces the model to learn the visual language of music notation without any supervision. This is the main reason we choose MAE: its high masking ratio naturally aligns with the structured, symbolic nature of sheet music, requiring the encoder to develop a deep understanding of musical notation in order to solve the reconstruction task. A qualitative reconstruction example is shown in Fig. 3.

Additionally, MAE reconstruction objective operates directly on pixel values and requires no external tokenizer or pre-trained teacher. This helps avoid domain mismatch, since no foundation model for sheet music currently exists. Its asymmetric encoder-decoder design, where the encoder processes only visible patches, also makes training efficient despite the long patch sequences that result from the fine-grained patch sizes required by our domain.

We now describe the pre-training procedure. Given an RGB input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, we split it into a grid of non-overlapping $P \times P$ patches, yielding $N = \frac{H}{P} \times \frac{W}{P}$ patches per image. We then follow the two-stage curriculum learning strategy, with varying masking ratios:

- **Stage 1 (Synthetic warm-up).** We initialize training on the *DeepScores V2* dataset [37], a synthetic corpus with lower visual variability and cleaner notation. This stage is not merely beneficial but necessary since we observed that training directly on real-world sheet music results in highly unstable and unreliable convergence, with the model predicting average pixel values rather than structured content. Therefore, starting on synthetic data allows the encoder to learn basic structures before encountering the high complexity of real-world scores. We use a masking ratio of 50% to provide more visible context and ease convergence. Training operates on random 512×512 crops with patch size $P = 16$, yielding $N = 32 \times 32 = 1{,}024$ patches per crop, ensuring each patch captures elementary notational content (e.g., a single notehead or accidental).
- **Stage 2 (Real-world adaptation).** We transfer to the full IMSLP corpus, exposing the model to the diversity of real-world sheet music. We increase the masking ratio to 70% and move from crops to full pages resized to $1{,}024\times 1{,}024$ pixels, maintaining patch size $P = 16$ and yielding $N = 64\times 64 = 4{,}096$ patches per image. This transition introduces full-page context—including global layout structure, multi-system arrangements, and page-level spatial relationships—while the consistent patch size ensures that the local features learned in Stage 1 remain applicable.

An ablation comparing the proposed curriculum against single-stage MAE training directly on IMSLP confirms that this two-stage design is necessary: without the synthetic warm-up, the encoder suffers dimensional collapse [18, 41]—the decoder predicts average patches rather than music-notation structures. A detailed analysis of the singular value spectrum and effective rank is provided in the "Supplementary Material".

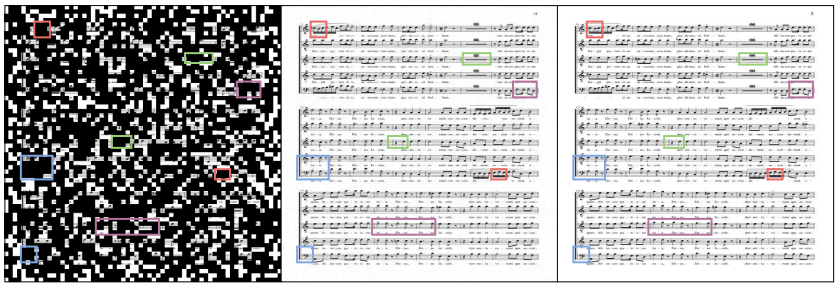

Fig. 3: MuSViT reconstruction example. **Left:** Masked music score image. **Middle:** MuSViT reconstruction. **Right:** The original sheet music. The coloured rectangles highlight different reconstruction components: *Red* indicates pitch sequence reconstruction (staff position); *Blue* represents clef reconstruction; *Green* marks rest reconstruction; and *Purple* denotes musical note sequence reconstruction (musical symbol aligned with staff position).

Unlike distillation-based approaches such as DINO [9, 25], which require aggressive augmentation (including Gaussian noise that can easily erase staff lines—destroying pitch information, a well-known challenge in music document processing [13]), the MAE reconstruction objective naturally encourages robust learning with minimal augmentation. We apply only random resized cropping to preserve the structural integrity of musical notation. Full pre-training details (learning rate, batch size, epochs, warm-up schedule, weight decay) are provided in the "Supplementary Material".

2.2 Architecture

The MuSViT encoder employs a ViT architecture [12], comprising 12 Transformer layers with approximately 85M parameters. Each $P \times P$ patch from the input image is flattened and linearly projected into a $d = 768$-dimensional embedding vector. Following MAE design [15] and prior work on music document recognition [32], we use 2D sinusoidal positional encodings (PE) rather than the standard 1D PE, as 2D PE explicitly encodes vertical position and helps the model identify elements along the height of the document—information that 1D PE forces the model to learn implicitly. During pre-training, only the visible (unmasked) patches are processed by the encoder, as described in Section 2.1.

Adhering to MAE design philosophy [15], a lightweight decoder is used exclusively during pre-training and discarded afterward. The encoded visible patch

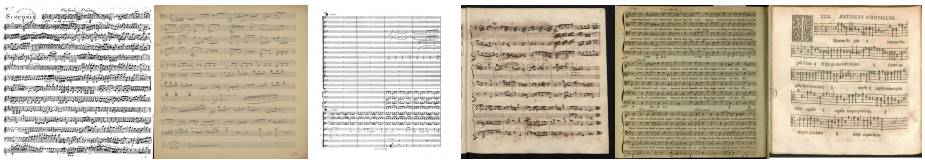

Fig. 4: Representative pages from the IMSLP pre-training corpus, illustrating its visual diversity. The collection spans historical periods, notation systems, engraving styles, and musical textures.

embeddings are projected to the decoder dimension, after which learnable mask tokens are inserted into the sequence at the positions of missing patches. As with the encoder, 2D PE is applied to the full sequence to maintain spatial correspondence. Each output at a masked position is linearly projected to reconstruct the patch pixel values. The training loss is the mean squared error (MSE) between predicted and target pixels, applied only to masked patches. Full architecture details are provided in the "Supplementary Material".

In addition to MuSViT, we train a lightweight variant, MuSViT$_{\text{Light}}$, which shares the same 12-layer Transformer architecture but uses a smaller embedding dimension ($d = 384$), totalling approximately 25M parameters. MuSViT$_{\text{Light}}$ targets limited computational resources scenarios, while MuSViT is intended for maximum performance.

2.3 IMSLP training data

MuSViT is pre-trained on a large-scale corpus of public-domain sheet music pages collected from the International Music Score Library Project (IMSLP).[3] The IMSLP hosts hundreds of thousands of scanned music scores spanning multiple centuries, composers, and engraving styles. Unlike synthetic or homogeneous corpora, this collection captures a wide spectrum of real-world sheet music variability—from monophonic vocal lines to dense orchestral scores, across both modern typeset and historical engraving conventions—providing the visual diversity necessary for learning generalizable representations. Specifically, we collect 9.7 million pages drawn from around 400,000 distinct musical works. Each page is rendered as an RGB image. Representative samples are shown in Fig. 4; additional examples are provided in the "Supplementary Material".

3 Evaluation on Downstream Tasks

Sheet music analysis encompasses a wide range of perceptual and structural challenges, from localizing individual symbols to transcribing entire pages and estimating high-level score properties. We probe MuSViT representations across four tasks that collectively cover this spectrum: *full-page music score recognition* [32], *staff-level music score recognition* [24], *music symbol detection* [21],

[3] All data used for pre-training is sourced from publicly available scans that are in the public domain or permissively licensed under the IMSLP terms of use.

and *score difficulty classification* [29]. We believe that performance across this diverse set of tasks serves as a reliable proxy for the generality of MuSViT representations for sheet music analysis. Table 1 summarizes the datasets and metrics. For each task, we consider two evaluation scenarios:

1. **Linear probing.** The MuSViT encoder remains frozen and only lightweight, task-specific heads are trained on top of the extracted representations. This protocol isolates the intrinsic quality of the learned features and provides a controlled comparison against other pre-trained vision encoders, all of them evaluated under identical frozen conditions.
2. **Fine-tuning.** The MuSViT encoder is unfrozen and trained jointly with the task-specific head, allowing the representations to adapt to each downstream objective. Since each task has its own established training and evaluation protocols, we follow the respective protocols in each case to ensure a fully comparable evaluation against state-of-the-art methods. Full training details are provided in the "Supplementary Material".

Table 1: Overview of the four downstream tasks used to evaluate MuSViT. Each task targets a different level of sheet music understanding—from sequence-level transcription (full-page and staff-level recognition) to dense spatial localization (symbol detection) and document-level semantics (difficulty classification). We report the datasets and primary evaluation metrics for each task.

Task	**Description**	**Datasets**	**Metric** (%)
Full-Page Music Score Recognition	Transcribe full music score pages	*Mozarteum*[4], *Polish Digital Scores*[5]	SER ↓
Staff-Level Music Score Recognition	Transcribe single staff-level images	*Capitan* [8], *Guatemala* [35], *Il Lauro Secco* [26], *AMDC* [23], *FMT* [31]	SER ↓
Music Symbol Detection	Localize and classify music symbols	*DeepScoresV2* [37]	mAP, w-mAP ↑
Score Difficulty Classification	Estimate performance difficulty from score images	*FreeScores* [29], *Can I Play It?* [29], *PianoStreet* [29]	Acc_0, Acc_1 ↑

Together, these two scenarios disentangle the contribution of pre-trained representations (linear probing) from the model capacity when allowed to specialize (fine-tuning), offering a comprehensive view of MuSViT as both a general-purpose foundation layer and a competitive task-specific model. Due to space constraints, we here report results averaged across datasets for all tasks; per-dataset breakdowns are available in the "Supplementary Material".

We compare MuSViT against four general-purpose vision encoders representing complementary pre-training paradigms: (i) **DINOv3-7B** [33], latest large-scale self-supervised vision model; (ii) **Qwen3-VL** [4], state-of-the-art vision-language model pre-trained on large-scale image–text corpora; (iii) **PaliGemma 2** [34], vision-language model that combines a SigLIP vision encoder pre-trained on diverse language–image pairs with a language model; and

[4] https://dme.mozarteum.at/movi/en
[5] https://github.com/pl-wnifc/humdrum-polish-scores

(iv) **Kosmos-2.5** [22], document-oriented multimodal model designed for structured visual inputs such as forms and documents. Only the vision encoder component is used for evaluation. To our knowledge, none of these encoders has been exposed to sheet music during pre-training, making them direct controls for measuring the value of music-specific visual representations.

Figure 2 provides a high-level summary of results across all four tasks and both evaluation scenarios. Under *linear probing*, MuSViT achieves the largest enclosed area in the radar chart, outperforming all general-purpose vision encoders across every task. Under *fine-tuning*, MuSViT matches or surpasses task-specific state-of-the-art methods. The following subsections present the detailed results for each task.[6]

3.1 Full-Page Music Score Recognition

Full-page music score recognition is the task of transcribing an entire score image into its corresponding symbol-level sequence, considering the correct reading order. This task has classically been approached using a two-stage pipeline—where a layout analysis step first segments individual staves, which are then transcribed independently—but recent work has enabled end-to-end models that process the full page in a single pass [32]. This end-to-end formulation is a particularly demanding test of representation quality: the encoder must capture both fine-grained notational detail at the symbol level and the global spatial organization of the page.

We evaluate on two datasets of scanned pianoform scores: *Mozarteum* (101 pages, printed CWMN) and *Polish Digital Scores* (117 pages, printed CWMN). We adopt an autoregressive Transformer decoder [32] on top of the MuSViT encoder, which attends autoregressively to the patch embeddings to generate the output sequence. Performance is measured using the Symbol Error Rate (SER), defined as the normalized edit distance between predicted and ground-truth sequences [32].

Table 2 reveals a striking gap under linear probing: all general-purpose encoders produce SER values in the 48–62% range, while MuSViT achieves 16.4%—more than 2.5× lower than the best baseline (PaliGemma 2, 48.6%). Remarkably, MuSViT already outperforms the state-of-the-art performance [32] (20.0%). These results show that MuSViT simultaneously encodes fine-grained symbol-level detail and the global reading-order structure spanning the entire page, whereas no general-purpose encoder provides representations sufficient for this dual requirement.

Table 3 shows that under fine-tuning MuSViT achieves 10.9% average SER, outperforming the state of the art [32] by 9.1 points. Per-dataset results (see

[6] Additionally, we fine-tuned all general-purpose models on two representative tasks to verify that the gap is not an artifact of the frozen evaluation protocol; MuSViT remains superior even against fine-tuned four baselines while being 5–82× smaller in parameters and 16–260× more efficient in GFLOPs. These results and detailed computational cost comparison are provided in the "Supplementary Material".

Table 2: Average Symbol Error Rate (SER %, ↓) across all full-page recognition datasets under the *linear probing* scenario (frozen encoder). Best result is in **bold**; second best is underlined.

	PaliGemma 2 [34]	Kosmos-2.5 [22]	Qwen3-VL [4]	DINOv3-7B [33]	MuSViT	MuSViT$_{\text{Light}}$
Avg.	48.6	62.4	51.0	56.9	**16.4**	20.9

Table 3: Average Symbol Error Rate (SER %, ↓) across all full-page recognition datasets under the *fine-tuning* scenario. Best result is in **bold**; second best is underlined.

	State-of-the-art [32]	MuSViT	MuSViT$_{\text{Light}}$
Avg.	20.0	**10.9**	11.8

"Supplementary Material") show that this improvement is consistent across both corpora, with a particularly large gain on *Polish Digital Scores* where the prior system struggled most, suggesting that pre-training on large-scale diverse scores yields a more robust initialization than task-specific supervision alone.

3.2 Staff-Level Music Score Recognition

Staff-level music score recognition shares the same transcription objective as full-page recognition but operates on individually segmented staff images rather than complete pages, requiring a prior layout analysis step. Although this intermediate segmentation makes the transcription pipeline less end-to-end, staff-level recognition is among the most established benchmarks in music transcription research, providing a complementary evaluation at finer spatial granularity.

We evaluate MuSViT on five corpora spanning both historical and modern notation, as well as printed and handwritten engraving styles: *Capitan* (828 staves, handwritten mensural) [8], *Guatemala* (3,263 staves, handwritten mensural) [35], *Il Lauro Secco* (1,136 staves, printed mensural) [26], *AMDC* (308 staves, printed CWMN) [23], and *FMT* (1,305 staves, handwritten CWMN) [31]. We follow the recurrent neural network-based recognition head of [24] and report SER values.

Table 4 shows that under linear probing MuSViT achieves 18.4% average SER—the best result among all encoders, 2.6 points ahead of Qwen3-VL (21.0%). Per-dataset results (see "Supplementary Material") follow the same trend. The gap over DINOv3-7B (32.1%) and Kosmos-2.5 (47.5%) is large, confirming that neither scale nor linguistic pre-training substitutes for music-specific visual representations.

Table 5 shows that under fine-tuning MuSViT reaches 8.6% average SER, within 0.6 points of the task-specific state of the art [24] (8.0%). This narrow gap—smaller than the advantage seen in full-page recognition—is consistent with the reduced spatial complexity of the staff-level setting: because each input contains a single row of notation, the encoder does not need to simultaneously

Table 4: Average Symbol Error Rate (SER %, ↓) across all staff-level recognition datasets under the *linear probing* scenario (frozen encoder). Best result is in **bold**; second best is underlined.

	PaliGemma 2 [34]	Kosmos-2.5 [22]	Qwen3-VL [4]	DINOv3-7B [33]	MuSViT	MuSViT$_{\text{Light}}$
Avg.	23.9	47.5	21.0	32.1	**18.4**	23.0

Table 5: Average Symbol Error Rate (SER %, ↓) across all staff-level recognition datasets under the *fine-tuning* scenario. Best result is in **bold**; second best is underlined.

	State-of-the-art [24]	MuSViT	MuSViT$_{\text{Light}}$
Avg.	**8.0**	8.6	9.8

manage global page layout and local symbol detail, leaving less room for music-specific pre-training to provide an additional edge.

3.3 Music Symbol Detection

Music symbol detection is the task of simultaneously localizing and classifying all individual notation elements within a score image, producing a set of bounding boxes with class labels for every visible symbol (noteheads, stems, accidentals, rests, clefs, *etc.*). Unlike the transcription tasks above, this is a dense spatial task: it requires the model to reason about 2D locations rather than reading order.

We evaluate on *DeepScoresV2* [37] (1,714 scores, 135 symbol classes) using a Faster R-CNN [30], reporting mAP and w-mAP for linear probing. Concerning fine-tuning, we consider mAP_{50} to mirror the current state of the art [21]; full implementation details are in the "Supplementary Material".

Table 6 shows that, under linear probing, MuSViT achieves 79.7% mAP and 80.7% w-mAP—both best among all encoders—with MuSViT$_{\text{Light}}$ within 0.6 and 0.3 points, respectively. The gap over DINOv3-7B is substantial on both metrics—9.3 points in mAP (70.4%) and 18.7 in w-mAP (62.0%). Vision-language models trail far behind. The large disparity between vision-only and language-aligned encoders suggests that language-aligned training actively degrades spatial precision on symbol-dense layouts.

Fine-tuning yields the most decisive gains in the entire evaluation as shown in Table 7: both MuSViT variants surpass 96% mAP_{50}, outperforming the state of the art [21] (90.5%) by more than 6 points. This margin is doubly notable: MuSViT uses a Faster R-CNN detector, considerably lighter than the Transformer-based architecture of the state of the art, and the encoder is only parameter-efficiently adapted via LoRA rather than fully fine-tuned (see "Supplementary Material")—both factors favour the baseline. The performance gap therefore reflects the quality of the pre-trained representations rather than a more powerful detection head or a more intense training.

Table 6: mAP and weighted mAP (↑) on the *DeepScores V2* dataset for music symbol detection under the *linear probing* scenario (frozen encoder). Best results are in **bold**; second best are underlined.

	PaliGemma 2 [34]	Kosmos-2.5 [22]	Qwen3-VL [4]	DINOv3-7B [33]	MuSViT	MuSViT$_{\text{Light}}$
mAP	31.7	42.7	67.1	70.4	**79.7**	79.1
w-mAP	39.0	47.4	61.0	62.0	**80.7**	80.4

Table 7: mAP$_{50}$ (%, ↑) on the *DeepScores V2* dataset for music symbol detection under the *fine-tuning* scenario. Best results are in **bold**; second best are underlined.

	State-of-the-art [21]	MuSViT	MuSViT$_{\text{Light}}$
mAP$_{50}$	90.5	**97.0**	96.6

3.4 Score Difficulty Classification

Score difficulty classification aims at estimating the performance difficulty of a musical piece directly from its sheet music images [29]. Prior research has primarily addressed this task using symbolic music representations, which offer high interpretability but are limited by their scarce availability and dependence on successful prior transcription steps [28]. In contrast, the image-based formulation operates directly on raw score pages, bypassing the need for intermediate transcription and enabling application to any digitized score. Unlike recognition or detection, this task demands representations that aggregate holistic page-level properties rather than resolving individual symbols or their spatial arrangement.

We evaluate MuSViT on three piano corpora: *FreeScores* (4,193 pieces, 5 difficulty levels) [29], *Can I Play It?* (652 pieces, 9 levels) [29], and *PianoStreet* (2,816 pieces, 9 levels) [29]. We report Acc_0 (exact match) and Acc_1 (within one adjacent level) [29]. We also mirror the evaluation protocol of [29], which compares simpler and more complex classification heads and reports best results with the latter. Therefore, for linear probing, we adopt the simpler head: per-page patch embeddings are mean-pooled into a document-level embedding and passed to an MLP classifier on top of the frozen encoder; whereas for fine-tuning, we adopt a more expressive head: a GRU-based recurrent network that processes the sequence of page embeddings sequentially.

Table 8 show that, under linear probing, MuSViT achieves 47.4% Acc_0 and 87.1% Acc_1, the best results among all encoders. Notably, MuSViT$_{\text{Light}}$ ties MuSViT on Acc_0, suggesting that even the lightweight variant captures the global page-level properties relevant to difficulty estimation. Both MuSViT variants already surpass the current state of the art [29] (38.4% Acc_0, 84.3% Acc_1) by 9.0 and 2.8 points respectively. The relatively small gap between MuSViT and the best general-purpose encoder (PaliGemma 2, 46.8% Acc_0) is consistent with the holistic nature of this task, where page-level visual statistics alone already provide a useful difficulty signal.

In the fine-tuning scenario, reflected by Table 9, MuSViT achieves 54.2% Acc_0 and 89.3% Acc_1, outperforming the state of the art [29] by 15.8 and 5.0

Table 8: Average Acc_0 and Acc_1 across all score difficulty classification datasets under the *linear probing* scenario (frozen encoder). Best results are in **bold**; second best are underlined.

	PaliGemma 2 [34]	Kosmos-2.5 [22]	Qwen3-VL [4]	DINOv3-7B [33]	MuSViT	MuSViT$_{\text{Light}}$
Avg. Acc$_0$	46.8	34.3	43.3	45.5	**47.4**	**47.4**
Avg. Acc$_1$	83.9	69.6	83.1	83.9	**87.1**	84.4

Table 9: Average Acc_0 and Acc_1 (%, ↑) across all score difficulty classification datasets under the *fine-tuning* scenario. Best results are in **bold**; second best are underlined.

	State-of-the-art [29]	MuSViT	MuSViT$_{\text{Light}}$
Avg. Acc$_0$	38.4	**54.2**	54.0
Avg. Acc$_1$	84.3	**89.3**	88.9

points, respectively. MuSViT$_{\text{Light}}$ matches these results closely (54.0% Acc_0, 88.9% Acc_1), with a margin of only 0.2 and 0.4 points. This represents the smallest gap between MuSViT variants observed across all tasks, further reinforcing that difficulty estimation is well served by holistic representations rather than fine-grained encoder capacity.

4 Embedding-Transcription Consistency Analysis

Although downstream task performance provides evidence of representation quality, it does not directly reveal whether the learned embeddings capture music-notation semantics. To address this question, we perform an embedding-transcription consistency analysis that quantifies how strongly the visual embedding space aligns with symbolic musical content (i.e., transcription). The hypothesis is simple: if MuSViT learns music-aware representations, then pairs of score images with similar transcriptions should produce similar embeddings, and pairs with different transcriptions should be farther apart in the embedding space. We expect MuSViT to exhibit stronger embedding–transcription alignment than general-purpose vision models, validating that music-specific pretraining captures symbolic musical structure rather than purely visual patterns.

We conduct this analysis on the *Mozarteum* and *Polish Digital Scores* datasets, for which both score images and symbolic transcriptions are available. For each image i, we extract an embedding vector $\mathbf{e}_i$ by flattening all token representations produced by the frozen encoder into a single vector. We compare MuSViT against PaliGemma 2 [34], Kosmos-2.5 [22], Qwen3-VL [4], and DINOv3-7B [33], all evaluated under identical frozen conditions.

To quantify similarity, we compute pairwise distances in two spaces. In the *embedding space*, we use the Euclidean distance between embedding pairs, yielding $D^{\text{emb}} \in \mathbb{R}^{N \times N}$ with $D^{\text{emb}}_{ij} = \|\mathbf{e}_i - \mathbf{e}_j\|_2$. In the *transcription space*, we compute two complementary measures: (1) **Edit distance** D^{edit}, applying Levenshtein distance to transcription pairs, capturing differences in symbolic content

Table 10: Pearson and Spearman correlation between pairwise embedding distances and pairwise transcription distances (edit and histogram) for each encoder. Higher positive values indicate stronger alignment between the visual embedding space and symbolic musical content. Best results are in **bold**; second best are underlined.

		PaliGemma 2	Kosmos-2.5	Qwen3-VL	DINOv3-7B	MuSViT	MuSViT$_{\text{Light}}$
Pearson Correlation	ρ_p^{ed}	-0.110	-0.080	-0.041	-0.080	0.606	**0.618**
	ρ_p^{h}	-0.127	-0.130	-0.052	-0.100	**0.665**	0.646
Spearman Correlation	ρ_s^{ed}	-0.120	-0.114	-0.009	-0.135	**0.691**	0.658
	ρ_s^{h}	-0.132	-0.153	-0.009	-0.152	**0.714**	0.662

and sequential ordering; and (2) **Histogram distance** D^{his}, computing distance between token frequency distributions, measuring compositional differences in notation while ignoring order.

To assess embedding–transcription consistency, we compute both Pearson (ρ_p) and Spearman (ρ_s) correlation [5,39] between corresponding entries of D^{emb} and each transcription distance matrix, yielding four values per encoder: ρ_p^{ed}, ρ_p^{h}, ρ_s^{ed}, and ρ_s^{h}. A higher positive correlation indicates that the embedding space preserves symbolic similarity, i.e., transcriptionally similar images are placed closer in the embedding space.

Table 10 reports the embedding–transcription correlation results. All general-purpose encoders yield low negative correlations across both distance measures and both correlation metrics, meaning their embedding spaces are not merely uncorrelated with musical content, but they are slightly anti-correlated: images with similar content might be placed farther apart in the embedding space. Because these encoders are optimized for visual appearance, they can be misled by surface-level cues that do not necessarily reflect the underlying musical structure. Therefore, visually similar pages may encode different musical content, and vice versa. In contrast, MuSViT variants yield consistently high positive correlations, confirming that music-specific models are necessary for the encoder to capture symbolic musical structure. The larger gap on histogram distance suggests that MuSViT models are particularly effective at capturing the distribution of music notation. We further complement this research with two analyses in the "Supplementary Material": an attention heat map that visualizes which image regions MuSViT focuses on, and a k-curve analysis that examines local embedding structure across neighborhood sizes.

5 Conclusions

We introduce MuSViT, the first foundation vision model for sheet music representation: a ViT pre-trained via MAE on 9.7 million public-domain score pages from the IMSLP through a two-stage curriculum progressing from synthetic to real-world data. As a result, MuSViT representations are grounded in the

structure of music notation. Across four downstream tasks, MuSViT generally outperforms other vision encoders under linear probing and task-specific state-of-the-art systems under fine-tuning. This advantage spans tasks with fundamentally different demands—from spatially precise symbol detection, which requires dense per-symbol localization, to classification-oriented score difficulty estimation, which requires global semantic understanding—demonstrating that the learned representations are broadly useful rather than narrowly specialized. Underlying this performance gap is a more fundamental result: sheet music is a genuinely distinct visual domain that general-purpose models fail to represent adequately. Our embedding-transcription consistency analysis shows that general-purpose encoders produce representations that do not correlate with musical content, while MuSViT encodes symbolic musical structure directly in its embedding space. Therefore, a domain-specific model is not merely beneficial for sheet music representation but necessary.

Acknowledgements

The authors gratefully acknowledge Edward Guo, on behalf of IMSLP/Petrucci Music Library, for providing access to the data used to train the models.

This publication is part of the LEMUR project PID2023-148259NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The first author is supported by the University of Alicante through the FPU Program (UAFPU22-19). The third author is supported by a predoctoral contract associated with the LEMUR project. The fourth author is supported by a predoctoral contract from grant CISEJI/2023/9 "Programa para el apoyo a personas investigadoras con talento (Plan GenT) de la Generalitat Valenciana".

References

1. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada (Jul 2025)
2. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation Models Defining a New Era in Vision: A Survey and Outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2245–2264 (2025)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen Technical Report (2023)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song,

S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report (2025)
5. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
6. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai, X.: PaliGemma: A versatile 3B VLM for transfer (2024)
7. Calvo-Zaragoza, J., Jr, J.H., Pacha, A.: Understanding Optical Music Recognition. ACM Computing Surveys **53**(4), 1–35 (2020)
8. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten Music Recognition for Mensural notation with convolutional recurrent neural networks. Pattern Recognition Letters **128**, 115–121 (2019)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9650–9660. IEEE Computer Society, Online (Jun 2021)
10. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J.: Qwen2-Audio Technical Report (2024)
11. Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., Zhou, J.: Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models (2023)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proceedings of the 9th International Conference on Learning Representations. Online (May 2021)
13. Fujinaga, I.: Staff detection and removal. In: Visual Perception of Music Notation: On-Line and Off Line Recognition, pp. 1–39. IGI Global Scientific Publishing (2004)
14. Goel, A., Ghosh, S., Kim, J., Kumar, S., Kong, Z., Lee, S.g., Yang, C.H.H., Duraiswami, R., Manocha, D., Valle, R., et al.: Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In: Proceedings of the 39th Annual Conference on Neural Information Processing Systems. Sydney, Australia (Dec 2025)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009. IEEE Computer Society, New Orleans, Louisiana (Jun 2022)
16. Hondru, V., Croitoru, F.A., Minaee, S., Ionescu, R.T., Sebe, N.: Masked image modeling: A survey. International Journal of Computer Vision **133**(10), 7154–7200 (2025)
17. Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J.Y., Ellis, D.P.W.: MuLan: A Joint Embedding of Music Audio and Natural Language. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference. pp. 559–566. ISMIR, Bengaluru, India (Dec 2022)

18. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348 (2021)
19. Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., Catanzaro, B.: Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria (Jul 2024)
20. LI, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., Fu, J.: MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria (May 2024)
21. Luo, F., Dai, Y., Fuentes, J., Ding, W., Zhang, X.: M-DETR: multi-scale DETR for optical music recognition. vol. 249, p. 123664. Elsevier (2024)
22. Lv, T., Huang, Y., Chen, J., Zhao, Y., Jia, Y., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., Wu, S., Wang, G., Zhang, C., Wei, F.: KOSMOS-2.5: A Multimodal Literate Model (2025)
23. Madueño, A., Ríos-Vila, A., Rizo, D.: Automatized incipit encoding at the Andalusian Music Documentation Center. In: Proceedings of the 8th International Conference on Digital Libraries for Musicology. Association for Computing Machinery, Online (Jul 2021)
24. Martinez-Sevilla, J.C., Rizo, D., Calvo-Zaragoza, J.: Towards universal Optical Music Recognition: A case study on notation types. In: Proceedings of the 25th International Society for Music Information Retrieval Conference. pp. 914–921. ISMIR, San Francisco, USA (Nov 2024)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2024)
26. Parada-Cabaleiro, E., Batliner, A., Schuller, B.W.: A Diplomatic Edition of Il Lauro Secco: Ground Truth for OMR of White Mensural Notation. In: Proceedings of the 20th International Society for Music Information Retrieval Conference. pp. 557–564. ISMIR, Delft, The Netherlands (Nov 2019)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. Online (May 2021)
28. Ramoneda, P., Eremenko, V., D'Hooge, A., Parada-Cabaleiro, E., Serra, X.: Towards explainable and interpretable musical difficulty estimation. In: Proceedings of the 25th International Society for Music Information Retrieval Conference. pp. 520—-528. ISMIR, California, USA (Nov 2024)
29. Ramoneda, P., Valero-Mas, J.J., Jeong, D., Serra, X.: Predicting performance difficulty from piano sheet music images. In: Proceedings of the 24th International Society for Music Information Retrieval Conference. pp. 708–715. ISMIR, Milan, Italy (Nov 2023)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc., Barcelona, Spain (Dec 2015)

31. Ríos-Vila, A., Esplà-Gomis, M., Rizo, D., Ponce de León, P.J., Iñesta, J.M.: Applying Automatic Translation for Optical Music Recognition's Encoding Step. Applied Sciences **11**(9), 3890–3912 (2021)
32. Ríos-Vila, A., Calvo-Zaragoza, J., Rizo, D., Paquet, T.: End-to-End Full-Page Optical Music Recognition for Pianoform Sheet Music. International Journal of Computer Vision **134**, 49–66 (2026)
33. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)
34. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., Zhai, X.: PaliGemma 2: A Family of Versatile VLMs for Transfer (2024)
35. Thomae, M.E., Cumming, J.E., Fujinaga, I.: Digitization of Choirbooks in Guatemala. In: Proceedings of the 9th International Conference on Digital Libraries for Musicology. pp. 19–26. Association for Computing Machinery, Prague, Czech Republic (Jul 2022)
36. Tuggener, L., Emberger, R., Ghosh, A., Sager, P., Satyawan, Y.P., Montoya-Zegarra, J.A., Goldschagg, S., Seibold, F., Gut, U., Ackermann, P., Schmidhuber, J., Stadelmann, T.: Real World Music Object Recognition. Transactions of the International Society for Music Information Retrieval **7**(1), 1–14 (2024)
37. Tuggener, L., Satyawan, Y.P., Pacha, A., Schmidhuber, J., Stadelmann, T.: DeepScoresV2 (Sep 2020). `https://doi.org/10.5281/zenodo.4012193`
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution (2024)
39. Wissler, C.: The Spearman correlation formula. Science **22**(558), 309–311 (1905)
40. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986. IEEE Computer Society, Paris, France (Oct 2023)
41. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems **35**, 27127–27139 (2022)