# MODALSHARE: CONTRIBUTION-AWARE BUDGET ALLOCATION FOR MULTIMODAL SPLIT LEARNING

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

## ABSTRACT

Multimodal models combine several input streams, yet training them where the data originates, such as on edge devices carrying cameras, microphones, and other sensors, is impractical when each modality requires its own heavy encoder. Split Learning (SL) makes such training feasible by keeping only the first layers on the device and offloading the rest to a server, but the resulting communication must carry intermediate activations for every modality at every step. Communication-efficient SL reduces that cost by compressing the activations, keeping only a fraction of each; however existing schemes hand every modality the same keep-ratio, so the shared budget is split in proportion to activation dimension, a quantity unrelated to how much each modality contributes to the fused prediction. We make that split an explicit decision, namely *inter-modality allocation*: under a fixed budget, every policy transmits the same expected payload and differs only in how that payload is divided across modalities. Our allocator, ModalShare, sets each modality's keep-ratio from a Shapley contribution score computed by the server over coalitions of activations it has already received, so measurement adds no transmission overhead, no client-side computation, and assumes no prior knowledge of which stream is which. ModalShare improves accuracy over equal keep-ratios by $15.4$ and $12.4$ percentage points on CREMA-D and MVSA at matched payload under $5\times$ compression, with consistent gains across three compressors, three datasets, and four budgets. Existing compressors leave these gains unclaimed in multimodal settings; contribution-aware allocation recovers them.

## 1 INTRODUCTION

Multimodal models have become the default choice for perception, and the sensors that feed them (e.g., cameras, microphones, accelerometers) are already standard on the devices people carry. Emotion recognition from speech and facial expression, sentiment analysis over paired image and text, and activity recognition from inertial streams all rest on combining modalities rather than choosing among them. A recurring finding is that the modalities are not equal partners: fused models weight them unequally and degrade differently when any single stream is constrained (Feng et al., 2026; Wei et al., 2025; Peng et al., 2022; Xu et al., 2025). This imbalance depends on the task and the data, so under a limited resource budget the streams should not be treated alike.

Resources are scarcest exactly where these models would be most useful: on the edge devices that hold the sensors and collect the training data, but also where memory and compute are tight. *Split Learning* (SL) removes this constraint (Vepakomma et al., 2018): the network is partitioned at a cut layer, so the client runs only the first few layers of each encoder and transmits the resulting intermediate ("smashed") activations to a server that completes the forward pass, computes the loss, and returns gradients at the cut for the client to complete the learning pass. The client's share can be made as small as the device demands, unlike Federated Learning (McMahan et al., 2017), which keeps data on device but still needs the full model to fit locally,thus making SL a natural fit for on-device multimodal training, as highlighted in edge architectures (Lin et al., 2024; Hafi et al., 2024).

The cost of SL is that the cut carries per-batch activations for the entire training run, and their size is the main bottleneck of the system (Lin et al., 2024; Oh et al., 2025), concentrated on the more constrained upstream direction. A substantial line of work therefore compresses what crosses the



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.

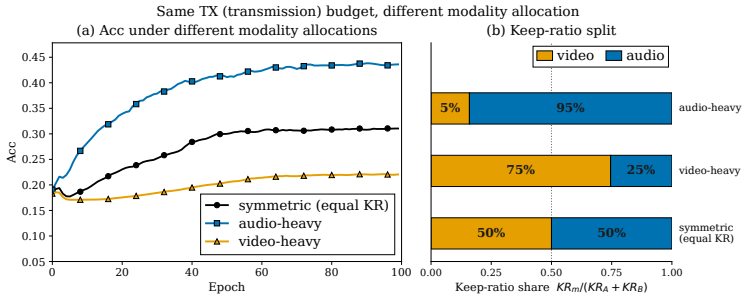



Figure 1: Fixed-budget keep-ratio allocations across modalities (CREMA-D, $5\times$ compression). All three arms transmit the same total payload under the same compressor and differ only in how it is divided (b); the resulting accuracy differs by a wide margin (a). Allocation, not only budget and compression, determines what a fixed uplink buys.

cut: sparsification, randomized selection, progressive pruning, channel-wise rate adaptation, and feature-wise dropout (Yuan et al., 2020; Zheng et al., 2023; Mudvari et al., 2024; Lin et al., 2026; Oh et al., 2025),each deciding how a single tensor is represented within its budget. Applied to a multimodal client, they are modality-agnostic by construction: one policy, the same keep-ratio handed to every stream.

A multimodal client, however, holds *multiple* tensors that share one uplink, which exposes a second decision the literature does not examine: how the shared budget is divided among them. Equal keep-ratios divide the payload in proportion to the cut dimensions, so a modality whose activations are $4\times$ larger silently receives $4\times$ the bandwidth; equal compression severity is not equal bandwidth. On MVSA-Single, an image–text dataset whose image cut is roughly four times the text cut, image takes $80\%$ of the budget before anything about the task is measured, and under a fixed budget, floats spent on a stream the prediction barely uses are floats taken from one it depends on. The effect is not an artifact of unequal dimensions: Figure 1 holds payload and compressor fixed on CREMA-D, an audio–video task whose two cuts are nearly equal in dimension, and varies only the split; an audio-heavy allocation converges well above the equal-keep-ratio default while a video-heavy one falls below it. Only the allocation differs.

The multimodal-learning literature has long recognized this asymmetry, but acts on it with tools that do not consider the budget constraints. Gradient modulation (Peng et al., 2022), imbalance-aware optimization (Wei et al., 2025), nested learning (Feng et al., 2026), and contribution-guided bottlenecks (Xu et al., 2025) change how the model is trained or how much capacity each modality is given; none changes how many bytes a modality sends, so under a fixed uplink budget they save nothing on the wire. The transmitted budget is a separate lever, and it is the main motivation here: which modality's activations are worth transmitting when the client cannot afford to send them all.

This motivates our question: under a fixed budget, does allocating it asymmetrically across modalities outperform uniform allocation, and under what conditions? We make the decision explicit as *inter-modality allocation* and design a policy for it. Our allocator, *ModalShare*, estimates each modality's marginal contribution with a Shapley probe evaluated at the server over coalitions of smashed activations the server already holds, and maps those scores to per-modality keep-ratios that preserve the total transmitted payload. Because the coalitions are formed by masking activations that have already crossed the cut, measurement incurs no added communication cost, adds no client-side computation, and needs no auxiliary channel; its only cost is server-side compute. Allocation sits above the compressor rather than replacing it: ModalShare decides how much budget each stream receives, while the compressor still decides which coordinates within that stream survive. The iso-budget formulation also keeps the comparison clean: every policy transmits the same expected payload under the same compressor, model, and schedule, so an accuracy difference is attributable to the split alone, not to a stronger backbone, more capacity, or more communication.

**Contributions.**

- **Inter-modality allocation**. We identify how a shared communication budget is divided across modalities as an unexamined degree of freedom, and formulate it as an iso-budget

problem: every policy transmits the same expected payload and differs only in its split, making any allocator comparable at matched payload.

- **ModalShare**. We set per-modality keep-ratios from a server-side Shapley contribution score computed over coalitions of already-received activations, so measurement adds no communication cost, no client-side computation, and no assumption about which stream is which. It improves accuracy over equal keep-ratios by up to $15.4$ pp across three compressors, three datasets, and four budgets; gains existing SL compressors leave unclaimed in multimodal settings.
- **The gain tracks measured contribution**. It persists on datasets with balanced cut dimensions, exceeds the dimension-skew component any fixed rule could recover, and vanishes where the contribution gap is small (UCI HAR); cut-local signals, such as activation norms, gradient norms and standalone utility do not recover it, since none observes how modalities combine.

## 2 Related Work

**Asymmetric multimodal learning.** The fact that modalities contribute unequally to a fused model, and that treating them uniformly is suboptimal, is by now well established in centralized training. Gradient-balancing methods such as on-the-fly gradient modulation (OGM) (Peng et al., 2022) equalize optimization effort across modalities. Wei et al. (2025) argue that balanced optimization is not itself the goal and instead align each modality's optimization dependency with the inverse of its variance ratio. Feng et al. (2026) decouple multimodal training into nested sub-processes that let each modality learn at its own pace, coordinating them through multi-timescale memory rather than a uniform monolithic update. Xu et al. (2025) guide asymmetry with an explicit contribution score, using it both to accelerate gradients and to set a per-modality information-bottleneck strength, while greedy modality selection (He et al., 2024) makes the coarser decision of which modalities to admit at all. Modality-balanced quantization (Li et al., 2025) makes a related observation at the level of numerical precision, calibrating vision and language tokens separately because a single quantization objective serves them unequally. These works reallocate optimization effort, representation capacity, or numerical precision, which are quantities set at training or design time, and none changes what the client must transmit. We reallocate the transmitted budget itself: the per-stream share of a fixed communication payload at the split-learning cut. The distinction is not which score is used but which quantity it controls: prior asymmetric-learning methods and ours can even draw on the same notion of contribution, yet an optimization- or capacity-side rule does not consider constraints such as communication budget.

**Contribution scoring.** Shapley-style attribution (Shapley et al., 1953) can help quantify how much each modality drives a prediction, and has been used to audit vision-language transformers (Parcalabescu & Frank, 2023), to separate unimodal from cross-modal contributions (Hu et al., 2022), and to report diagnostic weight in clinical models (Soenksen et al., 2022). Closest in estimator, Wei et al. (2024) compute Shapley-based contributions per sample and resample data for the weaker modality. Contribution is consistently found to be dataset- and task-dependent, which is exactly what makes a fixed budget split unsatisfactory and an online estimate useful. The question also arises in Federated Learning, where contribution estimates have been applied to communication (Yuan et al., 2026) and to modality imbalance across clients (Amini et al., 2025); there the network carries aggregated model updates rather than per-batch activations at a fixed cut, so the allocation decision we study does not arise. We evaluate the same coalitions during training and spend the resulting scores on transmitted payload: the estimator is shared with prior contribution work, but the actuator is the uplink budget. To our knowledge, this is the first use of a Shapley-based contribution score to allocate communication resources across modalities in multimodal learning.

**Split learning and communication compression.** Split Learning partitions a model so that clients transmit intermediate activations to a server that completes the forward and backward passes (Vepakomma et al., 2018), and the size of those intermediates is the dominant cost (Oh et al., 2025; Lin et al., 2024). A range of communication-efficient variants compress the transmitted tensor: Top-$K$ sparsification (Yuan et al., 2020), randomized variants (Zheng et al., 2023) that reduce selection and index cost, progressive pruning of the representation over training (Mudvari et al., 2024), adaptive feature-wise dropout at the cut (Oh et al., 2025), channel-wise compression that adapts the rate

across channels of a single tensor (Lin et al., 2026), batch-wise packing of multiple smashed vectors into one (Hsieh et al., 2022), and learned autoencoder bottlenecks at the cut (Meuwissen et al., 2026; Shao & Zhang, 2020). Each of these decides how a single tensor is represented: a sparsity level, a schedule, a channel-wise rate or a learned bottleneck. Multimodality adds a second: when several streams share one communication budget, it should be divided among them. That division is the subject of this work.

# 3 METHOD

We consider multimodal SL between two devices (one client and one server): the client holds the early modality encoders, the server the remaining backbone, fusion, and classifier. Because attribution and allocation act only on that client's cut and uplink, the same policy applies independently on every client–server link in a multi-client deployment. We formulate the sharing of the cut as an iso-budget allocation problem and give a contribution-aware policy for it; the policy is agnostic to the conserved budget, which here is the transmitted uplink payload. Notation is summarized in Appendix K.

## 3.1 MULTIMODAL SPLIT LEARNING

Let $\mathcal{D} = \{(x_i^{(1)}, \ldots, x_i^{(M)}, y_i)\}_{i=1}^N$ be a training set with modality index set $\mathcal{M} = \{1, \ldots, M\}$. A single client hosts $M$ modality encoders; for modality $m$, encoder $g_m$ with parameters $\theta_m$ maps its input to intermediate activations (smashed activations) at a fixed cut,

$$z_m = g_m(x^{(m)}; \theta_m) \in \mathbb{R}^{D_m}, \tag{1}$$

where $D_m$ is the flattened smashed-activation dimension. The server model $h$ with parameters $\theta_s$ receives the smashed activations, completes any remaining modality-specific layers, fuses the modalities, and predicts

$$\hat{y} = h(\{z_m\}_{m \in \mathcal{M}}; \theta_s). \tag{2}$$

For each mini-batch the client uploads $\{z_m\}$; the server evaluates the task loss, backpropagates through $h$, and returns the cut gradients $\nabla_{z_m} \ell$. All communication crosses this cut.

## 3.2 COMPRESSION OF SMASHED ACTIVATIONS

Let $\mathcal{C}$ denote a modality-level compressor with budget parameter $r_m \in (0, 1]$,

$$\hat{z}_m = \mathcal{C}(z_m, r_m), \tag{3}$$

where the keep-ratio $r_m$ is the expected payload of $\hat{z}_m$ as a fraction of $D_m$. This interface separates two decisions: the compressor determines *how* to represent modality $m$ within its budget, the allocator *how much* budget it receives; our rule is therefore independent of the selection rule inside $\mathcal{C}$.

Budget is accounted in transmitted smashed-activation floats,

$$\mathrm{TX} = \sum_{m \in \mathcal{M}} D_m r_m, \tag{4}$$

per sample and per direction. The return path carries gradients in the same compressed representation (i.e. coordinates the client did not send receive no gradient) so total cut traffic is $2 \cdot \mathrm{TX}$, and two allocations at equal TX exchange the same expected payload in both directions, side information (masks, indices, scales) included.

## 3.3 INTER-MODALITY ALLOCATION

Existing compression applies one rule to concatenated activations, or the same keep-ratio to every modality; neither asks how a shared budget should be divided according to the modalities' roles in the fused prediction. We expose that decision as inter-modality allocation. Define the normalized budget $\beta = \mathrm{TX} / \sum_m D_m$; all policies compared at $\beta$ satisfy the iso-budget constraint

$$\beta = \frac{\mathrm{TX}}{\sum_m D_m}, \qquad \sum_m D_m r_m = \beta \sum_m D_m, \tag{5}$$

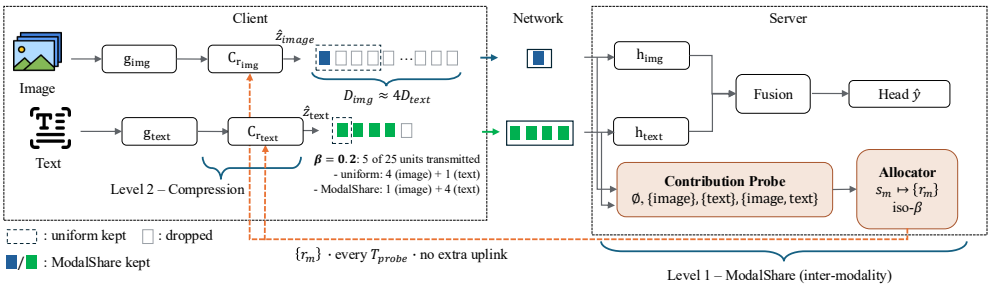



Figure 2: **ModalShare allocates a fixed cut budget across modalities.** Prior compression operates within a modality (Level 2); ModalShare divides the shared payload between them (Level 1) using a server-side Shapley probe on already-received activations, so no extra uplink is consumed. Strip lengths follow the MVSA cut dimensions; cell counts are illustrative at $\beta = 0.2$, where both policies transmit 5 of 25 units and differ only in the split.

so they exchange the same expected payload and differ only in its division. The payload share $q_m = D_m r_m / \mathrm{TX}$ (with $\sum_m q_m = 1$) lies on the probability simplex:

$$q_m = \frac{D_m r_m}{\mathrm{TX}}, \qquad q_m^{\mathrm{uni}} = \frac{D_m}{\sum_k D_k} \ \ (\text{for } r_m = \beta). \tag{6}$$

Equal compression severity therefore does not imply equal transmitted floats: a modality with four times the smashed-activation dimension receives four times the payload. Our objective is to choose $\mathbf{q}$ from fusion-level contribution while preserving equation 5.

### 3.4 ModalShare: Contribution-Aware Allocation

We now introduce *ModalShare*, a contribution-aware allocator that divides a fixed budget into per-modality keep-ratios based on a server-side contribution score, leaving the intra-modality compressor unchanged (Figure 2).

**Server-side contribution probe.** We define the coalition utility $u(S)$ as the mean true-label log-probability produced by the server when only modalities $S \subseteq \mathcal{M}$ are present, with absent modalities replaced by zeros, a masking convention used also in modality valuation (Wei et al., 2024). Coalition marginals are aggregated by the Shapley value (Shapley et al., 1953),

$$\phi_m = \sum_{S \subseteq \mathcal{M} \setminus \{m\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[ u(S \cup \{m\}) - u(S) \right], \tag{7}$$

evaluated exactly with $2^M$ coalitions, or by sampling coalition orderings for large $M$. For example, when $M = 2$ the four coalitions are $\emptyset, \{1\}, \{2\}, \{1, 2\}$.

Because batch-level utilities are noisy, negative contributions are floored, renormalized to a share, and then smoothed,

$$\tilde{\phi}_m^{(t)} = \frac{[\phi_m^{(t)}]_+}{\sum_k [\phi_k^{(t)}]_+ + \varepsilon}, \qquad s_m^{(t)} = \alpha \tilde{\phi}_m^{(t)} + (1 - \alpha) s_m^{(t-1)}, \tag{8}$$

with the probe evaluated every $T_{\mathrm{probe}}$ batches and cached scores used in between.

**From contribution to keep-ratios.** Given scores $\{s_m\}$, temperature-controlled weights

$$w_m = s_m^{1/\tau}, \qquad \pi_m = \frac{w_m}{\sum_k w_k}, \tag{9}$$

convert contribution into a share of the residual payload after a keep-ratio floor $r_{\min}$ reserves $D_m r_{\min}$ units per modality. With $\mathrm{TX}_{\mathrm{sur}} = \mathrm{TX} - r_{\min} \sum_k D_k > 0$,

$$c_m = D_m r_{\min} + \pi_m \mathrm{TX}_{\mathrm{sur}}, \qquad r_m = \frac{c_m}{D_m}, \tag{10}$$

followed by projection to $r_m \in [r_{\min}, 1]$ that preserves the total budget equation 5. Smaller $\tau$ sharpens equation 9: hard modality selection is its $\tau \to 0$ limit, and the floor $r_{\min}$ and temperature $\tau$ interpolate between that corner and uniform allocation.

**Confidence and freeze.** Live allocation follows equation 10 without gating keep-ratios on the contribution gap: while unfrozen, ModalShare remains adaptive. Confidence is measured separately by the normalized gap between the leading and runner-up scores,

$$g = \frac{s_{(1)} - s_{(2)}}{\sum_k s_k + \varepsilon}, \tag{11}$$

where $s_{(1)} \geq s_{(2)}$ are the top two values in $\{s_m\}$. For $M$=2 this reduces to $g = |s_1 - s_2|/(s_1 + s_2 + \varepsilon)$. After epoch $e_{\min}$, if $g \geq \gamma$ favors the same winning modality for $K$ consecutive epochs, ModalShare freezes the current keep-ratios $\{r_m\}$, preserves equation 5, and discontinues probing. A drop of $g$ below $\gamma$, or a change of winner, resets the streak. The threshold $\gamma$ therefore controls only when to stop adapting, not the online map itself.

### 3.5 Training procedure and communication cost

The full training loop, embedding the probe, map, and freeze rule, is given in Appendix L. During early training the uplink budget is ramped from full keep-ratios ($r_m$=1) to the target $\beta$ (schedules in Section 4); while unfrozen, ModalShare refreshes $\{r_m\}$ via equation 10 on the probe schedule, including during the ramp. At fixed $\beta$, compared policies share the same expected TX and differ only in $\mathbf{q}$.

Crucially, the probe adds no client–server messages: coalitions are formed by masking smashed activations already held at the server, so attribution consumes no part of equation 4 and needs no auxiliary channel. Its only cost is $2^M$ forward evaluations of the server submodel (four for $M$=2) every $T_{\text{probe}}$ batches while unfrozen; freezing removes even that for the rest of training once the confidence criterion is met (Appendix E).

## 4 Experimental Setup

Our evaluation tests an allocation hypothesis rather than a new feature compressor. At a fixed communication budget $\beta$ equation 5, we ask whether replacing modality-agnostic keep-ratios with contribution-aware keep-ratios improves accuracy when the expected transmitted payload is held constant. Every controlled comparison holds fixed the intra-modality compressor $\mathcal{C}$ and the training protocol, and changes only the mapping from budget to per-modality keep-ratios $\{r_m\}$. Budget is accounted in expected kept smashed-activation floats, $\text{TX} = \sum_m D_m r_m$, together with the matched cut-gradient support (Section 3.2).

### 4.1 Baselines

To the best of our knowledge, no prior method allocates a shared payload budget across modalities, whether in SL or otherwise: existing communication-efficient methods compress each transmitted tensor independently and leave the division of a shared budget among modalities unspecified. Our central comparison is therefore between ***ModalShare*** and the modality-agnostic default it replaces, ***uniform*** allocation, which assigns every modality the same keep-ratio ($r_m = \beta$ for all $m$) and represents the existing practice of compressing each modality independently. Both satisfy the iso-budget constraint equation 5 and differ only in how the shared payload is split.

### 4.2 Datasets

**CREMA-D** (Cao et al., 2014) is an audio-visual emotion recognition dataset of 7,442 clips (synchronized speech and facial expression) over six emotions, split into 6,698 training and 744 test samples. **MVSA-Single** (Niu et al., 2016) is an image–text sentiment dataset; we use a stratified subsample of 2,592 pairs split 60/20/20 into 1,555 training, 518 validation, and 519 held-out pairs, with the validation partition reported as test accuracy. **UCI HAR** (Anguita et al., 2013) is a six-class human activity recognition dataset from smartphone inertial sensors (30 volunteers), 10,299 samples split into 7,352 training and 2,947 test. Each sample provides two synchronized streams, accelerometer and gyroscope.

Table 1: Flattened cut dimensions $D_m$ at the early split and the payload share $q_m^{\text{uni}} = D_m / \sum_k D_k$ induced by equal keep-ratios.

| | CREMA-D | | MVSA | | UCI HAR | |
|---|---|---|---|---|---|---|
| | video | audio | image | text | accel | gyro |
| $D_m$ | 200,704 | 195,520 | 200,704 | 49,152 | 32 | 32 |
| $q_m^{\text{uni}}$ | 50.7% | 49.3% | 80.3% | 19.7% | 50.0% | 50.0% |

### 4.3 Models

**Encoders.** Each modality has a client-side encoder truncated at a fixed early cut, with the server completing the remaining modality-specific layers, fusion, and classification (Section 3.1). Visual streams (CREMA-D video, MVSA image) use ResNet-18; MVSA text uses DistilBERT fine-tuned end-to-end; CREMA-D audio is a log-magnitude STFT spectrogram (22,050 Hz, $n_{\text{fft}}$=512, hop 353, 257×188) fed to a ResNet-18 trained from scratch; UCI HAR uses a small 1D CNN per stream.

**Fusion.** Fusion is concatenation followed by a linear classifier head, with no cross-modal attention. Unlike attention-based fusion, it does not reweight streams after the cut, so an accuracy difference between allocations at matched $\beta$ is attributable to the allocation itself.

**Splitting point.** In realistic SL the client cannot host most of the model: it transmits activations from an early cut, which is also where the payload is large enough for an uplink budget to bind and per-modality keep-ratios to matter. We use the early split (layer-1) by default: the client holds only the first stage of each encoder (ResNet-18 stem plus first residual block for CREMA-D video/audio and MVSA image; DistilBERT layers 0–1 for MVSA text; the first 1D-CNN block for UCI HAR), and the server runs the rest. A late split keeps the full encoder on the client and exchanges only pooled embeddings, leaving little payload to allocate. Table 1 lists the resulting cut dimensions: CREMA-D is nearly balanced, while MVSA's image stream already takes four-fifths of the payload under the modality-agnostic default.

### 4.4 Hyperparameters and evaluation

CREMA-D is trained with SGD (learning rate $10^{-3}$, weight decay $10^{-4}$, momentum 0.9) for 100 epochs with a 5-epoch learning-rate warmup and a 5-epoch compression ramp. MVSA is trained with AdamW (learning rate $3\times10^{-4}$, weight decay $10^{-3}$) for 30 epochs with one-epoch warmups. Batch size is 64 in both cases. Allocation uses floor $r_{\min}$=0, probe period $T_{\text{probe}}$=10, EMA coefficient $\alpha$=0.3, and temperature $\tau$=1, each shared across all datasets, budgets, and compressors. Freezing uses confidence threshold $\gamma$=0.90, streak length $K$=3, and earliest freeze epoch $e_{\min}$=10, likewise shared. For the underlying compressor we use SplitFC (Oh et al., 2025), unless otherwise specified. Final test accuracy is averaged over $N$=12 seeds.

**Experimental environment.** We implemented our framework and all baselines using PyTorch (Paszke et al., 2019). All experiments were conducted on a server equipped with a 32-core AMD Ryzen Threadripper PRO CPU, 504 GB of memory, and one NVIDIA RTX 4090 GPU.

## 5 Results

### 5.1 Same budget, better accuracy

Table 2a reports accuracy against uplink compression $1/\beta$ under SplitFC with the early split at matched total payload. ModalShare strongly outperforms uniform on CREMA-D and MVSA at every compression level. At 5× it lifts CREMA-D from 32.2 to 47.6 % and MVSA from 48.0 to 60.4 % at matched payload, recovering most of the gap to the no-compression references (53.1 and 61.8) that the uniform arm forfeits to compression. UCI HAR is milder but still favours the allocator at 5/6 budgets, settling at a slight imbalance rather than a corner – its two sensor streams contribute roughly equally, leaving little asymmetry to exploit.

Plotted against transmitted payload rather than a keep-ratio label (Figure 3a), the same pattern appears: Δ peaks near moderate compression (about +15 pp on CREMA-D and +12 pp on MVSA) and declines toward small budgets, approaching zero at 40×, where both arms hit the trivial floor, CREMA-D just above six-class chance (16.7), MVSA uniform on the majority-class rate (≈40), leaving allocation little to differentiate. Reallocation pays when the budget binds but still carries signal.

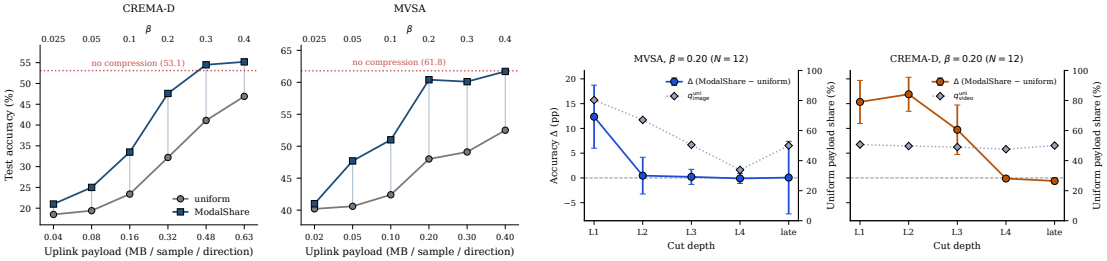
Table 2: ModalShare against uniform under SplitFC, early-split multimodal SL. (a) Accuracy (%) across compression rates, with the accuracy difference $\Delta$; the shaded $\Delta$ row is bold where the gain exceeds 1pp. (b) At $\beta$=0.2, the modality the probe favors and the payload share $q$ it receives ($q^{\mathrm{mod}}$: ModalShare, $q^{\mathrm{uni}}$: uniform), with the probe's contribution gap $g$. The gain tracks the measured gap: large on CREMA-D and MVSA, near-zero on the balanced UCI HAR.

| Dataset (no comp.) | Allocation | Compression rate | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2.5× | 3.3× | 5× | 10× | 20× | 40× |
| CREMA-D (53.1) | Uniform | 46.9±6.2 | 41.1±6.4 | 32.2±5.2 | 23.4±2.5 | 19.4±1.6 | 18.5±1.6 |
| | ModalShare | 55.2±2.1 | 54.5±3.5 | 47.6±5.9 | 33.5±6.5 | 25.0±4.8 | 21.0±4.4 |
| | $\Delta$ | **+8.3**±4.9 | **+13.4**±4.1 | **+15.4**±4.3 | **+10.1**±5.1 | **+5.6**±5.1 | **+2.5**±4.9 |
| MVSA (61.8) | Uniform | 52.5±2.5 | 49.1±3.8 | 48.0±2.1 | 42.4±3.3 | 40.6±2.5 | 40.2±2.6 |
| | ModalShare | 61.7±2.8 | 60.1±4.6 | 60.4±6.3 | 51.0±7.4 | 47.7±4.7 | 41.0±3.3 |
| | $\Delta$ | **+9.2**±2.5 | **+11.0**±5.4 | **+12.4**±6.4 | **+8.6**±8.1 | **+7.1**±5.5 | +0.8±4.5 |
| UCI HAR (63.7) | Uniform | 51.2±2.4 | 47.3±2.6 | 43.6±2.3 | 39.4±2.1 | 39.2±2.3 | 37.3±2.2 |
| | ModalShare | 51.5±2.2 | 48.0±2.5 | 44.7±3.5 | 40.2±3.1 | 39.1±2.1 | 37.5±2.1 |
| | $\Delta$ | +0.3±0.7 | +0.7±1.0 | **+1.1**±1.3 | +0.8±1.7 | −0.1±1.6 | +0.2±0.6 |

(a) Accuracy versus uplink compression (no-compression in parentheses).

| Dataset | Modalities | $q^{\mathrm{mod}}$ | $q^{\mathrm{uni}}$ | Gap $g$ | $\Delta$ |
|---|---|---|---|---|---|
| CREMA-D | audio | 0.99 | 0.49 | 1.00 | +15.4 |
| | video | 0.01 | 0.51 | | |
| MVSA | text | 0.88 | 0.20 | 0.90 | +12.4 |
| | image | 0.12 | 0.80 | | |
| UCI HAR | accel | 0.56 | 0.50 | 0.12 | +1.1 |
| | gyro | 0.44 | 0.50 | | |

(b) Probe contribution versus realised allocation.




(a) Accuracy vs. uplink payload (physical cost).

(b) Split-depth ablation at $\beta$=0.20.

Figure 3: **(a)** Accuracy (%) versus uplink payload; red dotted is the no-compression reference, top axis is budget $\beta$; the advantage peaks at moderate compression and closes toward the floor. **(b)** $\Delta$ = ModalShare − uniform versus cut depth (MVSA, CREMA-D); grey diamonds (right axis) are the uniform payload share, and $\Delta$ is largest at the early cut where the payload is large.

### 5.2 Does the probe measure contribution?

Does *ModalShare* help because it measures contribution, or would any aggressive asymmetric split have helped? Table 2b places the probe's measurement beside the allocation it produces, summarising each dataset by the contribution gap $g$ of equation 11, a property of the scores before they are mapped to $\{r_m\}$. The allocation follows it: on CREMA-D and MVSA the probe reads a large gap ($g$≈1.0 and 0.90) and concentrates the budget on the dominant modality (audio and text) for double-digit $\Delta$, driving the audio share to 0.99 on CREMA-D and the text share to 0.88 on MVSA, against the 0.49 and 0.20 that equal keep-ratios give them; on balanced UCI HAR it reads a small gap ($g$≈0.12) and splits near-evenly. UCI HAR is the discriminating case: a method forcing asymmetry regardless of the data would commit there too and lose accuracy, whereas ModalShare stays near uniform, so the allocation is a function of measured contribution, not a fixed asymmetric prior, and the gain follows a property the probe reads in the data rather than any generic reward for deviating from uniform.

### 5.3 Contribution versus dimension skew

On MVSA the image cut is roughly four times the text cut (Table 1), so equal keep-ratios already place 80.3% of the payload on image, and the MVSA gain could reflect skew correction rather than contribution. To separate the two we hold the budget fixed and split it three ways: uniform (proportional to cut dimension), an equal-payload control (half the floats each, isolating skew), and ModalShare (by contribution). Table 3 reports $\beta$=0.20. Moving from uniform to equal payload recovers 5.17 pp, the skew term; moving from equal payload to ModalShare recovers a further 7.23 pp, the contribution term, for 12.40 pp total. The larger part is contribution, which no dimension-based rule can reach. The CREMA-D result confirms this from the other side (Table 2b): its cuts differ by 1.4 pp in payload share, so skew is absent by construction, yet its contribution gain is the larger of the two datasets. The skew term thus appears only under unequal dimensions, while the contribution term is present in both.

Table 3: MVSA static allocation references at $\beta$=0.20. $q_{\text{image}}/q_{\text{text}}$ is the image/text float share. $\Delta$ is relative to uniform. Equal payload (per modality) isolates dimension skew; the further gap to ModalShare is the contribution effect.

| Allocation | $q_{\text{image}}/q_{\text{text}}$ | Acc. (%) | $\Delta$ (pp) |
|---|---|---|---|
| Uniform | 0.803/0.197 | 48.00$\pm$2.1 | — |
| Equal payload | 0.500/0.500 | 53.17$\pm$3.9 | +5.17 |
| ModalShare | 0.122/0.878 | 60.40$\pm$6.3 | +12.40 |

Table 4: Proxy screen at $\beta$=0.20. For each dataset, $\Delta$ is versus equal keep-ratios (pp) and $q$ is the payload share of the superscripted modality; uniform row ($\Delta$=0).

| | CREMA-D | | MVSA | | UCI HAR | |
|---|---|---|---|---|---|---|
| Allocator | $\Delta$ | $q^{\text{video}}$ | $\Delta$ | $q^{\text{image}}$ | $\Delta$ | $q^{\text{accel}}$ |
| Shapley (ModalShare) | $+\mathbf{15.4}$ | 0.01 | $+\mathbf{12.4}$ | 0.12 | +1.1 | 0.56 |
| Activation-norm | +4.3 | 0.31 | −1.5 | 0.55 | +0.1 | 0.50 |
| Gradient-norm | −0.3 | 0.58 | −1.3 | 0.52 | 0.0 | 0.51 |
| Solo utility | +0.3 | 0.50 | −1.7 | 0.52 | +0.1 | 0.50 |
| Uniform | 0.0 | 0.51 | 0.0 | 0.80 | 0.0 | 0.50 |

Table 5: Effect of ModalShare across intra-modality compressors: accuracy difference $\Delta$ = ModalShare − uniform (pp) at matched payload.

| | CREMA-D | | | | MVSA | | | | UCI HAR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Compressor | 5× | 10× | 20× | 40× | 5× | 10× | 20× | 40× | 5× | 10× | 20× | 40× |
| SplitFC | +15.4 | +10.1 | +5.6 | +2.5 | +12.4 | +8.6 | +7.1 | +0.8 | +1.1 | +0.8 | −0.1 | +0.2 |
| Top-$S$ | −5.11 | +7.71 | +7.67 | +4.97 | +6.52 | +13.21 | +9.91 | +7.26 | +0.08 | +1.32 | −0.21 | −0.04 |
| RandTop-$S$ | −2.51 | +4.99 | +9.65 | +8.18 | +4.21 | +11.50 | +10.05 | +8.93 | +0.77 | −0.40 | −0.05 | +0.03 |

### 5.4 Contribution Proxies

The probe costs a forward pass, so a natural question is whether a cheaper cut-local signal can drive the same map. Table 4 screens three at $\beta$=0.20: activation norm, cut-gradient norm, and solo utility $u(\{m\})$. None reaches the helpful allocation. On CREMA-D the best, activation norm, shifts only part way toward audio ($q_{\text{video}}$=0.31 vs. ModalShare's 0.01) for +4.3 against +15.4 pp; gradient norm drifts the wrong way and solo utility barely moves. On MVSA the scores hardly separate, so the allocation stays near equal payload and at or below uniform; on UCI HAR every arm is near-null. The reason is structural: activation and gradient norms are pre-fusion and cannot see how the modalities combine, while solo utility scores each modality in isolation, where one strong alone may be redundant and one weak alone complementary. Only the coalition difference in equation 7 measures marginal value, which is why the full probe recovers a gain every cheaper signal misses.

### 5.5 Robustness across compressors, split depth, and probe cost

The effect is not tied to one compressor. Under Top-$S$ and RandTop-$S$ as well as SplitFC, ModalShare beats matched uniform in 29/36 configurations across three compressors, three datasets, and four budgets (Table 5); the exceptions are within 1 pp on balanced UCI HAR and two negatives on CREMA-D at the mildest compression, where hard selection leaves little support to reallocate. The same table shows existing SL compressors sacrifice multimodal accuracy that ModalShare recovers. The gain is largest at the early cut and vanishes server-ward as the payload shrinks (Figure 3b). Confidence-freezing matches the always-online diagnostic while paying the probe only until the split is decided (Appendix E).

## 6 Discussion

**Scope of the formulation.** The formulation is defined for arbitrary $M$, but our experiments cover $M$=2, where the effect is cleanest to isolate, and evidence is strongest at early splits, where the cut carries enough payload for allocation to matter (Appendix D). The same link-local design places ModalShare under multi-client SL and split federated learning (Thapa et al., 2022; Hafi et al., 2024); cross-client scoring, clients with different modality subsets, non-stationary rankings (Appendix F), and scaling to $M\geq$3 remain outside this paper.

## 7 Conclusion

Multimodal SL inherits unimodal compression, whose only decision is how a single tensor is represented. A second decision goes unmade: when several streams share one uplink, equal keep-ratios divide the budget by cut dimension, a quantity unrelated to what each modality contributes. Casting that division as an iso-budget problem makes it a tunable degree of freedom, and *ModalShare* sets it from a server-side contribution score, with double-digit gains at matched payload and near-uniform allocation where the gap is small. Under a fixed communication cost, how a system divides its budget is as consequential as how tightly it compresses.

### Reproducibility statement

All information required to reproduce our results is contained in the main text and appendix. The multimodal split-learning setup, per-modality encoders, cut points, and fusion head are specified in Section 4.3; datasets and their splits in Section 4.2; and every optimizer, schedule, and allocation hyperparameter, all shared unchanged across datasets, budgets, and compressors in Section 4.4. The allocation rule is given in closed form in Section 3 (Equations 7–11 and Algorithm 1 in Appendix L), and the iso-budget accounting that makes every comparison payload-matched is defined in Section 3.2. All reported numbers are averaged over $N{=}12$ seeds with standard deviations. Source code will be released. We provide the source code under supplementary material.

### AI use statement

In preparing this work we used generative AI tools for minor language editing and LaTeX formatting assistance. We did not use generative AI tools for research ideation, experimental design, or the analysis of results, which were carried out by the authors. We have reviewed all AI-assisted text and take full responsibility for the final content of this work, including all claims, figures, and artifacts.

## References

Hadi Amini, Md Jueal Mia, Yasaman Saadati, Ahmed Imteaj, Seyedsina Nabavirazavi, Urmish Thakker, Md Zarif Hossain, Awal Ahmed Fime, and SS Iyengar. Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions. *arXiv preprint arXiv:2503.16585*, 2025.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3–4, 2013.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Yanglin Feng, Yang Qin, Dezhong Peng, Rui Wang, Xiaomin Song, and Peng Hu. Multimodal nested learning for decoupled and coordinated optimization. In *Forty-third International Conference on Machine Learning*, 2026.

Houda Hafi, Bouziane Brik, Pantelis A Frangoudis, Adlen Ksentini, and Miloud Bagaa. Split federated learning for 6g enabled-networks: Requirements, challenges, and future directions. *IEEe Access*, 12:9890–9930, 2024.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47): 1–39, 2024.

Cheng-Yen Hsieh, Yu-Chuan Chuang, and An-Yeu Wu. C3-sl: Circular convolution-based batch-wise compression for communication-efficient split learning. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2022.

Pengbo Hu, Xingyu Li, and Yi Zhou. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*, 2022.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4167–4177. IEEE, 2025.

Zehang Lin, Zheng Lin, Miao Yang, Jianhao Huang, Yuxin Zhang, Zihan Fang, Xia Du, Zhe Chen, Shunzhi Zhu, and Wei Ni. Sl-acc: A communication-efficient split learning framework with adaptive channel-wise compression. *IEEE Transactions on Vehicular Technology*, 2026.

Zheng Lin, Guanqiao Qu, Xianhao Chen, and Kaibin Huang. Split learning in 6g edge networks. *IEEE Wireless Communications*, 31(4):170–176, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Bas Meuwissen, Vasileios Tsouvalas, and Nirvana Meratnia. Autoencoder-compressed parallel split learning for pre-trained model fine-tuning. *arXiv preprint arXiv:2607.17913*, 2026.

Akrit Mudvari, Antero Vainio, Iason Ofeidis, Sasu Tarkoma, and Leandros Tassiulas. Adaptive compression-aware split learning and inference for enhanced network efficiency. *ACM Transactions on Internet Technology*, 24(4):1–26, 2024.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pp. 15–27. Springer, 2016.

Yongjeong Oh, Jaeho Lee, Christopher G Brinton, and Yo-Seb Jeon. Communication-efficient split learning via adaptive feature-wise compression. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10844–10858, 2025.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4032–4059, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

Jiawei Shao and Jun Zhang. Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6. IEEE, 2020.

Lloyd S Shapley et al. A value for n-person games. 1953.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. SplitFed: When federated learning meets split learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8485–8493, 2022.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27338–27347, 2024.

Zijing Xu, Yunfeng Kou, Kunming Wu, and Hong Liu. Contribution-guided asymmetric learning for robust multimodal fusion under imbalance and noise. *arXiv preprint arXiv:2510.26289*, 2025.

Binhang Yuan, Song Ge, and Wenhui Xing. A federated learning framework for healthcare iot devices. *arXiv preprint arXiv:2005.05083*, 2020.

Liangqi Yuan, Dong-Jun Han, Su Wang, Devesh Upadhyay, and Christopher G Brinton. Communication-efficient multimodal federated learning: Joint modality and client selection. *IEEE Transactions on Mobile Computing*, 2026.

Fei Zheng, Chaochao Chen, Lingjuan Lyu, and Binhui Yao. Reducing communication for split learning by randomized top-k sparsification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4665–4673, 2023.

# APPENDIX

## A CONTRIBUTION VERSUS DIMENSION SKEW

Section 5.3 already separates the MVSA $\beta$=0.20 gain into a skew term and a contribution term (Table 3), and notes that CREMA-D has almost equal cut dimensions, so there is no skew to undo. Two clauses not stated there: selecting a winning modality is the $\tau \to 0$ limit of the same map, not an alternative to it, and still requires knowing which stream wins; no cut-local signal in Table 4 recovers that direction, and on UCI HAR the map self-abstains at $g \approx 0.12$.

## B CONTRIBUTION PROXIES

Table 4 is reported in Section 5.4. None of the three cheaper scores reaches the helpful split.

## C ALLOCATION IS ORTHOGONAL TO THE COMPRESSOR

Section 5.5 already reports the $29/36$ win rate under Top-$S$, RandTop-$S$, and SplitFC (Table 5). Absolute numbers under hard sparsification are in Table 6.

Under Top-$S$ and RandTop-$S$ on MVSA the gains are often larger, and two Top-$S$ cells exceed the uncompressed reference. We keep SplitFC as the primary compressor because its uniform arm has the highest absolute accuracy in most budgets.

Table 6: Absolute accuracy (%) under hard column sparsification, $N$=12 seeds per cell, matched to the budgets of Table 2a. ModalShare is the confidence-freeze configuration used in the main results (same policy as Table 2a). Underline marks a cell above the per-dataset early-split no-compression reference used in Table 2a (CREMA-D 53.1, MVSA 61.8, UCI HAR 63.7). Comparisons are within a compressor and allocation pair; rows for different compressors are not controlled against one another.

| Compressor | Dataset | Allocation | 5× | 10× | 20× | 40× |
|---|---|---|---|---|---|---|
| Top-$S$ | CREMA-D | Uniform | 46.34 | 29.04 | 23.53 | 22.96 |
| | | ModalShare | 41.23 | 36.75 | 31.19 | 27.93 |
| | MVSA | Uniform | 57.43 | 50.64 | 48.39 | 45.01 |
| | | ModalShare | 63.95 | 63.85 | 58.30 | 52.27 |
| | UCI HAR | Uniform | 48.85 | 39.72 | 38.04 | 37.03 |
| | | ModalShare | 48.93 | 41.03 | 37.84 | 36.99 |
| RandTop-$S$ | CREMA-D | Uniform | 44.46 | 29.92 | 25.52 | 23.58 |
| | | ModalShare | 41.94 | 34.92 | 35.17 | 31.76 |
| | MVSA | Uniform | 56.05 | 50.76 | 46.96 | 42.63 |
| | | ModalShare | 60.26 | 62.26 | 57.01 | 51.56 |
| | UCI HAR | Uniform | 44.73 | 39.99 | 38.24 | 37.05 |
| | | ModalShare | 45.50 | 39.59 | 38.19 | 37.08 |

## D DEPENDENCE ON SPLIT DEPTH

Figure 3b (Section 5.5) already shows $\Delta$ at $\beta$=0.20 versus depth. Two facts not in that paragraph: on MVSA the uniform image share drops from $80.3\%$ to $33.8\%$ by layer 4 as the payload shrinks; CREMA-D stays dimension-balanced at every depth and still gains through layer 3, so that gain cannot be skew correction. Both datasets go to zero $\Delta$ at the late embedding cut, where almost no payload remains to divide; UCI HAR is flat throughout, consistent with its near-balanced probe. The operating regime is the early split SL is built for.

## E FREEZING VERSUS ONLINE

The probe costs $2^M$ server forwards every $T_{\text{probe}}$ batches while unfrozen (four for $M{=}2$), and no extra uplink. Table 7 compares uniform, always-online Shapley, and confidence freeze at $\beta{=}0.20$. Freeze matches or exceeds the online diagnostic on both datasets and still beats uniform by a wide margin, then stops probing.

CREMA-D separates early, so locking at epoch $e_{\min}{+}K$ is cheap. MVSA separates more slowly, which is why the rule waits for a sustained gap rather than a fixed clock. Score and keep-ratio trajectories are in Appendix I.

Table 7: Online versus early-freeze ModalShare at $\beta{=}0.20$. Accuracy (%) is the best post-warmup checkpoint (mean $\pm$ std). The last column is freeze minus always-online (paired), not freeze minus uniform. The freeze arm adapts for a short post-ramp window, then locks $\{r_m\}$ and stops probing.

| Dataset | Uniform | Online | Freeze | $\Delta_{\text{f}-\text{o}}$ |
|---|---|---|---|---|
| CREMA-D | 32.2±5.2 | 41.9±8.2 | 47.6±5.9 | +5.7±6.2 |
| MVSA | 48.0±2.1 | 56.9±6.4 | 60.4±6.3 | +3.5±6.9 |

## F NON-STATIONARY CONTRIBUTION

The freeze rule assumes the contribution ranking eventually stabilizes. On CREMA-D and MVSA the winner is consistent for a streak before the split locks. UCI HAR is stable and near-balanced, so $g$ never meets $\gamma$ and the split never locks; keep-ratios stay near $r_m = \beta$. Before a lock, a change of winner resets the streak; the always-online variant (Appendix E) never locks.

A ranking that oscillates faster than the probe and EMA in equation 8 can track is outside what we test. That regime would need a shorter EMA (larger $\alpha$) or a policy that never freezes.

## G FUSION-LEVEL CONTRIBUTION-SIGNAL VARIANTS

Table 8 swaps only the fusion-level score, keeping an online iso-budget map. Ungated Shapley is competitive on both datasets. The deployed method is confidence freeze on that ungated map (Appendix E), not the gated tanh row, which is an online score ablation.

The fixed-heavy row uses the winning direction known in advance. It is a known-direction control, not a bound, and can sit below the online probe (CREMA-D $+6.8$ versus ungated $+9.6$).

Table 8: Fusion-level contribution-signal variants at $\beta{=}0.20$ under SplitFC ($N{=}12$, seeds 50–61, paired; appendix). Each cell is $\Delta$ versus equal keep-ratios (percentage points; mean$\pm$std) with the win count against uniform. All rows keep an online (unfrozen) policy so that only the score changes. The deployed method is confidence freeze on the ungated Shapley map (Appendix E), not the gated tanh row. The fixed-heavy row uses the winning direction identified in advance (audio-heavy on CREMA-D; text-heavy on MVSA). It is a known-direction control, not a bound, frozen before scoring on these seeds, not a per-seed maximum over a share grid.

| Allocator | CREMA-D $\Delta$ | Wins | MVSA $\Delta$ | Wins |
|---|---|---|---|---|
| Equal keep-ratios (uniform) | — | — | — | — |
| Leave-one-out marginal | +10.3±4.6 | 12/12 | +6.7±6.7 | 10/12 |
| Ungated Shapley (online) | +9.6±5.0 | 12/12 | +8.9±6.9 | 11/12 |
| Share-normalised Shapley | +11.2±5.6 | 12/12 | +5.2±8.4 | 9/12 |
| Confidence-gated (tanh) | +10.7±4.8 | 12/12 | +9.5±6.1 | 12/12 |
| Fixed heavy (static) | +6.8±3.8 | 12/12 | +10.5±4.7 | 12/12 |

## H PER-SEED RESULTS AT $\beta$=0.20

Tables 9 and 10 expand Table 2a at $\beta$=0.20 (early SplitFC, confidence freeze, seeds 50–61). CREMA-D wins on $12/12$ seeds and locks audio-heavy ($q_{\text{audio}} \geq 0.98$) on every run.

MVSA wins on $11/12$. The miss is seed 50, which locks image-heavy ($q_{\text{text}}$=0.14) and loses 4.09 pp. That is a seed-level residual of a slower probe (Appendix I), not a reversal of the mean.

Table 9: Per-seed confidence-freeze ModalShare results at $\beta$=0.20 under SplitFC (CREMA-D, early split, seeds 50–61). Uniform is the seed-aligned equal-keep-ratio reference used in Table 2a. $q_{\text{audio}}$ is the realized audio payload share at the best post-warmup checkpoint. $\Delta$ is ModalShare minus uniform (percentage points). Mean±std and win count ($\Delta > 0$) are in the final rows.

| Seed | Uniform (%) | ModalShare (%) | $\Delta$ (pp) | $q_{\text{audio}}$ |
|---|---|---|---|---|
| 50 | 32.22 | 51.16 | +18.94 | 0.999 |
| 51 | 38.63 | 50.06 | +11.43 | 0.996 |
| 52 | 23.99 | 34.93 | +10.94 | 1.000 |
| 53 | 30.52 | 46.16 | +15.64 | 0.999 |
| 54 | 36.87 | 47.18 | +10.31 | 1.000 |
| 55 | 28.75 | 52.84 | +24.09 | 1.000 |
| 56 | 28.41 | 48.47 | +20.06 | 0.993 |
| 57 | 32.33 | 47.33 | +15.00 | 1.000 |
| 58 | 35.09 | 54.04 | +18.95 | 0.980 |
| 59 | 24.26 | 37.73 | +13.47 | 0.999 |
| 60 | 39.18 | 53.41 | +14.23 | 1.000 |
| 61 | 36.52 | 47.59 | +11.07 | 1.000 |
| Mean | 32.23 | 47.57 | +15.34 | 0.997 |
| Std | 5.21 | 5.90 | 4.35 | 0.006 |
| Wins ($\Delta > 0$) | | 12/12 | | |

Table 10: Per-seed confidence-freeze ModalShare results at $\beta$=0.20 under SplitFC (MVSA, early split, seeds 50–61). Uniform is the seed-aligned equal-keep-ratio reference used in Table 2a. $q_{\text{text}}$ is the realized text payload share at the best post-warmup checkpoint. $\Delta$ is ModalShare minus uniform (percentage points). Mean±std and win count ($\Delta > 0$) are in the final rows.

| Seed | Uniform (%) | ModalShare (%) | $\Delta$ (pp) | $q_{\text{text}}$ |
|---|---|---|---|---|
| 50 | 47.10 | 43.01 | -4.09 | 0.140 |
| 51 | 49.19 | 63.05 | +13.86 | 0.999 |
| 52 | 48.11 | 58.34 | +10.23 | 0.999 |
| 53 | 47.64 | 61.27 | +13.63 | 0.988 |
| 54 | 47.68 | 60.31 | +12.63 | 0.738 |
| 55 | 47.76 | 63.17 | +15.41 | 0.998 |
| 56 | 43.28 | 63.55 | +20.27 | 1.000 |
| 57 | 47.45 | 60.15 | +12.70 | 0.786 |
| 58 | 48.96 | 65.41 | +16.45 | 0.999 |
| 59 | 49.50 | 69.03 | +19.53 | 1.000 |
| 60 | 47.10 | 56.99 | +9.89 | 0.965 |
| 61 | 52.20 | 60.04 | +7.84 | 0.882 |
| Mean | 48.00 | 60.36 | +12.36 | 0.874 |
| Std | 2.06 | 6.35 | 6.37 | 0.249 |
| Wins ($\Delta > 0$) | | 11/12 | | |

# I CONTRIBUTION-SCORE DYNAMICS

Figures 4 and 5 show epoch-end EMA scores $\{s_m\}$ and the keep-ratios they produce at $\beta$=0.20 ($N$=12). Freeze epochs are reconstructed from stored per-epoch EMA scores using the rule in Section 3.4 ($\gamma$=0.90, $K$=3, $e_{\min}$=10), not from a logged freeze event. CREMA-D audio dominates by epoch 10 and every seed freezes at epoch 12 on an audio corner; the freeze-time winner matches the final winner on 12/12. UCI HAR stays near 0.56/0.44, never meets $\gamma$, and never freezes, so keep-ratios stay near $r_m = \beta$.

MVSA is slower: freeze on 11/12 seeds (median epoch 14), and the freeze-time winner matches the final winner on all eleven locks. Ten locks are text-heavy; one is image-heavy; one seed never meets $\gamma$. The accuracy miss in Table 10 is the image-heavy lock (seed 50), not the unfrozen seed and not oscillation after the lock.

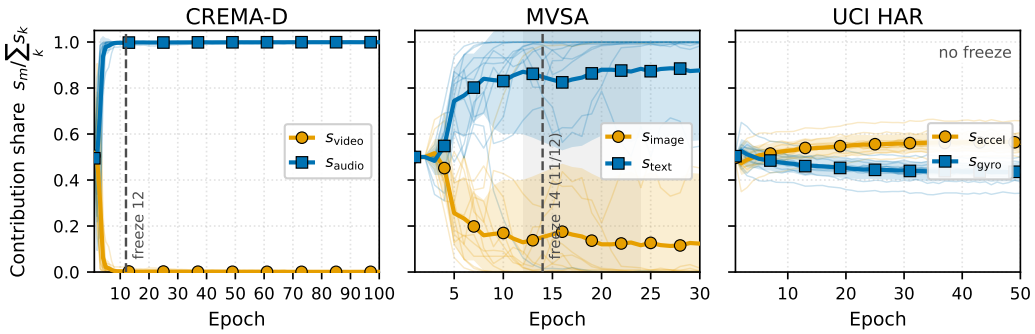



Figure 4: EMA Shapley contribution shares at $\beta$=0.20 ($N$=12). Thick lines: mean; thin lines: seeds; shaded: mean ± std. Vertical dashed: reconstructed freeze (median epoch among seeds that lock), not a logged freeze event. CREMA-D audio dominates by epoch 10 and freezes at epoch 12 on every seed; MVSA moves toward text more slowly (freeze on 11/12 seeds); UCI HAR stays near-balanced and never freezes.

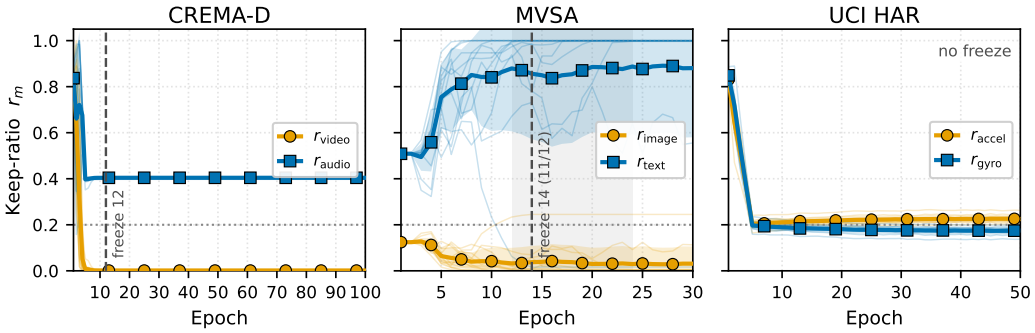



Figure 5: Keep-ratios $\{r_m\}$ corresponding to Figure 4. Dotted horizontal: equal keep-ratio $r_m = \beta$. Vertical dashed and grey band: reconstructed freeze as in Figure 4. CREMA-D locks an audio corner; MVSA locks a text corner on most seeds; UCI HAR remains near uniform because the freeze rule never fires.

# J GRADIENT BALANCING DOES NOT REPLACE ALLOCATION

A natural objection is that on-the-fly gradient modulation (OGM) (Peng et al., 2022) could close the uniform-to-asymmetric gap without changing keep-ratios. Table 11 tests OGM-GE at the smashed-data cut on CREMA-D at $\beta$=0.20. Equal keep-ratios stay equal; Shapley uses the live map equation 10.

OGM-GE on equal keep-ratios does not recover the allocation gap (+1.0 pp). The modulation coefficients stay at identity ($k_{\text{video}}\approx 1.00$, $k_{\text{audio}}\approx 0.998$): unimodal true-class softmax never exceeds the fused score, so Peng's versus-fusion rule never damps a stream. We do not combine Shapley with OGM: at identity coefficients the modulator is equalization noise on an already-allocated split, not a second allocator. Gradient modulation at the cut does not change the uplink split and does not replace contribution-aware keep-ratios.

Table 11: OGM-GE at the cut versus keep-ratio allocation on CREMA-D, early SplitFC, $\beta$=0.20. Accuracy (%) is the best post-warmup checkpoint. OGM-GE does not change $\{r_m\}$. Shapley is the live map without OGM.

| Keep-ratios | OGM-GE | Acc. (%) |
|---|---|---|
| Equal ($r_m$=$\beta$) | – | 35.6 |
| Equal ($r_m$=$\beta$) | yes | 36.6 |
| Shapley | – | 48.7 |

## K NOTATION

Table 12: Notation.

| Symbol | Meaning |
|---|---|
| $M$ | number of modalities |
| $D_m$ | flattened cut dimension of modality $m$ |
| $z_m$ | smashed activations of modality $m$ |
| $r_m$ | keep-ratio of modality $m$ (the allocated quantity) |
| $\beta$ | uplink budget as a fraction of full cut payload, $\text{TX}/\sum_k D_k$ |
| $q_m$ | payload share of modality $m$, $D_m r_m/\text{TX}$ |
| $\mathcal{C}$ | intra-modality compressor (SplitFC primary) |
| $u(S)$ | coalition utility (server true-label log-prob with only modalities $S$) |
| $\phi_m$ | Shapley contribution score of modality $m$ |
| $s_m$ | EMA-smoothed contribution share of modality $m$ |
| $\pi_m$ | normalized allocation weight from $s_m$ |
| $\tau$ | allocation temperature |
| $r_{\min}$ | keep-ratio floor |
| $g$ | contribution confidence gap |
| $\gamma$ | freeze confidence threshold |
| $K$ | consecutive confident epochs required to freeze |
| $e_{\min}$ | earliest epoch at which freezing is allowed |

## L TRAINING PROCEDURE

Algorithm 1 places the probe, keep-ratio map, and freeze rule in the training loop.

**Algorithm 1** ModalShare iso-budget allocation.

1. Compute smashed activations $z_m \leftarrow g_m(x^{(m)})$, $m \in \mathcal{M}$.
2. Every $T_{\text{probe}}$ batches while unfrozen: evaluate server coalitions, update $\{s_m\}$, and map to $\{r_m\}$ under equation 10.
3. If the freeze criterion on $g$ is met, hold $\{r_m\}$ fixed and skip further probes; otherwise continue adapting.
4. The client compresses and uploads $\hat{z}_m \leftarrow \mathcal{C}(z_m, r_m)$.
5. Complete server forward/backward, return cut gradients, and update $\{\theta_m\}$ and $\theta_s$.