Designing a Universal Modular Multi-Modal Student Model

Building a smaller "student" network that matches a large teacher's capabilities requires carefully addressing several challenges. First, the *teacher–student gap* (vast parameter and capacity differences) must be reduced so knowledge can transfer effectively. Recent work proposes inserting lightweight **adapter modules** into the teacher to "soften" its outputs and better match the student's scale. For example, the *Parameter-Efficient and Student-Friendly KD* (PESF-KD) framework trains adapter layers on the teacher so its predictions align more closely with what a smaller student can learn [1] [2]. The authors show this narrows the output gap, making distillation easier [1]. In general, carefully smoothing or combining the teacher's logits (e.g. via temperature scaling, feature-map distillation, or intermediate-layer matching) can help a tiny model absorb the teacher's knowledge.

Another key concern is *retaining all of the teacher's capabilities* without catastrophic forgetting. Standard fine-tuning of a small model on some tasks often degrades its performance on others. Techniques like **Learning without Forgetting** (LwF) explicitly address this by using only the new-task data while imposing a distillation loss on the teacher's original outputs [3] [4]. LwF trains the student to produce the same responses as the teacher on the old tasks, even while learning new ones [3] [4]. In practice this means adding a loss that keeps the student's pre-softmax outputs close to the teacher's for shared tasks, which preserves the teacher's pre-existing skills (e.g. its original classifier or voice-generation behavior) while adapting to new constraints [3] [4]. Together, smoothing the teacher and using retention-focused loss terms can mitigate capacity mismatch and forgetting.

Figure: A modular architectural design allows independent training and reuse of components [5] [6]. Another powerful strategy is **modular, multi-tower architectures**. Instead of one giant end-to-end network, we can compose many smaller modules or “towers” that handle sub-tasks or modalities. Castillo-Bolado et al. argue that monolithic end-to-end networks face scaling limits, and propose *modular training*: each neural module is trained independently and then assembled into a larger network [5]. This enables reuse (trained modules can be kept for future tasks) and easier debugging. A recent survey also emphasizes that modularity improves *interpretability, scalability, and reusability*, since components can be swapped or recombined [6] [7]. For example, two-tower (dual-encoder) models – common in vision-language systems – use one “image tower” and one “text tower” to project each modality into a shared space [8] [9]. These separate towers can each be trained on its modality, then jointly aligned. Likewise, **Mixture-of-Experts (MoE)** architectures split the network into many smaller expert modules: a gating network routes each input to one or more experts, so only a subset of parameters is active per example. An MoE can dramatically increase capacity without linear cost, since each expert (module) learns a sub-task [7]. In short, designing the student as a set of reusable, independently trained components – possibly arranged as parallel “towers” – keeps the model flexible, lean, and capable of capturing diverse teacher behaviors.

In the multi-modal setting (e.g. vision + language + audio), specialized distillation methods help preserve all of the teacher’s features. For instance, the **EPIC** framework compresses a multimodal language model by introducing *progressive consistency distillation*. It partitions the distillation problem along token and layer dimensions, using the teacher to guide each stage. In practice EPIC imposes **token-level** and **layer-level** consistency losses that keep the student’s intermediate representations close to the teacher’s [10]. This stepwise approach significantly reduces training difficulty and cost while maintaining accuracy [11] [10]. Similarly, *DIME-FM* distills a large vision-language model (e.g. CLIP) into a small dual-encoder student: it freezes the teacher and uses unpaired image/text data to match embedding similarity scores, ensuring the student retains transferability [12]. Notably, such approaches allow choosing a much smaller feature dimension for the student’s encoders than the teacher’s, yet still learn a shared space that works well on downstream tasks. For speech and audio, **DM-Codec** offers an elegant solution: it unifies language and speech representations by using the teacher’s language model as a “guide” during distillation. The student speech model is trained so that its [CLS] token and other features align with contextual embeddings from the teacher LM [13] [14]. Crucially, the LM and large speech teacher are only used during training – they do not increase the student's runtime cost [14]. This yields a lightweight model that nonetheless captures rich contextual understanding. In practice, DM-Codec's distilled text-to-speech model (*DM-Codec-TTS*) achieved state-of-the-art voice quality: for example it reached a mean opinion score (MOS) of ~3.70 on LibriSpeech and 3.78 on VCTK, beating larger baselines despite its smaller size [15]. In summary, multi-modal distillation methods like EPIC, DIME-FM, and DM-Codec can compress high-capacity vision, language, and audio models into fast, compact students **without sacrificing key features** (e.g. voice cloning performance or zero-shot generalization).

Finally, we can **inject domain knowledge** to further preserve teacher features. Techniques like **Neuron-Importance-aware Weight Transfer (NIWT)** have been proposed to ground new model neurons in human-understandable concepts [16]. NIWT maps free-form class or attribute descriptions onto internal neurons of a deep net, then uses these neuron importance vectors to initialize or adjust the student's weights. In effect, the student explicitly "knows" semantic concepts that the teacher encodes (e.g. voice characteristics, visual attributes, etc.), which guides learning novel tasks. Incorporating such guided transfer ensures that specialized capabilities of the teacher – like speaker identity traits for voice cloning – can be maintained in the student's parameters.

Key strategies summary: To build the universal compact model as described, one must (a) apply **student-friendly distillation** that narrows the teacher–student gap (e.g. adapter-based PESF-KD [1]); (b) use **continual-learning losses** (e.g. LwF) to prevent forgetting [3] [4]; (c) embrace a **modular multi-tower architecture** so that each modality or function can be developed and scaled independently [5] [6]; (d) employ **multimodal distillation techniques** (EPIC, DIME-FM, DM-Codec) that leverage the teacher's guidance on compressed inputs to retain accuracy [10] [13]; and (e) optionally inject **semantic priors** (e.g. via NIWT [16]) to lock in the teacher's special features. Combining these approaches in a carefully engineered pipeline will mitigate known pitfalls (capacity mismatch, forgetting, modality gaps) and avoid introducing new ones, yielding a small, fast student that faithfully preserves the teacher model's rich capabilities [1] [15].

Sources: We draw on recent distillation research (e.g. PESF-KD [1], EPIC [10], DIME-FM [12], DM-Codec [13] [15]), modular network design studies [5] [6], and continual-learning methods (LwF [3] [4]) to inform these strategies. These works demonstrate principled ways to shrink model size and inference cost while retaining the original model's performance.

---

[1] [2] arxiv.org
https://arxiv.org/pdf/2205.15308

[3] [4] arxiv.org
https://arxiv.org/pdf/1606.09282

[5] Design and independent training of composable and reusable neural modules - ScienceDirect
https://www.sciencedirect.com/science/article/pii/S0893608021001222

[6] [7] (PDF) Modularity in Deep Learning: A Survey
https://www.researchgate.net/publication/373249858_Modularity_in_Deep_Learning_A_Survey

[8] Two-Tower Embedding Model - MLOps Dictionary | Hopsworks
https://www.hopsworks.ai/dictionary/two-tower-embedding-model

[9] [12] DIME-FM : DIstilling Multimodal and Efficient Foundation Models
https://openaccess.thecvf.com/content/ICCV2023/papers/Sun_DIME-FM__DIstilling_Multimodal_and_Efficient_Foundation_Models_ICCV_2023_paper.pdf

[10] [11] Efficient Multi-modal Large Language Models via Progressive Consistency Distillation | OpenReview
https://openreview.net/forum?id=gZjPllL9jM&referrer=%5Bthe%20profile%20of%20Zhaorun%20Chen%5D(%2Fprofile%3Fid%3D~Zhaorun_Chen1)

[13] [14] [15] aclanthology.org
https://aclanthology.org/2025.findings-emnlp.1394.pdf

[16] nips2018vigil.github.io
https://nips2018vigil.github.io/static/papers/accepted/38.pdf