Proposed Multi-Modal OrthoAI Experiments for OrthoAI Phase-1 (MedGemma-4B Backbone)

1. Objective

The goal of this experimental plan is to develop and rigorously evaluate a unified multi-modal OrthoVLM (OrthoAI) for orthodontic decision support, built on top of **MedGemma-4B**. The model should jointly reason over intraoral photographs, panoramic X-rays (OPG), and lateral cephalometric images to predict clinically relevant orthodontic findings (skeletal pattern, malocclusion class, crowding/spacing, etc.) and support explainable downstream tasks (VQA-style clinical reasoning, structured reports).

---

2. Base Model and Data

- **Backbone:** MedGemma-4B multimodal (MedSigLIP image encoder + Gemma LLM).
- **Initialization:** I start from our **existing MedGemma-4B checkpoint fine-tuned on panoramic orthodontic X-rays**, which encodes domain-specific dental radiology priors.
- **Phase-1 Dataset:** Curated multi-modal orthodontic cohort with:
  - Intraoral RGB views,
  - Panoramic OPG,
  - Lateral ceph,
  - Expert-curated diagnostic labels and textual descriptions.
- Data remains de-identified and used exclusively under the approved research framework.

---

3. Experimental Conditions

I will conduct a controlled set of experiments to isolate the value of fusion and validate our chosen design.

(E1) Single-Modality Baseline (OPG-only)

Use the pano-tuned MedGemma-4B to predict diagnostic labels from OPG alone.

Purpose: establishes a **strong radiology-only baseline** and quantifies the incremental value of additional views.

(E2) Naïve Late Fusion (Per-Modality Heads + Ensemble)

Train lightweight classifiers for:

- intraoral-only,
- OPG-only,
- ceph-only;

  then fuse predictions at decision level (e.g., averaging or learned logistic fusion).

  Purpose: baseline multi-modal strategy with **no shared representation**; tests whether simple ensembling is sufficient.

(E3) MedGemma Multi-Image Prompting (Unmodified Encoder)

Feed multiple images (intraoral + OPG + ceph) as separate inputs to MedGemma-4B with an instruction prompt (e.g., “Image 1 is intraoral, Image 2 is OPG…”).

Purpose: **token-level / attention-based fusion** without architecture changes; serves as a strong, implementation-aligned baseline.

(E4) Proposed Early Fusion with Adapted Vision Adapter (Primary Model)

Design a unified early-fusion pipeline aligned with Dr Yutong’s guidance:

1. Normalize all modalities to a shared resolution.
2. Construct a **stacked multi-channel tensor** (e.g., intraoral RGB + OPG + ceph → 5 channels).
3. **Adapt the MedGemma-4B vision input layer**:
   - Start from our pano-tuned checkpoint.

- Extend the first conv/patch-embedding to accept the stacked channels, seeding new channels from existing radiology weights and/or stable MedGemma filters.
- Freeze most of the encoder initially, fine-tuning selectively (plus a small multi-task classification / VQA head).

4. Train with supervised and VQA-style objectives to ensure:
    - accurate multi-label diagnosis,
    - consistent cross-view reasoning.

Purpose:

This model **implements principled early fusion**:

- a single shared encoder,
- modality-specific information preserved via channels,
- efficient parameter usage suitable for our dataset size,
- consistent with state-of-the-art multi-modal fusion practices.

---

4. Evaluation and Justification

Each configuration will be evaluated on:

- Multi-label diagnostic accuracy, F1, AUROC (per condition and overall),
- Calibration and consistency between modalities,
- Case-level ablations (with/without specific views),
- Qualitative VQA outputs for clinical plausibility.

Why this plan is reasonable:

1. **Leverages strong priors:** builds on MedGemma-4B and our pano-tuned checkpoint instead of training from scratch.
2. **Clinically aligned:** mimics how orthodontists jointly interpret intraoral photos, OPG, and ceph rather than treating them in isolation.
3. **Data-efficient:** early fusion with one encoder avoids the parameter explosion of three separate backbones.

4. **Scientifically honest:** includes clear baselines (single-modality, naive fusion, multi-image prompting) to show that any gains from the proposed fusion are **earned and measurable**, not hand-waved.