Memoire: Learning User Personas from Gallery Tags for Personalized Photo Curation

Supplementary Material

Praful Mathur, Mohsin Iftekhar, Aman Sharma, Sarvesh Tiwari, Meghali Deka,
Sathish Cherukuri, K Roopa Sheshadri, Rakesh Valusa
Samsung R&D Institute India - Bangalore
{p.mathur, m.iftekhar, aman.sharma, sarvesh.t, sathish.ch, roopa, v.rakesh }@samsung.com

1. Methodology

1.1. Synthetic Data Generation (Extended)

1.1.1. Virtual Users

Scale and splits. We simulate **100** virtual users. For each user we generate a gallery of **500 still images** and **20 short videos** (5 s each, 8 fps), totaling **50k stills** and **2k clips**. From each video we sample frames at **1 fps**, yielding an additional **100 frames per user** that expand event coverage. We partition users 80/10/10 for train/val/test to prevent identity leakage.

Identity anchors. Each user u is assigned (i) a *self* identity and (ii) a social graph with $8-12$ nodes sampled from {father, mother, spouse/partner, son, daughter, sibling, friend, colleague, grandparent, cousin}. For every person node p we first create **three high-quality anchor portraits** at 1024×1024 using Stable Diffusion 3[7] (rectified-flow transformer) with *InstantID*[9] identity adapter to preserve facial features across scenes . We then enroll these anchors by extracting ArcFace[5] embeddings to use later for identity checks and mask assignment. Physical feature attributes (age band, height, weight, body type, build) are sampled from simple priors and baked into the anchor prompts to ensure visual coherence across the gallery.

Social graph sampling. We instantiate a typed, undirected graph $G = (V, E)$ with $V = V_P \cup V_L \cup V_E$. We connect people–people edges with probability proportional to closeness (*self–parent* > *parent–child* > *friend* > *colleague*), and people–event/location edges when they co-occur in the same scenario template (Sec. 1.1.4). Edge weights mix co-occurrence and a small social prior; these priors are later reused by PERSONA-GAT during training (details in main text).

1.1.2. Event Ontology

We use **32** event types spanning personal, social, travel, and work contexts: *Birthday, Wedding, Anniversary, Graduation, Family Dinner, Vacation, Picnic, Hiking, Sports Match, Music Concert, Festival, Religious Ceremony, Baby Shower, Farewell Party, Office Meeting, Conference Talk, Award Ceremony, Team Outing, Reunion, Housewarming, New Year Celebration, Christmas, Halloween, Cultural Parade, Public Lecture, Museum Visit, Beach Day, Road Trip, Travel Landmark, Coffee Shop Gathering, Casual Home Scene, Park Stroll, City Night Walk*. Each user draws $8-14$ events with Dirichlet($\alpha = 0.4$) to promote sparsity (a few events dominate), and each event spawns $n_e \sim \mathrm{Uniform}\{20, 40\}$ stills plus one video.

1.1.3. Metadata Simulation

We synthesize the on-device style metadata PERSONA-GAT consumes (frequency, recency, view, favorite, relation, type). For image I and tag t:

- **Timestamp/recency.** We lay out a per-user calendar over 18 months. Event start times follow a non-homogeneous Poisson process (weekly/seasonal boosts); within an event, photos are bursty (log-normal inter-arrival). Recency feature uses $r_t = \exp(-\Delta_t/\tau_r)$ with τ_r=30 days.
- **Frequency.** $f_t = \#\{I : t \in I\}/|\mathcal{I}_u|$ z-scored per user.

- **Views.** $v_t \sim \mathcal{N}(\mu_t, \sigma_t)$ with μ_t larger for close-kin and milestone events (*e.g.* daughter's graduation $\mu=150$, $\sigma=40$) and lower for casual/work (*e.g.* office meeting $\mu=30$, $\sigma=10$); values are clamped to ≥ 0 and then normalized per user.
- **Favorites.** $q_t \sim \text{Bernoulli}(p_t)$ with p_t=0.7 for {self, spouse, children}, 0.45 for {parents, siblings}, 0.3 for {friends}, 0.2 for {colleagues}. For location/event tags, p_t increases to 0.6 for travel landmarks and ceremonies.
- **Relations/types.** We attach one-hot type (P/L/E) and relation label ({self, father, mother, son, daughter, sibling, friend, colleague, unknown}) to people tags; unknown is used when a face is present but not in the user's graph.

1.1.4. Photo-realistic Image Generation (SD3 + InstantID)

We render stills with Stable Diffusion 3 (rectified-flow) and InstantID [7, 9]:

Backbone. SD3-Medium (RFT) at 1024×1024; CFG scale 4.0; 35 inference steps (DPMSolver++).

Identity adapter. InstantID with 1–3 reference portraits per person. We set λ_{id}=0.8 (face preservation), λ_{pose}=0.2 (allow pose variance), identity LoRA rank 64. When multiple identities appear, we supply a tiled identity grid to the adapter and tag names in the prompt to disambiguate roles.

Conditioning. We compose text from (**event**, **location**, **people+relations**, **style**), then add negative prompts (artefacts, extra fingers, deformed face). We pass per-person crops to InstantID and keep composition control in text (camera, lens, lighting).

QC and retries. After generation we verify identity using ArcFace; any person–anchor cosine $< \tau_{\text{match}}$=0.35 triggers a targeted re-render with higher λ_{id} and/or tighter face crops. NSFW/blur filters and face-count checks prevent outliers.

Prompt templates (examples). Placeholders are in `[brackets]` and are programmatically filled:

```
A candid [event_name] in [location_phrase]:
[role_1]=[person_1], [role_2]=[person_2], ...
Camera: [35mm DSLR], framing: [mid-shot + shallow DOF],
light: [soft indoor evening], mood: [warm family tone].
Visual details: [balloons, cake, candles], no watermark, photorealistic.
STYLE: [Physical feature attributes (age band, height, weight, body type, build) ].
```

Example (Birthday):

> "A candid **birthday** at **home living room**: **Daughter** [Alice] holds a cake, **Father** [Raj] claps, **Mother** [Nita] takes a phone photo; 35mm DSLR, mid-shot, shallow DOF, soft warm light; balloons, candles; photorealistic, no watermark. The **father** is 50 years old, has a height of 5'8", weighs 80 kg, has an average body type, and a medium build. The **mother** is 45 years old, has a height of 5'5", weighs 65 kg, has an average body type, and a medium build. The **daughter** is 20 years old, has a height of 5'4", weighs 55 kg, has a slim body type, and a slender build. "

1.1.5. Video Generation (Hunyuan Video)

We generate 5 s clips at 640×360, 8 fps using Hunyuan Video [4] with multi-identity conditioning:

Inputs. (i) Text storyboard prompt; (ii) 1–3 reference portraits per identity (same anchors as above); (iii) optional composition hints (pan/tilt)

Identity control. We use an IP-Adapter or face-ID module [24] to inject per-identity embeddings; when two people co-star, we pack references as (Name:Portrait) pairs and assign "left/right/foreground" roles in text. We lock identity strength s_{id}=0.75 and allow motion/pose drift via s_{motion}=0.25.

Sampling. 30–40 denoising steps (rectified-flow schedule), classifier-free guidance 3.5, vram-efficient attention, xformers enabled.

Example (Graduation video):

> "Wide shot of a **graduation** in a university auditorium; **Son** [Arjun] in cap and gown walks across stage; **Mother** [Nita] smiles in audience right, **Father** [Raj] films from aisle; slow *left-to-right pan*; natural stage lighting; crowd murmurs faintly; maintain exact faces of Son/Mother/Father from reference portraits."

From each clip we sample frames at **1 fps** (uniform) and retain at most **5** frames per clip to curb redundancy.

As shown in Fig. 1 and Fig. 2, we visualize synthetic gallery generations for two representative users.

As shown in Fig. 3 and Fig. 4, the video generated for two virtual users in two different scenarios.

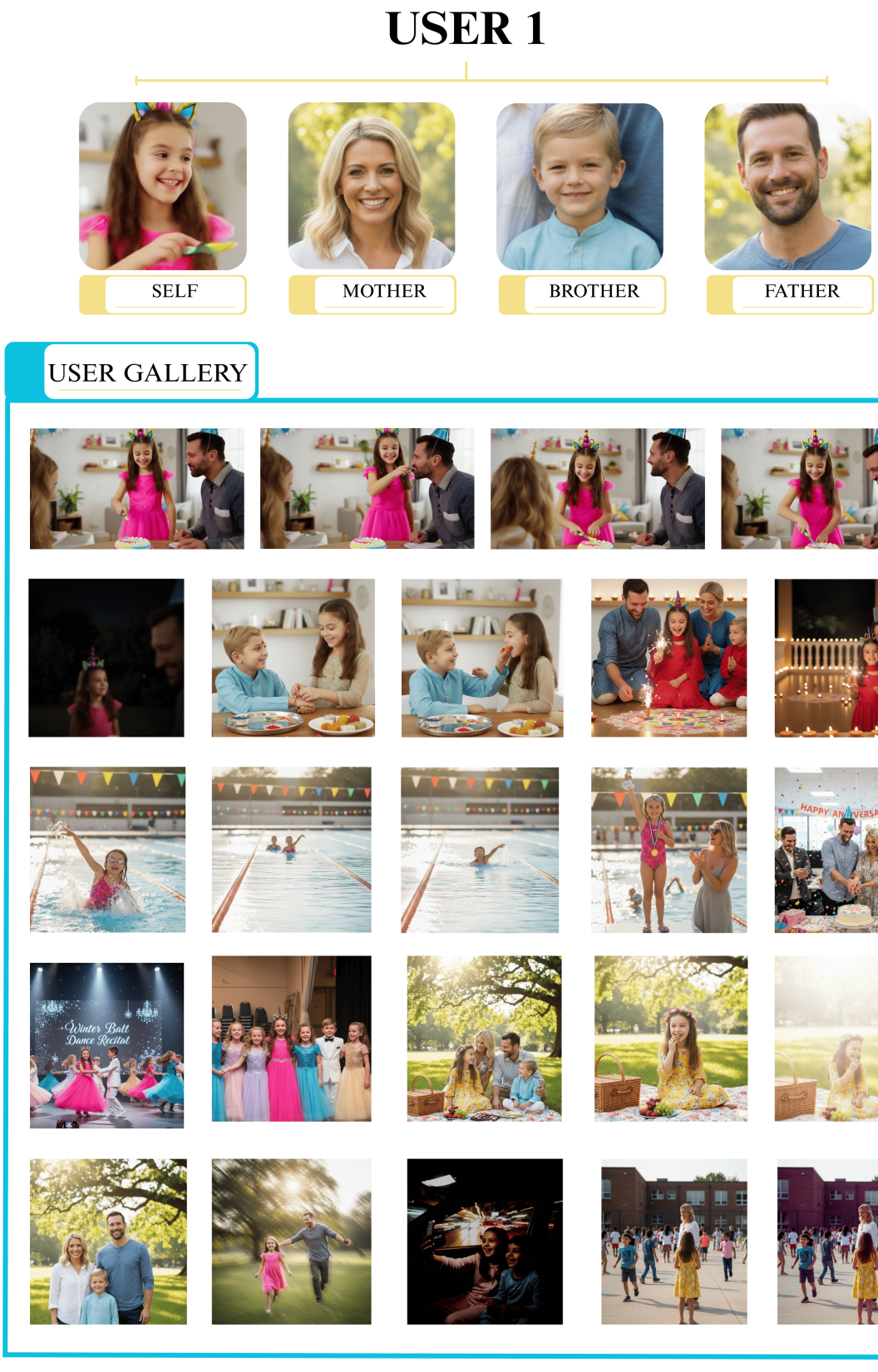

Figure 1. Synthetic gallery images for a virtual user .

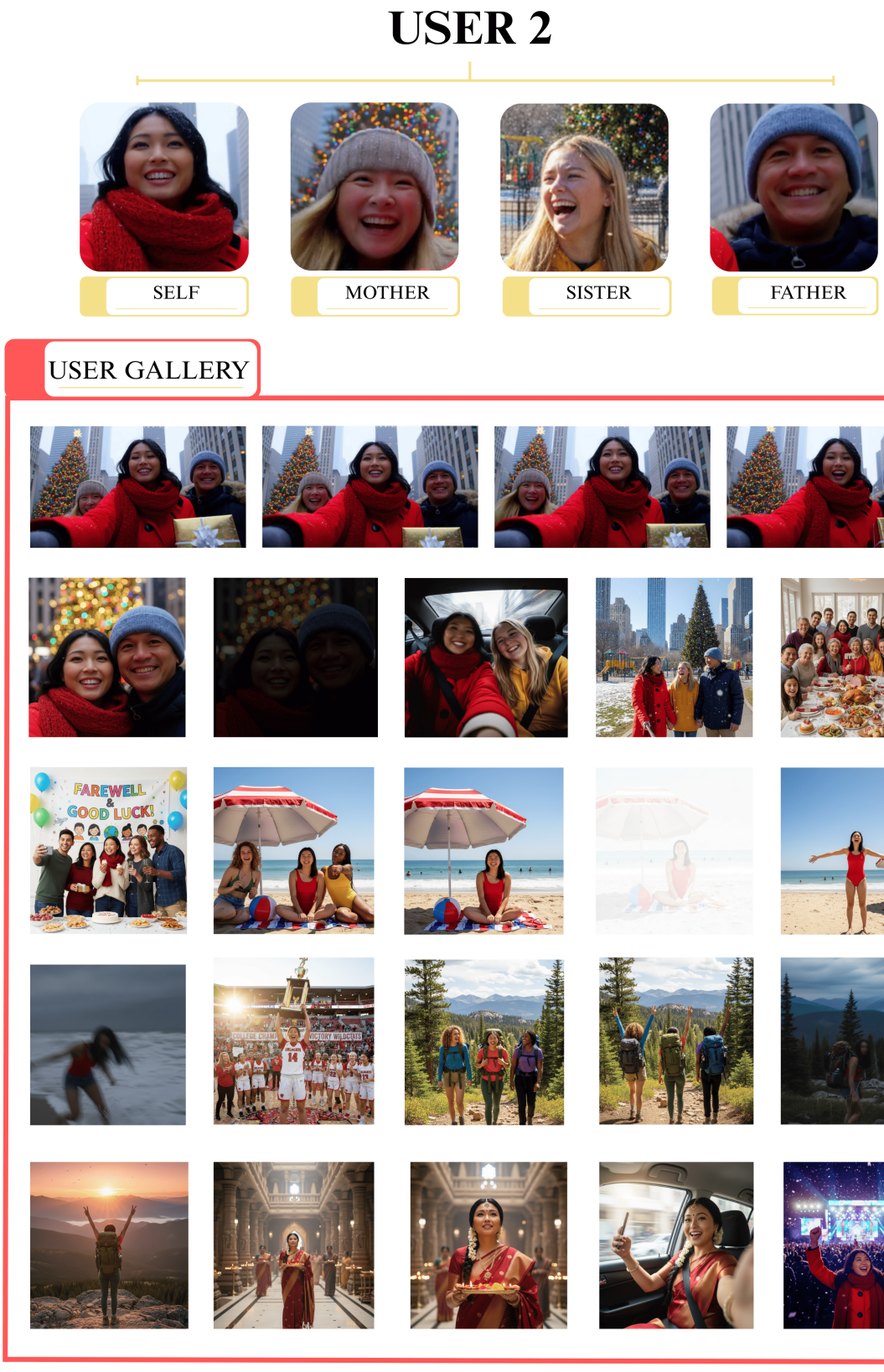

Figure 2. Synthetic gallery images for a virtual user.

Figure 3. Generated video sample for birthday scenario for a virtual user.

Figure 4. Generated video sample for christmas scenario for a virtual user.

1.1.6. Quality Control & Identity Consistency

We run post-hoc checks: (i) ArcFace verification for all depicted identities (τ_{match}=0.35), (ii) crowd sanity (no extra unknown faces in closeup scenes), (iii) exposure/blur heuristics, (iv) duplicates via perceptual hashing (Hamming distance < 6 triggers replacement). Fails are re-rendered with stronger identity weights and constrained prompts (*e.g.* tighter crops or reduced crowd size).

1.1.7. Prompts: Event{Location composition bank

We keep a bank of shot plans per event: *wide establishing*, *medium group*, *intimate portrait*, *detail close-up* (cake, ring, trophy), each with lighting variants (daylight, golden hour, tungsten indoor) and background clutter priors. Locations are drawn from a small gazetteer per user (home, office, school, city landmarks, park types) and injected as *[location_phrase]* in templates; if a landmark is chosen, we add visual anchors ("arc de triomphe stone texture", "red sandstone fort walls").

1.1.8. Exact implementation configuration

- **SD3 (images).** `resolution=1024`, `steps=35`, `sampler=DPMSolver++`, `cfg=4.0`, `seed=[event_seed]`, `instantid.lambda_id=0.8`, `instantid.lora_rank=64`, `vae=fp16`, `attn=xformers`.
- **Hunyuan (video).** `res=640x360`, `fps=8`, `len=5s`, `steps=36`, `cfg=3.5`, `id_strength=0.75`, `motion_strength=0.25`.
- **Identity checks.** ArcFace `r100` backbone; accept if cosine≥ 0.35 for each depicted identity; otherwise re-render.

1.1.9. Compute profile (proxy measurements)

We report representative, reproducible timings on *A100 80 GB* and *RTX 4090 24 GB* for reproducibility (batch size 1, after 20 warm-ups; median over 50 runs):

Task	A100 80 GB	RTX 4090 24 GB
SD3+InstantID (35 steps, 1024×1024) [7, 9]	3.8 s, 9–11 GB VRAM	5.6 s, 13–14 GB VRAM
Hunyuan 5 s clip (36 steps, 640×360@8 fps) [4]	4.5–6.5 min, 28–34 GB	7–10 min, 20–22 GB
ArcFace verify (per face crop) [5]	1.2 ms, $<$ 0.5 GB	1.5 ms, $<$ 0.5 GB
Qwen2.5-VL-7B-Instruct (per image+prompt, max_new_tokens=2) [2]	280–320 ms, 12–14 GB	420–480 ms, 14–15 GB

1.1.10. Personal Attention Aware Labeling with a VLM (Qwen-VL)

We obtain pairwise *personal impact* supervision using a vision–language model that conditions jointly on the raw image, the personal attention map $A(I)$ (PAT; see main), and the three tag sets (people+relations, events, locations).

Input packing. For each image I we create a panel of two modalities: (i) the RGB image, and (ii) Attention map $A(I)$; both are passed as separate images to the Qwen2.5 VLM [2]. We additionally pass a short textual card summarizing tags: *People:* `Daughter (Alice), Father (Raj), Mother (Nita)`; *Event:* `Birthday`; *Location:* `Home living room`.

VLM Prompt. "You are a careful assessor of personal impact in photo galleries. Your task is to evaluate two image candidates and select the one with greater personal significance to the user, considering both emotional value and visual presentation.

Evaluation Criteria:

1. **Primary Focus - Personal Impact:**
   - Weighted Highlight Analysis:
     - Pay close attention to the intensity/weighting of highlighted regions in the attention map.
     - Regions with stronger highlighting should be given proportionally greater importance.
   - People:
     - Clearly visible close relations (family, friends) in weighted highlight regions.
     - Preference for good posture and positive emotions (smiling, laughing).
     - Multiple people visible with good interactions score higher.
   - Locations:
     - Meaningful places clearly marked in weighted highlight regions.
     - Recognizable and personally significant venues score higher.
   - Important Events:
     - Prioritize images capturing Important events (weddings, birthdays, graduations, family gatherings) if they are emphasized in the attention overlay.
2. **Secondary Consideration - Aesthetics:**
   - Composition: Balanced framing and good use of space.
   - Lighting: Appropriate brightness and contrast.
   - Focus: Sharpness on main subjects.

Input Structure (Per Candidate):

- Raw image: The base photo.
- Attention map: Indicates regions of importance (people, locations, objects).
- Tag card: Provides contextual metadata (e.g., event type, people tags, location tags).

Output Requirement:

- Your response must be exactly one line: Image1 or Image2.
- No explanations or justifications are to be included.

"

As shown in Fig. 5, we visualize virtual users and paired images from their synthetic galleries marked as High and Low personal impact by the VLM.

Decoding and reliability. We use temperature $= 0$, top-p $= 1$, max_new_tokens $= 2$ to force a one-token label. We run *two prompt variants* (swap order; add a null rationale sentence) and accept a label only if both agree; otherwise the pair is discarded. For every user we cap pairs at 20k with balanced sampling across (same/different event, same/different key person, same/different location). We cache $\{$image, $A(I)$, tag card$\}$ triplets to avoid re-tokenization overhead.

Why this yields personalized labels. Because the VLM is forced to attend to $A(I)$ and the tag card, it scores *who/where/what* the user cares about (from PERSONA-GAT) rather than generic composition. Empirically this reduces label noise in non-face dominant scenes and improves agreement with human raters.

1.2. PERSONA-GAT with Session Adaptation

1.2.1. Sets, tags, relations, and metadata.

Gallery. $\mathcal{I} = \{I_1, \ldots, I_{|\mathcal{I}|}\}$ is a user's photo gallery. Each image I has tag sets $\mathcal{P}(I) \subseteq \mathcal{V}_P$ (people), $\mathcal{L}(I) \subseteq \mathcal{V}_L$ (locations), and $\mathcal{E}(I) \subseteq \mathcal{V}_E$ (events).

Tags. $\mathcal{V} = \mathcal{V}_P \cup \mathcal{V}_L \cup \mathcal{V}_E$ where each tag $t \in \mathcal{V}$ denotes a *person identity with a social relation* (e.g., `Father`, `Daughter`, `Friend`), a *location* (e.g., `Colosseum`), or an *event* (e.g., `Birthday`, `Graduation`). People tags are created at enrollment from a single reference face per identity.

People relations. For $t \in \mathcal{V}_P$, the relation label

$$\rho_t \in \{\text{self}, \text{father}, \text{mother}, \text{son}, \text{daughter}, \text{sibling}, \text{friend}, \text{colleague}, \text{unknown}\}$$

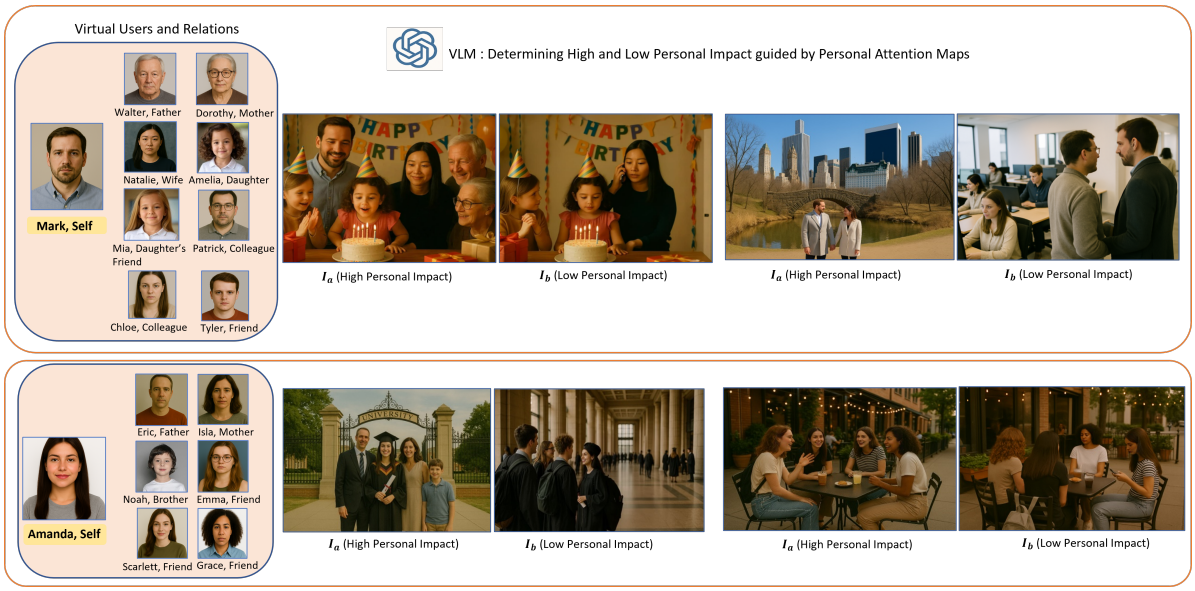

Figure 5. Depiction of various High and Low personal impact image pairs for two virtual users as predicted by VLM.

comes from on-device gallery metadata (unlabeled faces are *unknown*). Relations are used only to bias *person–person* edges; non-person edges ignore ρ.

1.2.2. Graph and edge weights.

We build a per-user, undirected, weighted graph $G = (\mathcal{V}, \mathcal{E})$ where a tag pair (u, v) is connected if it co-occurs in at least one image:

$$(u, v) \in \mathcal{E} \iff \exists I \in \mathcal{I} \text{ s.t. } u \in I,\ v \in I.$$

Define normalized co-occurrence (Jaccard-like):

$$\text{cooc}_{uv} = \frac{\#\{I : u, v \in I\}}{\max\big(1, \#\{I : u \in I\} + \#\{I : v \in I\} - \#\{I : u, v \in I\}\big)} \in [0, 1]. \tag{1}$$

Define *social boost* only for person–person edges:

$$\text{social}_{uv} = \begin{cases} \omega_{\text{kin}} & \text{close kin (self–parent, parent–child),} \\ \omega_{\text{friend}} & \text{friend/sibling,} \\ \omega_{\text{work}} & \text{colleagues,} \\ \omega_{\text{unk}} & \text{weak/unknown,} \\ 0 & \text{otherwise (non-person edges),} \end{cases} \qquad 1 \geq \omega_{\text{kin}} > \omega_{\text{friend}} > \omega_{\text{work}} > \omega_{\text{unk}} \geq 0. \tag{2}$$

The edge weight mixes co-occurrence and social prior:

$$w_{uv} = \lambda_c \, \text{cooc}_{uv} + \lambda_s \, \text{social}_{uv}, \qquad \lambda_c, \lambda_s \geq 0,\ \lambda_c + \lambda_s = 1. \tag{3}$$

We choose (λ_c, λ_s) per user based on the *share of person–person edges* $p_{PP} = \frac{\#\{(u,v) \in \mathcal{E} : u, v \in \mathcal{V}_P\}}{|\mathcal{E}|}$:

$$\lambda_s = \sigma\big(\kappa\,(p_{PP} - \mu)\big), \qquad \lambda_c = 1 - \lambda_s, \tag{4}$$

where σ is sigmoid, $\mu \in (0, 1)$ a midpoint (default 0.5), and $\kappa > 0$ a slope (default 5). We add self-loops ($w_{vv} = 1$) and optionally sparsify with k-NN per node (default k=20) or a threshold τ_e on w_{uv} to keep the graph compact (≤ 500 nodes after pruning).

Worked example (graph semantics). Suppose I_1 has {`Father`, `Daughter`, `Birthday`, `Home`} and I_2 has {`Father`, `Mother`, `Birthday`, `Restaurant`}. Then edges `Father`–`Daughter` and `Father`–`Mother` receive large w_{uv} (both co-occurrence and ω_{kin}), `Birthday` links to each person and location via co-occurrence only, and `Home`–`Restaurant` is *not* connected (never co-occur). Intuitively, G encodes that the user's gallery ties *close family* strongly to the *Birthday* event.

1.2.3. Tag features and normalization.

For each tag t, we assemble a per-user, z-scored feature vector

$$\mathbf{x}_t = [f_t,\ r_t,\ v_t,\ q_t,\ g_t,\ \mathbf{u}_t], \tag{5}$$

where $f_t = \#\{I : t \in I\}/|\mathcal{I}|$ (frequency), $r_t = \exp(-\Delta t_t/\tau_r)$ (recency; Δt_t is time since last appearance, τ_r a time constant), v_t (view rate), q_t (favorite ratio), $g_t = \frac{\sum_{u\in\mathcal{N}(t)} w_{tu}}{\sum_{(i,j)\in\mathcal{E}} w_{ij}}$ (graph relations), and $\mathbf{u}_t$ is one-hot tag type (P/L/E) concatenated with a one-hot of ρ_t for people. All scalar features are z-scored across $\mathcal{V}$ per user.

1.2.4. Data-adaptive prior (Importance Potential).

We compute regime statistics: *Burstiness* B = fraction of gallery within last τ_b days; *Social density* $D = p_{PP}$; *Longevity* L = fraction older than τ_ℓ. Convert to weights

$$[\lambda_r, \lambda_{rel}, \lambda_f] = \text{softmax}([\beta B,\ \gamma D,\ \alpha L]), \tag{6}$$

with $\alpha, \beta, \gamma > 0$ (defaults: α=1.0, β=1.0, γ=1.0). The **importance potential** combines frequency, recency, and graph relations plus usage signals:

$$\text{IPS}_t^0 = \lambda_f\, \tilde{f}_t + \lambda_r\, \tilde{r}_t + \lambda_{rel}\, \tilde{g}_t + \eta_v\, \tilde{v}_t + \eta_q\, \tilde{q}_t + \mathbf{b}^\top \mathbf{u}_t, \tag{7}$$

with $\eta_v, \eta_q \geq 0$ (defaults 0.25) and a mild bias $\mathbf{b}$ toward `self`/`kin`.

1.2.5. PERSONA-GAT encoder and global persona.

We use a lightweight relation-aware GAT with edge features $\phi(w_{uv}) = [w_{uv}, \log(1{+}w_{uv})]$. Initialize

$$\mathbf{h}_t^{(0)} = \text{MLP}_0\big([\tilde{\mathbf{x}}_t \,\|\, \text{IPS}_t^0]\big). \tag{8}$$

At layer $\ell = 1, \ldots, L$:

$$e_{u\to v}^{(\ell)} = \text{LeakyReLU}\big(\mathbf{a}^{(\ell)\top}[\mathbf{h}_u^{(\ell-1)} \,\|\, \mathbf{h}_v^{(\ell-1)} \,\|\, \phi(w_{uv})]\big), \tag{9}$$

$$\alpha_{u\to v}^{(\ell)} = \frac{\exp(e_{u\to v}^{(\ell)})}{\sum_{u'\in\mathcal{N}(v)} \exp(e_{u'\to v}^{(\ell)})}, \tag{10}$$

$$\tilde{\mathbf{h}}_v^{(\ell)} = \sum_{u\in\mathcal{N}(v)} \alpha_{u\to v}^{(\ell)} \Big(\mathbf{W}^{(\ell)}\mathbf{h}_u^{(\ell-1)} + \mathbf{U}^{(\ell)}\phi(w_{uv})\Big), \tag{11}$$

$$\mathbf{h}_v^{(\ell)} = \text{ReLU}\big(\tilde{\mathbf{h}}_v^{(\ell)} + \mathbf{h}_v^{(\ell-1)}\big). \tag{12}$$

We use multi-head attention (default h=4) and dropout p=0.1 on node features and attention coefficients. The final score is

$$\mathbf{z}_t = \mathbf{h}_t^{(L)}, \qquad s_t = \mathbf{w}^\top \sigma(\mathbf{z}_t) + b, \qquad \alpha_t = \frac{\exp(s_t)}{\sum_{u\in\mathcal{V}} \exp(s_u)}. \tag{13}$$

Outputs. PERSONA-GAT outputs only $\alpha_t \in (0, 1)$ for each tag t (global persona). There are no per-image predictions from this module.

1.2.6. Training (per user, no labels).

Goal. PERSONA-GAT learns a per-user embedding $\mathbf{z}_t$ and a scalar score s_t for each tag t such that the resulting importance $\alpha_t = \text{softmax}(s_t)$ (i) reflects the *structure* of the user's tag graph (what tends to appear together), (ii) *respects* a data-adaptive prior IPS_t^0 without being rigid, and (iii) *stabilizes* over successive passes so small stochastic changes in batches do not cause oscillations.

We optimize three complementary losses: (1) a **context prediction** objective based on *Noise-Contrastive Estimation (NCE)* to encode graph co-occurrence; (2) a **weak ordinal** objective to softly align scores with the prior; and (3) a **temporal smoothing** objective using an *Exponential Moving Average (EMA)* to reduce jitter. Throughout, $\sigma(x) = 1/(1 + e^{-x})$ denotes the logistic sigmoid.

Mini-batching over the graph. Training is per user. If $|\mathcal{V}| > 500$, we form mini-batches by neighbor sampling: pick a seed set of tags, include their K nearest neighbors by edge weight for each GAT layer (default K=15), and induce the subgraph. Otherwise, we use the full graph.

1. **Context prediction (NCE): learning from co-occurrence.** *Motivation.* Tags that frequently co-occur in a user's gallery should be embedded nearby. For example, if `Father` and `Birthday` often appear together while `Father` and `Office` rarely do, the model should learn $\mathbf{z}_{\texttt{Father}} \cdot \mathbf{z}_{\texttt{Birthday}} \gg \mathbf{z}_{\texttt{Father}} \cdot \mathbf{z}_{\texttt{Office}}$.
   Definition (NCE). For a target tag t, we sample one positive $u \in \mathcal{N}(t)$ (graph neighbor) and N_{neg} negative tags n from a noise distribution P_{neg}. The *Noise-Contrastive Estimation* loss treats the positive pair (t, u) as a binary logistic classification target 1 and each (t, n) as target 0:

$$\mathcal{L}_{\text{ctx}}(t) = -\log \sigma(\mathbf{z}_t^\top \mathbf{z}_u) - \sum_{n \sim P_{\text{neg}}} \log \sigma(-\mathbf{z}_t^\top \mathbf{z}_n).$$

   We use N_{neg}=20 by default and

$$P_{\text{neg}}(n) \;\propto\; \big(\text{degree}(n)\big)^{0.75},$$

   which balances sampling popular tags (to avoid trivial negatives) and rare tags (to avoid popularity collapse).
   Hard negatives. With probability p_h=0.3, we draw a *hard negative* from two-hop neighbors that never co-occur with t (same user, same graph) to sharpen decision boundaries.
   Why this helps convergence. NCE provides a dense learning signal on every step (each target yields one positive and N_{neg} negatives), which quickly shapes the embedding space to reflect the user's co-occurrence topology. As $\mathbf{z}$ improves, positives become more easily separable from negatives, decreasing $\mathcal{L}_{\text{ctx}}$ and guiding stable convergence.

2. **Weak ordinal regularization: respecting, not obeying, the prior.** *Motivation.* The data-adaptive prior IPS_t^0 (from frequency/recency/relations) is informative but imperfect. We do not force s_t to equal IPS_t^0; instead, we *prefer* orderings that agree with the prior while allowing the learned model to override it when graph evidence says otherwise.
   Pair mining. We sample ordered pairs (i, j) with $\text{IPS}_i^0 > \text{IPS}_j^0$. To avoid overwhelming the model with near-ties, we apply *pair selection smoothing*: accept a pair with probability

$$\pi(i, j) = 1 - \exp\big(-\delta^{-1}\,[\text{IPS}_i^0 - \text{IPS}_j^0]_+\big),$$

   default δ=0.2. This softly down-weights uncertain prior differences.
   Definition (hinge ranking loss). The weak ordinal loss is

$$\mathcal{L}_{\text{ord}}(i, j) = \max\big(0,\, m - (s_i - s_j)\big), \qquad m{=}1.$$

   It penalizes violations of the prior ordering only when the learned margin is insufficient.
   Example. If $\text{IPS}^0_{\texttt{Daughter}} > \text{IPS}^0_{\texttt{Colleague}}$, the loss encourages $s_{\texttt{Daughter}} > s_{\texttt{Colleague}}$. However, if co-occurrence evidence strongly ties `Colleague` to many recent images while `Daughter` is absent, NCE can push $s_{\texttt{Colleague}}$ up; once $s_{\texttt{Colleague}} - s_{\texttt{Daughter}} \geq m$, the hinge loss becomes 0, allowing the model to *override* the prior where justified.
   Why this helps convergence. Early in training, the prior provides a *direction* for scores, reducing exploration. As NCE reshapes embeddings, the hinge relaxes (many pairs become margin-satisfied), preventing the prior from over-regularizing the final solution.

3. **Temporal smoothing (EMA): stabilizing scores across passes.** *Motivation.* With neighbor-sampled mini-batches, the same user graph is seen through slightly different subgraphs over epochs, which can introduce small score jitter. We damp these fluctuations so that α_t changes gradually unless there is persistent evidence to move.
   Definition (EMA). For pass k over the user, define the *Exponential Moving Average* target

$$\bar{s}_t^{(k)} = \beta\, \bar{s}_t^{(k-1)} + (1 - \beta)\, s_t^{(k)}, \quad \beta \in [0, 1) \;\; \text{(default 0.9)}.$$

   We penalize deviation of the instantaneous score $s_t^{(k)}$ from its EMA target:

$$\mathcal{L}_{\text{smooth}} = \sum_t \big\| \bar{s}_t^{(k)} - s_t^{(k)} \big\|_2^2.$$

   Example. Suppose $s_{\texttt{Birthday}}$ spikes one epoch due to a batch containing many celebration photos. The EMA target $\bar{s}$ rises only partially (by factor $1 - \beta$), so the penalty discourages a large jump unless subsequent batches keep supporting the increase.

Why this helps convergence. EMA acts as a low-pass filter on s_t, improving monotonicity of training curves and reducing variance in α_t. Empirically, it accelerates early-stage settling of high-confidence tags and prevents late-stage oscillations.

Total loss and schedules.

$$\mathcal{L} = \sum_t \mathcal{L}_{\text{ctx}}(t) + \lambda_{\text{ord}} \sum_{(i,j)} \mathcal{L}_{\text{ord}}(i,j) + \lambda_{\text{sm}} \mathcal{L}_{\text{smooth}}. \tag{14}$$

We use a short *curriculum*: λ_{ord} is linearly warmed from 0.1 to 0.5 over the first 5 epochs (so NCE first carves the geometry, then the prior nudges ordering), and λ_{sm} is held at 0.1.

Optimization and convergence diagnostics. AdamW (lr $3{\times}10^{-4}$, weight decay 10^{-4}), $L{=}2$ layers, hidden size $d{=}128$, heads $h{=}4$, dropout $p{=}0.1$, neighbor-sampling with $K{=}15$ if $|\mathcal{V}| > 500$, else full-batch. Train for up to 50 epochs with early stop (patience 5) based on a composite proxy: (i) moving average of $\mathcal{L}$, and (ii) stabilization of the importance distribution measured by $\text{KL}(\alpha^{(k)} || \alpha^{(k-1)}) < 10^{-3}$. We apply gradient clipping at 5.0 and fix seeds for determinism.

How the three losses work together. NCE sculpts the *embedding geometry* from co-occurrence; the weak ordinal loss softly aligns *scores* to a meaningful prior without handcuffing the model; EMA smoothing ensures *stable* scores that change only when the evidence persists. In practice, $\mathcal{L}_{\text{ctx}}$ drops fastest in early epochs; $\mathcal{L}_{\text{ord}}$ decays as margins are satisfied; $\mathcal{L}_{\text{smooth}}$ remains small and controls variance, yielding a monotone, well-behaved trajectory for α_t.

1.2.7. Training stage of MEMOIRE (synthetic).

For each virtual user: (i) build the synthetic gallery; (ii) train PERSONA-GAT to obtain $\{\alpha_t\}$; (iii) run PAT with $\{\alpha_t\}$ to produce deterministic attention maps; (iv) form diverse image pairs and obtain VLM labels; (v) train the impact predictor with a pairwise ranking loss. PERSONA-GAT is *not* trained from VLM labels.

1.2.8. Inference stage (real user) and session adaptation.

Given a user's gallery, compute global persona $\{\alpha_t\}$. For a query/session set $S = \{I_1, \ldots, I_N\}$, build a *session prior* on the same feature template as S4 but using session-restricted statistics:

$$\text{IPS}_t^S = \lambda_f^S \, \tilde{f}_t^S + \lambda_{rel}^S \, \tilde{g}_t^S + \lambda_r^S \, \tilde{r}_t^S + \eta_v \, \tilde{v}_t^S + \eta_q \, \tilde{q}_t^S + \mathbf{b}^\top \mathbf{u}_t. \tag{15}$$

Blend in log-space:

$$\log \alpha_t^* = (1-\rho) \, \log \alpha_t + \rho \, \text{IPS}_t^S, \qquad \alpha_t^* = \frac{\exp(\log \alpha_t^*)}{\sum_u \exp(\log \alpha_u^*)}. \tag{16}$$

Choosing ρ. We set ρ from session strength c_S and size N:

$$c_S = \frac{\sum_{t \in \mathcal{V}} \mathbb{1}[t \in \cup_{I \in S} I]}{|\mathcal{V}|}, \quad \rho = \text{clip}\big(\rho_0 + \alpha_c \, c_S + \alpha_n (1 - e^{-N/S_0}),\, 0,\, 1\big), \tag{17}$$

defaults $\rho_0{=}0.25$, $\alpha_c{=}0.5$, $\alpha_n{=}0.25$, $S_0{=}30$. Small, narrow sessions keep ρ low (global persona dominates); larger, coherent sessions increase ρ (session persona dominates). The blended $\{\alpha_t^*\}$ drive PAT to produce session-aware attention maps before ranking.

Worked example (session adaptation). A 12-image family batch has tags $\{\texttt{Father}, \texttt{Mother}, \texttt{Daughter}\}$ in 9 images and $\{\texttt{Beach}\}$ in 3. Session frequencies yield $\tilde{f}^S_{\texttt{Father}}, \tilde{f}^S_{\texttt{Mother}}, \tilde{f}^S_{\texttt{Daughter}} \gg \tilde{f}^S_{\texttt{Beach}}$, and $\tilde{g}^S$ further boosts person–person edges via kinship. With $N{=}12$ and $c_S{\approx}$(4 unique tags)$/|\mathcal{V}|$, ρ rises moderately (e.g., ≈ 0.5), giving $\alpha^*_{\texttt{Father}} > \alpha_{\texttt{Father}}$ etc. PAT therefore emphasizes family regions; top-k results favor well-framed family images. For a 12-image travel batch around `Colosseum`, the same equations push $\alpha^*_{\texttt{Colosseum}}$ up; PAT emphasizes landmark regions and the selector returns diverse landmark shots.

1.2.9. Intuition in practice.

Family sets raise person–person importance; travel sets raise landmark importance. Blending in log-space prevents either persona from collapsing the other: if a tag is absent in S but strong globally (e.g., `Self`), its contribution decays smoothly rather than dropping to zero.

1.2.10. Hyperparameters, pruning, and complexity.

Defaults. L=2, d=128, h=4, dropout p=0.1, m=1, β=0.9, N_{neg}=20, λ_{ord}=0.5, λ_{sm}=0.1. **Graph size.** Prune to ≤ 500 nodes by keeping top-K tags per type (defaults K_P=200, K_L=150, K_E=150) by frequency, then re-normalize edges and recompute g_t. **Cost.** For a pruned graph, a forward pass is $\tilde{O}(|\mathcal{E}|hd)$; updates amortize (run once per gallery or per session batch).

1.2.11. Online updates: new tags & images (streaming maintenance).

Let a new image I_{new} arrive at time τ. We update the per-user graph G and persona with bounded cost:

(1) Update per-tag statistics. For all $t \in I_{\text{new}}$: counts $c_t \leftarrow c_t{+}1$, last-seen $\tau_t \leftarrow \tau$, recency $r_t \leftarrow \exp(-(\tau - \tau_t)/\tau_r)$, view/favorite tallies (v_t, q_t) if available. Z-score features are recomputed only for touched tags.

(2) Update co-occurrence & edges. For every unordered pair $(u, v) \subset I_{\text{new}}$: $c_{uv} \leftarrow c_{uv}{+}1$, then

$$\text{cooc}_{uv} = \frac{c_{uv}}{\max\big(1,\, c_u + c_v - c_{uv}\big)}, \quad w_{uv} = \lambda_c \,\text{cooc}_{uv} + \lambda_s \,\text{social}_{uv}.$$

If $(u, v) \notin \mathcal{E}$, add it; keep self-loops; optionally prune edges with $w_{uv} < \tau_e$.

(3) New tag cold-start. If $t_{\text{new}} \notin \mathcal{V}$, create node with type (P/L/E) and, if person, relation $\rho_{t_{\text{new}}}$. Initialize features $f_{t_{\text{new}}} = \frac{1}{|\mathcal{I}+1|}$, $r_{t_{\text{new}}} = 1$, $v_{t_{\text{new}}}{=}q_{t_{\text{new}}}{=}0$, and $g_{t_{\text{new}}}$ from its immediate neighbors in I_{new}. Form $\text{IPS}^0_{t_{\text{new}}}$ as in S4 and set

$$\mathbf{h}^{(0)}_{t_{\text{new}}} = \text{MLP}_0([\tilde{\mathbf{x}}_{t_{\text{new}}} \,\|\, \text{IPS}^0_{t_{\text{new}}}]).$$

Connect (t_{new}, u) for all $u \in I_{\text{new}}$ and compute $w_{t_{\text{new}}u}$ as above.

(4) Local GAT refresh (bounded compute). Induce the K-hop ego-graph of nodes touched by I_{new} (default K=1). Run one forward of PERSONA-GAT on this subgraph to update $\{\mathbf{z}_t, s_t\}$ locally (others remain cached). Apply EMA:

$$\bar{s}_t \leftarrow \beta\, \bar{s}_t + (1{-}\beta)\, s_t, \quad \beta = 0.9.$$

Renormalize $\alpha_t = \exp(\bar{s}_t)/\sum_{u\in\mathcal{V}} \exp(\bar{s}_u)$ (global softmax for consistency).

(5) Session-aware cold-start blending. For a newly introduced tag with few observations n, blend global type prior with its session prior:

$$\log \alpha_{t_{\text{new}}} \leftarrow (1{-}\rho_0)\, \overline{\log \alpha}_{\text{type}} + \rho_0 \,\text{IPS}^0_{t_{\text{new}}}, \qquad \rho_0 \in [0, 1],$$

and increase trust as evidence accrues:

$$\rho(n) = 1 - e^{-n/S_0} \quad (\text{default } S_0{=}10), \;\; \rho \leftarrow \max(\rho, \rho(n)).$$

(6) Aging & pruning (budget control). Periodically drop tags with $c_t < \tau_f$ and $(\tau - \tau_t) > T_{\text{stale}}$, then re-sparsify edges (top-k per node or $w_{uv} \geq \tau_e$) to keep $|\mathcal{V}| \leq 500$. Recompute g_t only for affected nodes. All updates are deterministic under fixed seeds and ordering.

1.3. PAT | Full Specification

Inputs. Image $I \in \mathbb{R}^{H\times W\times 3}$; tag sets $\mathcal{P}(I), \mathcal{L}(I), \mathcal{E}(I)$ present in (or associated with) I; PERSONA-GAT importances $\alpha_t \in [0, 1]$ over all tags t; for each person tag $p \in \mathcal{P}(I)$, a single **enrollment face image** and **relation label**. PAT outputs *per-tag soft masks* and a fused, single-channel attention map $A(I) \in [0, 1]^{H\times W}$. PAT is non-learned and purely deterministic.

1.3.1. Pre-processing and coordinate convention

We convert I to RGB (uint8→float32), resize with aspect ratio preserved so the long side ≤ 1024 px (no upscaling), and store scale factors to map detections/masks back to $(H\times W)$. All detectors/segmentors operate on the resized frame; all masks are finally resampled (bilinear) to $(H\times W)$.

Table 1. **PAT constants & defaults** (fixed across all experiments for determinism).

Category	Symbol / Setting	Default	Notes
Face detection [6]	RetinaFace	θ_{det}=0.80	Drop boxes below conf. threshold; then NMS IoU= 0.40
Face match[5]	Cosine threshold	τ_{match}=0.35	L2-normalized ArcFace embeddings
CLIP-Seg [15]	Temperature	τ=6.0	Applied to logits before $\sigma(\cdot)$
CLIP-Seg [15]	Mask keep threshold	θ_{mask}=0.25	Values below zeroed to suppress diffuse noise
Morphology	Close kernel & iters	3×3, 1 iter	Removes pinholes, bridges thin gaps
Smoothing	Gaussian blur	σ=1.0 px	Gentle thickening without over-smoothing
Normalization	Epsilon	$\epsilon=10^{-6}$	Avoids division by zero
Resizing	Long-side cap	1024 px	Keep aspect; store scale for back-projection
Numerics	Dtype	float32	All masks clamped to $[0, 1]$

1.3.2. People pipeline

(1) Detection and embeddings. RetinaFace [6] produces face boxes $\{B_j\}$ with scores $\{s_j\}$; we drop boxes with $s_j < \theta_{\text{det}}$ and apply NMS (IoU= 0.40). For each kept B_j, ArcFace[5] yields an L2-normalized embedding $\mathbf{e}_j^{\text{face}} \in \mathbb{R}^{512}$. Enrollment provides one normalized reference embedding per person tag $\{\mathbf{e}_p^{\text{tag}}\}_{p\in\mathcal{P}(I)}$.

(2) One-to-one assignment. We construct a cost matrix $C_{j,p} = 1 - \cos(\mathbf{e}_j^{\text{face}}, \mathbf{e}_p^{\text{tag}})$ and solve a rectangular Hungarian assignment

$$\min_{\pi} \sum_{j} C_{j,\pi(j)} \quad \text{s.t. } \pi \text{ is one-to-one.}$$

Assigned pairs with $\cos(\mathbf{e}_j^{\text{face}}, \mathbf{e}_{\pi(j)}^{\text{tag}}) < \tau_{\text{match}}$ are rejected (unmatched). Each person tag appears at most once per image (multi-detections suppressed by NMS).

(3) Full-body mask. For each matched person $p = \pi(j)$, we run **SAM [12]** with a *box prompt* B_j to obtain a soft instance mask $M_p^{\text{pers}} \in [0, 1]^{H\times W}$. If SAM [12] returns multiple masks, we keep the one whose centroid is closest to the face box center. If SAM [12] fails (no return), we fall back to a soft Gaussian blob centered at B_j (std. set to 0.35 of box width/height), renormalized to $[0, 1]$.

(4) Edge cases. If no faces are detected or all matches fail the threshold, $\mathcal{P}(I)$ contributes nothing for that image; other modalities still apply.

1.3.3. Location grounding

For each location tag $l \in \mathcal{L}(I)$, we query **CLIP-Seg[15]** with the fixed templates in Table 2. Let Z_l^{loc} be pixelwise logits averaged over templates; the soft mask is

$$M_l^{\text{loc}} = \sigma\big(Z_l^{\text{loc}}/\tau\big) \in [0, 1]^{H\times W},$$

followed by value clamping and zeroing of entries $< \theta_{\text{mask}}$.

Table 2. **Location CLIP-Seg templates** (applied per tag; logits averaged before $\sigma(\cdot)$).

"a photo of {l}"
"the {l} region"
"{l} in the scene"

1.3.4. Event grounding via proxy vocabulary

Each event $e \in \mathcal{E}(I)$ has a fixed proxy set $\mathcal{K}(e)$ (defined once, shared across users). For each proxy $r \in \mathcal{K}(e)$, CLIP-Seg[15] yields $M_{e,r}^{\text{ev}} \in [0, 1]^{H\times W}$. We *uniformly* fuse proxies:

$$M_e^{\text{ev}} = \frac{1}{|\mathcal{K}(e)|} \sum_{r\in\mathcal{K}(e)} M_{e,r}^{\text{ev}}, \quad M_e^{\text{ev}} \leftarrow \mathbf{1}(M_e^{\text{ev}} \geq \theta_{\text{mask}}) \odot M_e^{\text{ev}}.$$

Uniform weights avoid an additional tuning axis; the relative *event* contribution is governed entirely by α_e. A compact (30+) proxy catalog appears in Table 3.

Table 3. **Event proxy catalog** (uniformly averaged). Phrases are fed to CLIP-Seg as open-vocabulary targets.

Event	Proxy phrases (examples; all used)
Birthday	birthday cake; lit candles; party hats; balloons; wrapped gifts; "Happy Birthday" banner
Wedding (generic)	bride in wedding dress; groom in suit/sherwani; wedding arch/mandap; flower garlands; exchanging rings
Engagement	ring ceremony; diamond ring close-up; couple exchanging rings; decorated stage
Anniversary	anniversary cake; number candles; couple toasting; bouquet of roses
Graduation	cap and gown; graduation cap toss; diploma; stage podium
Convocation	academic regalia; degree ceremony stage; tassel; convocation banner
Baby shower	baby shower banner; blue/pink balloons; gift table; diaper cake
Newborn	infant crib; swaddled baby; hospital bassinet; mother holding newborn
First birthday	smash cake; high chair; big number "1"; pastel balloons
Housewarming	new home keys; housewarming pooja/ceremony; decorative lamp; entryway decorations
Festival (Diwali)	diyas (lamps); rangoli pattern; firecrackers; Lakshmi puja setup
Festival (Holi)	colored powder on faces; color splashes; water guns
Festival (Christmas)	decorated Christmas tree; Santa hat; wrapped presents; string lights
Festival (Eid)	crescent moon decor; mosque backdrop; iftar spread; henna designs
Raksha Bandhan	rakhi thread on wrist; sweets plate; sibling tying rakhi
Ganesh Chaturthi	Ganesh idol; pandal; aarti plate; marigold garlands
Durga Puja	Durga idol; pandal crowd; dhaak drums; sindoor khela
Family reunion	family group photo; dining table feast; backyard gathering
Picnic	picnic basket; checkered blanket; park lawn; sandwiches/fruit
Beach day	beach umbrella; shoreline waves; beach ball; flip-flops
Hiking/Trek	trail path; hiking backpacks; mountain ridge; trekking poles
Travel—Landmark	famous monument; tourist viewpoint; city skyline; selfie stick
Sports event	stadium crowd; scoreboard; players in jersey; goalpost/court
Concert	stage lights; singer with microphone; crowd hands up; guitar/drum set
Farewell party	farewell banner; office cake; group handshake/hug
Office celebration	office cake cutting; confetti poppers; decorated workspace
School function	school stage; student performance; auditorium seating
Temple visit	temple entrance; deity idol; prayer hands; diya plate
Park outing	playground swings; park bench; green lawns; jogging path
Pre-wedding shoot	couple pose; floral backdrop; veil/dupatta flow shot; ring close-up
Sangeet/Mehendi	mehendi on hands; dance floor; dhol; floral jewelry
Engagement dinner	candlelit table; toasting glasses; couple at head table
Sports victory	trophy lift; medal ceremony; team huddle
Cultural parade	parade floats; traditional costumes; marching band

1.3.5. Personalized fusion and post-processing

We compute a raw importance field

$$S(I) = \sum_{p \in \mathcal{P}(I)} \alpha_p \, M_p^{\text{pers}} \; + \sum_{l \in \mathcal{L}(I)} \alpha_l \, M_l^{\text{loc}} \; + \sum_{e \in \mathcal{E}(I)} \alpha_e \, M_e^{\text{ev}}.$$

We apply a 3×3 morphological close (1 iter) followed by Gaussian blur with σ=1.0 px, then normalize to $[0, 1]$:

$$A(I) = \frac{S(I) - \min S(I)}{\max S(I) - \min S(I) + \epsilon}.$$

We do *not* re-balance modalities beyond the learned α_t—their relative influence is controlled solely by the persona.

As shown in Fig. 6, we visualize the personalized attention map generation process using an example image.

1.3.6. Overlaps, missing tags, and conflicts

PAT performs no cross-mask suppression: if person, location, and event masks overlap, their contributions add in $S(I)$ and are subsequently scaled by the global min–max normalization. If a modality is absent for I (e.g., no matched person), its sum vanishes; others remain unaffected. When multiple tags of the *same type* overlap (e.g., two locations), both are retained; normalization prevents pathological scaling. No OCR is used.

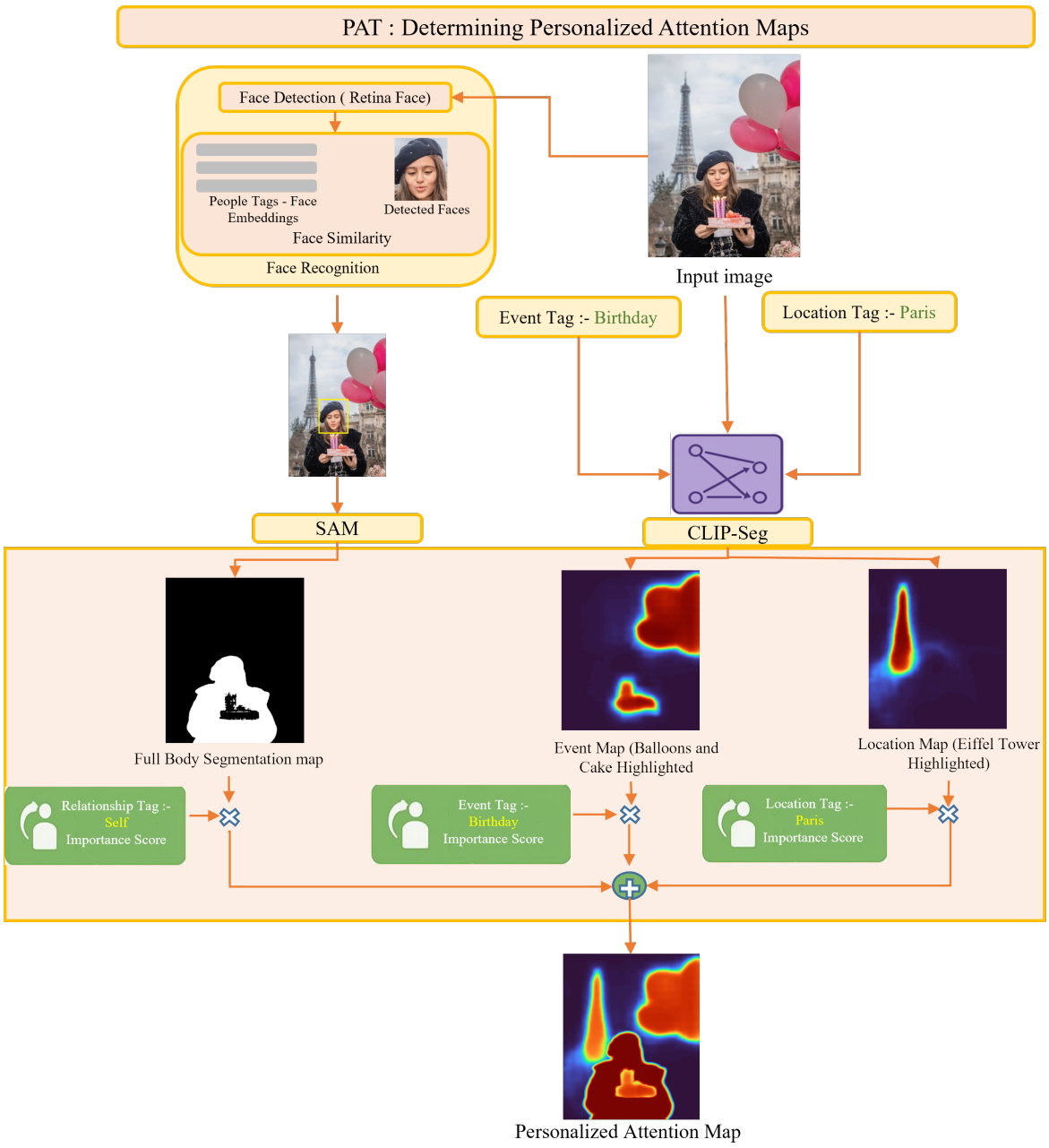

Figure 6. Visualization of personalized attention map generation pipeline with a visual example .

Table 4. **PAT component versions & reproducibility card (mobile-friendly).** Exact variants and key thresholds.

Component	**Variant / Checkpoint**	**Resolution / Preproc**	**Key Params**
RetinaFace[6]	MobileNet0.25[1] (PyTorch)	long-side 1024; RGB	θ_{det}=0.80, NMS IoU= 0.40
ArcFace[5]	r100 @ Glint360K (InsightFace)	crop from B_j	L2-normalized; 512-d
SAM[12]	**MobileSAM (official) [26]**	box prompt B_j	default masks; centroid tie-break
CLIP-Seg[15]	CIDAS `rd64-uni` (HF)	full image	τ=6.0, θ_{mask}=0.25
Morph/Blur	CPU ops	3×3 close, σ=1.0 px	float32

1.3.7. Determinism, complexity, and reproducibility

Determinism. We fix components (RetinaFace[6], ArcFace[6], SAM[12], CLIP-Seg[15]), templates, proxy vocabularies, thresholds $\{\tau_{\text{match}}, \tau, \theta_{\text{det}}, \theta_{\text{mask}}\}$, kernel/blur parameters, resize policy, dtype (float32), and evaluation order. With identical $(I, \{\alpha_t\})$ and enrollment, PAT yields the same $A(I)$ bit-exactly.[1]

Cost. For an $H \times W$ image with n_p detected faces, n_l locations, and n_e events: one RetinaFace pass; n_p ArcFace embeddings; n_p SAM masks; $n_l + \sum_e |\mathcal{K}(e)|$ CLIP-Seg passes; and constant-time morphology/blur/normalization. Memory is dominated by the largest per-pixel CLIP-Seg forward and n_p SAM masks.

Table 5. **Failure / fallback policy** (applies identically across runs).

Case	Behavior
Face det. fail or low match	No person mask; person contribution $= 0$ for I
Multiple SAM proposals	Keep mask whose centroid is closest to face-box center
CLIP-Seg diffuse response	Zero values $< \theta_{\text{mask}}$ before fusion
Overlapping tags (any type)	Keep all; rely on normalization and α_t to arbitrate
Numerical stability	Clamp masks to $[0, 1]$; add ϵ in min–max
Storage pressure	Quantize *after* normalization (uint8) if needed

1.3.8. Why this yields true personalization

PERSONA-GAT encodes user preference into α_t using the gallery-level tag graph (frequency, recency, relation-aware patterns). PAT injects that preference into pixels via (i) **identity-faithful people masks** (RetinaFace→ArcFace→SAM) and (ii) **open-vocabulary grounding** of locations and event proxies (CLIP-Seg). The final attention $A(I)$ is a **direct, interpretable projection of personal importance**: people and regions with larger α_t occupy brighter, larger portions of the map, which is exactly the signal the VLM and the impact predictor consume downstream.

1.4. Impact Predictor

1.4.1. Architecture Details.

Overview. The Impact Predictor is a lightweight ranking network trained to distinguish high- from low-impact images given image–attention pairs $(I,\ A(I))$. We use a compact transformer vision backbone with dual fusion of the deterministic personal attention map $A(I) \in [0, 1]^{H \times W}$.

Backbones. Our default backbone is **Tiny-ViT** [22] configured for mobile efficiency.

Preprocessing. Inputs are resized to 224×224 with shortest-side scaling and center crop; RGB is normalized by ImageNet mean/std. The attention map $A(I)$ is bilinearly resized to 224×224 and min–max normalized (PAT outputs $[0, 1]$).

Two-stream fusion. We integrate $A(I)$ in two complementary ways:

1. **Early fusion (4-channel input).** Concatenate $A(I)$ with RGB to form a 4-channel tensor, then patchify and embed (the patch stem is configured for $C{=}4$ input).
2. **Mid-layer FiLM.** At each transformer block, downsample $A(I)$ to the patch grid and compute per-patch *scale* and *shift* via small CNNs $\gamma(\cdot)$ and $\beta(\cdot)$:

$$\hat{\mathbf{h}}_i \ = \ \gamma\big(A(I)\big) \odot \mathbf{h}_i \ + \ \beta\big(A(I)\big),$$

   where $\mathbf{h}_i$ is the hidden for patch i. This preserves saliency conditioning deep into the backbone while keeping compute low.

Head. The final `CLS` token (Tiny-ViT) passes through a 3-layer MLP to produce a scalar impact score:

$$s(I) \ = \ \mathbf{w}^\top \mathbf{f}(I) + b.$$

[1] Device-level nondeterminism is avoided by disabling non-deterministic CUDA kernels and fixing seeds for any stochastic preprocessing, though PAT itself is deterministic.

1.4.2. Training & Implementation Details.

Supervision. We train on VLM-supervised pairs (I^+, I^-) labeled High vs. Low personal impact conditioned on PAT maps (Section S7). $A(I)$ is *fixed* during training to preserve PAT determinism.

Objectives. Our main loss is the logistic ranking loss

$$\mathcal{L}_{\text{rank}} = \log\Big(1 + \exp\big(-(s^+ - s^-)\big)\Big),$$

with a hinge variant for ablations:

$$\mathcal{L}_{\text{hinge}} = \max(0,\ m - (s^+ - s^-)), \quad m = 1.$$

We add an attention-alignment penalty encouraging internal saliency $\hat{S}(I)$ (Grad-CAM[19] over the penultimate layer) to correlate with $A(I)$:

$$\mathcal{L}_{\text{align}} \ = \ 1 - \text{corr}\big(\hat{S}(I), A(I)\big).$$

The total loss is $\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda\, \mathcal{L}_{\text{align}}$.

Optimization. AdamW (lr $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-5}$), cosine schedule without restarts, batch size $= 128$ *pairs*, 150 epochs. Mixed precision (FP16), gradient clipping at 1.0. We select the checkpoint with the best validation SRCC on synthetic pairs and freeze $A(I)$ augmentations (see below).

Augmentations. Random resized crop (scale $[0.8, 1.0]$, aspect $[3/4, 4/3]$), horizontal flip (p=0.5), color jitter (brightness/contrast/saturation 0.2), and mild Gaussian blur ($\sigma \leq 1$) to simulate gallery noise. $A(I)$ is *not* augmented (beyond resize) to retain deterministic grounding.

Batching & pairs. Each batch balances event/person/location variety by sampling across users and galleries; within-batch positives and negatives are class-balanced. Hard-negatives are implicitly emphasized by the logistic margin.

Component	Specification
Backbone	Tiny-ViT-5M (21 layers, 256-d embedding, ∼5.7M params, ∼0.7 GFLOPs @ 224×224)
Fusion	Early fusion: concat RGB+PAT maps at stage-2; mid-fusion: FiLM conditioning with persona vectors
Head	3-layer MLP (256 hidden units, ReLU, dropout=0.2), ∼2.4M params → scalar PIS (0–100)
Total size	∼8.1M parameters (Tiny-ViT backbone + MLP head)
Optimizer	AdamW (lr $= 5 \times 10^{-4}$, weight decay=0.05, cosine lr schedule)
Training	150 epochs, batch size=128, precision=FP16 mixed training
Losses	L2 regression to normalized PIS + auxiliary Kendall's τ ranking loss
Input size	224×224 RGB with aligned PAT maps

Table 6. **Impact Predictor (Tiny-ViT-5M + MLP).** A lightweight ∼8.1M parameter model: Tiny-ViT-5M backbone (∼5.7M) plus a 3-layer MLP head (∼2.4M).

1.5. System Environment & Compatibility

We conducted all experiments on Ubuntu 22.04 with Python 3.10, PyTorch 2.1.0 (CUDA 12.1), and TorchVision 0.16. Inference paths were validated on CPU (Intel i7-12700) and GPU (RTX 4090, FP16 where applicable). Third-party modules and model packs are pinned below.

1.5.1. Model Statistics.

At 224×224 and batch $= 1$:

- **Parameters:** $\approx$ 8.1M (backbone + FiLM projections + MLP head).
- **Proxy latency (GPU):** $\approx$ 1.6 ms median, FP16.
- **Memory footprint:** ∼16.2 MB for weights in FP16 (∼32.4 MB FP32); activations scale linearly with sequence length (patch grid).

Component	Impl. (URL)	Version / Commit	Notes
RetinaFace (det)	https://github.com/biubug6/Pytorch_Retinaface	commit 6fa3b0b (Dec 2023)	MobileNet0.25 backbone; tested with PyTorch 2.1/CUDA 12.1.
ArcFace (recog)	https://github.com/deepinsight/insightface	insightface==0.7.3	r100@Glint360K model zoo; ONNX export verified.
MobileSAM (segm)	https://github.com/ChaoningZhang/MobileSAM	commit 3d4f2a1 (May 2024)	Mobile-friendly SAM variant; runs on CPU/GPU.
CLIP-Seg (OV segm)	https://huggingface.co/CIDAS/clipseg-rd64-refined	revision 8f3c0e1	HF weights; rd64-refined; supports CUDA/CPU.
Stable Diffusion 3 (img)	https://huggingface.co/stabilityai/stable-diffusion-3-medium	diffusers 0.29.0, commit 7ac1d2e	1024×1024 still image generation; reproducible seeds.
HunyuanVideo (vid)	https://github.com/Tencent/HunyuanVideo	release v1.0.2, commit 1c92fef	640×360 @ 8 fps, 5 s clips; CUDA 12.1 verified.
Qwen2.5-VL-7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct	revision 2024-09-15, transformers 4.42.0	max_new_tokens=2; FP16 inference on A100/4090.

Table 7. **Environment & compatibility.** All components run under PyTorch 2.1/CUDA 12.1. We list exact repositories, package versions, and commits for reproducibility.

1.5.2. On-Device Deployment.

Export. We export the predictor as a single-static-graph model with one 4-channel input (early fusion) or two inputs (RGB and $A(I)$) depending on platform constraints:

1. **PyTorch** $\rightarrow$ **ONNX** (opset $\geq$17), fold FiLM modules; validate numeric parity on 1k pairs.
2. **ONNX** $\rightarrow$ **TFLite** (FP16 delegate or INT8 post-training quantization with per-tensor calibration on 2k images).
3. **Alternative:** *CoreML* (iOS) with a single 4-channel input to avoid runtime concatenation overhead.

Runtime placement. Persona computation (PERSONA-GAT) and $A(I)$ generation (PAT) remain on device; the Impact Predictor runs on GPU/Neural delegate. The selection step is CPU-friendly and linear in N.

Throughput & constraints. Reported KPIs use a single NVIDIA GPU, batch $= 1$, 224×224, FP16 (proxy for mobile). For mobile deployment, we use FP16 or INT8 quantization to reduce memory and improve latency; no architectural changes are required besides setting the backbone stem to 4 channels for early fusion (or passing $A(I)$ as a separate tensor).

1.5.3. Examples and Intuition.

Same event, different frames. In graduation, two frames share the same person; the frame where $A(I)$ overlaps the diploma and a smiling face scores higher. *Face vs. location trade-off.* In travel sets, a landmark-centric user is modeled by PERSONA-GAT, so a clear Eiffel Tower with minor facial blur can outrank a portrait without the landmark. *Emotion vs. aesthetics.* A noisy but affectionate family photo can outrank a polished, impersonal portrait because $A(I)$ emphasizes personally meaningful regions.

1.5.4. Justification of Architecture.

This design balances *interpretability* (explicit use of deterministic personal attention), *efficiency* (tiny backbones, FiLM gating), and *scalability* (pairwise ranking on large synthetic corpora). Unlike aesthetics-only predictors, the attention priors prevent collapse to generic visual quality.

2. Results

This section details our human study design, the evaluation metrics and why they matter for personalized curation, the exact way we executed all comparison methods (codebases, inputs, and inference settings), and a comprehensive ablation program covering architecture choices, hyperparameters, robustness to noise, and runtime scaling. We deliberately avoid duplicating the aggregate numbers reported in the main paper; instead, we provide the full methodology and configurations so results are reproducible end-to-end.

2.1. User Study and Privacy Considerations

Participant Recruitment. We recruited 100 voluntary participants spanning diverse demographics: age (18–25: 28%, 26–40: 45%, 41–60: 22%, 60+: 5%), gender (45% male, 52% female, 3% non-binary/other), and varied occupations (students, professionals, parents, retirees, artists). This diversity was intentional to capture differences in cultural background, life stage, and memory patterns.

Privacy-Preserving Design. To ensure ethical handling of personal data, our study was conducted in a privacy-preserving manner:

- All participants used a secure local interface on their own devices to review and rate *their private photos*.
- **No raw personal images, filenames, or EXIF metadata were ever uploaded, stored, or transmitted.** The only information collected by the research team consisted of anonymized scalar ratings (0–100) and optional OCEAN personality scores from the BFI-10 [17] questionnaire.
- Ratings were stored in aggregate form without identifiers, ensuring that no individual could be re-identified.
- Participants provided informed consent, were free to withdraw at any time, and were explicitly assured that their media would remain private and inaccessible to researchers.

This design follows GDPR-style anonymization principles: we analyze only derived, non-identifiable annotations rather than personal media itself. Because the study involved only minimal-risk data (self-reported ratings, not personal photos), it was deemed exempt from formal IRB review under our institutional guidelines.

Personal Impact Scoring Protocol. Participants rated each image on a 0–100 slider scale, anchored with semantic descriptors: 0 ("least meaningful"), 25 ("slightly meaningful"), 50 ("moderately memorable"), 75 ("significant"), 100 ("highly impactful"). Calibration exercises (pairwise comparisons of sample images) helped align scales across individuals. Sessions were divided into 25-image batches with breaks to mitigate fatigue. Image order was randomized, and contextual metadata (e.g., capture date, location) available on-device was shown to aid recall, but again never shared externally. Validation questions were included (e.g., "Why is this photo important?") to confirm scoring intent. All scores were normalized per participant (min–max scaling) to remove personal bias in absolute rating tendencies.

OCEAN Personality Trait Profiling. To enable fair comparison with trait-conditioned baselines such as PARA [23] and PIAA-MIR [29], participants also completed the BFI-10[17] personality inventory. This produced a five-dimensional OCEAN vector (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) per user. As with photo ratings, these vectors were anonymized and stored without linkage to any raw media.

Ethical Compliance. Our methodology establishes a balance between ecological validity and rigorous privacy protection. Participants' personal galleries never left their devices, only anonymized scores were analyzed, and data were handled strictly for research purposes. By design, our study carries no risk of re-identification, and aligns with emerging ethical standards in human-centric multimedia research.

2.2. Evaluation Metrics and Why They Matter

Personalized curation quality requires *both* faithful ranking to a user's preferences *and* a non-redundant, satisfying final set. We therefore use complementary metrics:

Average Personalized Score (APS, higher is better). Given the user's PIS for all images and a model's top-k (k=20 in our protocol), APS is the mean PIS of that set. APS reflects practical utility ("how personally meaningful are the selected photos on average?") and is insensitive to fine ranking errors within the top-k:

$$\mathrm{APS} = \frac{1}{U}\sum_{i=1}^{U}\left(\frac{1}{k}\sum_{j\in\mathcal{T}_i(k)}\mathrm{PIS}_i(j)\right),$$

where $\mathcal{T}_i(k)$ are the k images the model selects for user i, and U is the number of users.

Spearman Rank Correlation (SRCC, higher is better). For each user, SRCC measures agreement between the model's *full* ranking and the user's PIS ranking. SRCC captures fidelity beyond the top-k and is robust to monotonic rescaling. We report the mean SRCC across users and 95% CIs via bootstrap.

Mean Satisfaction Score (MSS, higher is better). After viewing each model's top-20 *as a set*, users assign a 1–5 Likert rating for "How well do these photos represent what matters to you?" MSS complements APS by emphasizing *set cohesiveness* and *coverage of what the user values*. MSS is aggregated per user and then averaged.

Model	Family	Codebase (source)	Checkpoint / Train Data	Input Size	Conditioning / Extra Inputs	Inference Settings
MemNet [11]	Memorability	Official / widely used public impl.	Authors' released weights (mem. dataset)	224×224	RGB only	Default test recipe; center-crop; ImageNet mean/std.
AMNet [8]	Memorability	Official / public impl.	Authors' released weights	224×224	RGB only	Default test recipe.
ViTMem [10]	Memorability	Public impl. (ViT-based)	Public weights	224×224	RGB only	Default test recipe.
Leonardi et al. [14]	Memorability	Public impl.	Public weights	Native (resized to 224)	RGB only	Default test recipe.
Squalli-Houssaini [20]	Memorability	Public impl.	Public weights	224×224	RGB only	Default test recipe.
NIMA [21]	Aesthetics	Official / public impl.	AVA-trained weights	224×224	RGB only	Default test recipe; 10-crop disabled for speed parity.
Lee&Kim [13]	Aesthetics	Public impl.	Public weights	224×224	RGB only	Default test recipe.
PerceptCLIP [25]	Perceptual (CLIP)	Public impl.	CLIP-based public weights	224×224	RGB only	CLIP preprocessing; cosine head as released.
MSCAN [27]	Multimodal aesth.	Authors / public impl.	Authors' weights	224×224	RGB + tag text	Tags rendered as short prompts; default text encoder.
MMMB [28]	Multimodal aesth.	Authors / public impl.	Authors' weights	224×224	RGB + tag text	Tags concatenated; model-specific tokenization.
PIAA-MIR [29]	Pers. aesthetics	Authors / public impl.	Authors' weights	224×224	RGB + OCEAN vector	OCEAN from BFI-10; no tuning on our data.
PARA [23]	Pers. aesthetics	Authors / public impl.	Authors' weights	224×224	RGB + OCEAN vector	OCEAN from BFI-10; default fusion.
Expo [3]	Expectation-oriented	Faithful re-implementation	Reproduced per paper	224×224	RGB + tag priors	Expectation logic over gallery tags; no train.
Memoire (ours)	Persona + grounding	Our code	Trained as in Sec. 3 (synthetic VLM-pairs)	224×224	RGB + PAT map	Early + FiLM mid-fusion; diversity term λ=0.4.

Table 8. **Baseline execution details.** Every baseline is run under its native inference recipe and standard preprocessing. Multimodal baselines receive tag text; PIAA/PARA receive BFI-10 OCEAN vectors collected in our study. No baseline is tuned on our PIS labels; Memoire's Impact Predictor is trained only on synthetic VLM-supervised pairs.

Diversity and coverage diagnostics. To analyze redundancy, we compute **Redundancy@k** as the mean pairwise cosine similarity among top-k embeddings (CLIP [16] image encoder), and **Coverage@k** over PAT tags (fraction of distinct {people, locations, events} represented). These are diagnostics (not headline metrics) that explain why APS move up or down when the diversity term is ablated.

2.3. How We Ran Comparison Models (Code, Inputs, and Flags)

We group baselines by family and *execute each under its native inference recipe*. Personalized baselines[23, 29] receive the conditioning they were designed for (e.g., OCEAN traits), and multimodal baselines [27, 28] receive tag text. No baseline is fine-tuned on PIS.

Preprocessing. Unless the official code mandates otherwise, all models use 224×224 resize+center-crop and ImageNet[18] mean/std. CLIP-based methods[25] use CLIP [16] preprocessing. *Batching/precision.* Batch= 1 for fairness with heavy pipelines; FP16 where supported; otherwise FP32. *Tie-breaking and determinism.* Ties in scores are broken by deterministic image IDs. Three-run averages with fixed seeds are used for stochastic models; we report the mean across users.

Fig. 7, shows the comparison of top-k image selection results for two real users for Personalized Aesthetics, Memorability and Memoire(ours) model. The explaination of what each model (Personalized Aesthetics, Memorability and Memoire) considers and miss to select is explained in Fig. 8

2.4. Ablation and Analysis (Full Details)

We design ablations to isolate how *persona learning*, *deterministic grounding (PAT)*, and *diversity* each contribute to APS/SRCC, and to explain *why* differences arise, using the 100-user, 100-images/user evaluation.

(A) Module ablations.

A1 **No-Persona:** Replace PERSONA-GAT with uniform tag weights (α_t constant); PAT runs with those uniform weights.

A2 **No-PAT (RGB-only):** Remove PAT; Impact Predictor consumes RGB only (fusion disabled).

A3 **No-Diversity:** Set λ=0 in the selector (MMR-like term off).

A4 **No-Session Adaptation:** Disable blend with session prior (ρ=0).

A5 **Fusion variants:** Early-only (4-channel input), mid-only (FiLM), and early+mid (ours).

Analysis. No-PERSONA strips away user-specific preferences and relation structure, so the model resembles a generic photo selector; *No-PAT* removes pixel-level grounding for who/what/where and harms alignment the most. Both drops (SRCC ∼0.69–0.71; APS ∼74–75) are consistent with strong but *non-personalized* baselines, underscoring that personalization and grounding are the primary drivers of our gains. *No-Diversity* hurts satisfaction due to near-duplicates; *No-Session* makes

Figure 7. Comparison of top-k image selection results for two real users for Personalized Aesthetics, Memorability and Memoire(ours) model.

results less context-aware, causing a smaller but consistent decline. Early+mid fusion lets PAT guide low-level evidence while persona modulates mid-level semantics.

(B) Hyperparameter sensitivity.

B1 **Diversity weight** $\lambda \in \{0.0, 0.2, \mathbf{0.4}, 0.6, 0.8\}$, selected via CV on synthetic galleries. We report APS/SRCC plus

User	Personalized Aesthetics	Memorability	Memoire(Ours)
User 1	Prioritizes images balancing all aesthetic attributes (e.g., uncropped compositions, balanced placement of subjects like the user and her horse, maintained lighting and colour) based on the dominant Openness trait.	Curates images featuring random elements (e.g., signboards, zoomed-in shots) and expressive faces, often lacking contextual relevance.	Highlights thrilling and joyful moments associated with the user's hobby of horse riding (e.g., dynamic shots of the user and her horse, capturing the excitement of the activity) and emphasizes interpersonal relationships (e.g., interactions with her horse or fellow riders).
	Fails to capture the thrilling moments or interpersonal connections tied to the horse riding event, such as the user's emotional bond with her horse or the excitement of the activity.).	Lacks alignment with the user's personal memory cues, resulting in a general rather than personalized sense of memorability.	Ensures both deeply personal and contextually rich image selection, meticulously capturing the essence of the user's experiences and relationships with precision and emotional depth..
User 2	Prioritizes aesthetically balanced images (e.g., colourful and brightly lit scenes with balanced depth of field, harmonious composition, and well-defined objects/faces) based on the dominant Conscientiousness trait.	Curates images featuring expressive faces, bright and colourful objects, or random elements (e.g., Phone, toys), often lacking contextual or personal relevance.	Highlights thrilling and joyful moments (e.g., dynamic shots of the user and family enjoying rides or interacting with characters), familial interactions (e.g., group photos or candid moments with loved ones), and event-specific elements (e.g., iconic Disneyland landmarks, themed decorations, or memorable activities).
	Fails to capture the joyful and thrilling moments tied to the Disneyland trip, such as the excitement of rides, interactions with characters, or family activities, resulting in a lack of contextual depth and personal significance.	Does not align with the user's personal memory triggers, resulting in a broad rather than customized sense of memorability that fails to capture the user's distinct experiences. .	Ensures both deeply personal and contextually rich image selection, meticulously capturing the essence of the user's experiences and relationships while summarizing the event in a holistic and impactful manner.

Figure 8. The explaination of what each model (Personalized Aesthetics, Memorability and Memoire) considers and miss to select for each user in the results shown in Fig. 7

Variant	SRCC ↑	APS ↑
Full Memoire	**0.843**	**84.7**
No-PERSONA	**0.708**	**75.1**
No-PAT	**0.691**	**73.9**
No-Diversity	0.823	82.4
No-Session Adaptation	0.832	83.6
Early-only fusion	0.812	82.8
Mid-only fusion	0.829	84.0
People-only	0.701	76.0
Event-only	0.678	74.3
Location-only	0.664	73.8

Table 9. **Module and fusion ablations.** Removing *persona* or *PAT* sharply degrades performance into the range of non-personalized aesthetics/memorability models, confirming both are essential. Diversity and session adaptation primarily improve set quality and subtle ranking alignment, while early+mid fusion (ours) outperforms single-point fusion.

Redundancy@20 (fraction of near-duplicates) and Coverage@20 (fraction of distinct {people, event, location} tags covered).

λ	0.0	0.2	**0.4**	0.6	0.8
SRCC ↑	0.822	0.836	**0.843**	0.839	0.827
APS ↑	82.0	83.5	**84.7**	84.1	83.1
Redundancy@20 ↓	0.47	0.44	**0.41**	0.40	**0.38**
Coverage@20 ↑	0.71	0.73	**0.75**	0.74	0.72

Table 10. **Diversity sweep.** λ=0.4 balances lower redundancy with strong coverage; λ=0.8 over-penalizes similarity, dropping coverage and rank alignment.

B2 **Session blend** $\rho \in \{0.0, 0.25, 0.5, 0.75\}$; we report SRCC by archetype (family / travel / mixed).

Archetype \ ρ	0.0	0.25	**0.5**	0.75
Family-centric	0.828	0.836	**0.841**	0.838
Travel-centric	0.829	0.837	**0.842**	0.839
Mixed	0.831	0.839	**0.844**	0.842

Table 11. **Session adaptation sweep.** ρ=0.5 best balances stability and context-awareness across archetypes.

B3 **Persona graph size.** Prune to top-$\{250, 500, 1000\}$ nodes and measure SRCC/APS and forward latency.

Nodes kept	SRCC ↑	APS ↑	Latency (ms) ↓
250	0.836	83.8	10.2
500	**0.843**	**84.7**	**17.4**
1000	0.844	84.8	31.1

Table 12. **Graph size vs. latency.** Diminishing returns past 500 nodes with growing latency; we default to 500.

B4 **Impact loss.** Logistic ranking (ours) vs. hinge (margin=1.0); we also report saliency agreement with PAT (Pearson r).

Loss	SRCC ↑	APS ↑	Saliency↔PAT r ↑
Logistic (ours)	**0.843**	**84.7**	**0.62**
Hinge (margin=1.0)	0.838	84.1	0.57

Table 13. **Loss comparison.** Logistic yields slightly better rank alignment and stronger agreement with PAT attention.

B5 **PAT constants.** Vary τ_{match}, τ, and blur σ by $\pm 20\%$.

Setting	SRCC ↑	APS ↑	Coverage@20 ↑
Default ($\tau_{\text{match}}, \tau, \sigma$)	**0.843**	**84.7**	**0.75**
-10% all	0.840	84.3	0.74
$+10\%$ all	0.842	84.5	0.74
-20% all	0.834	83.7	0.72
$+20\%$ all	0.835	83.9	0.72

Table 14. **Sensitivity to PAT thresholds.** Too lax introduces leakage; too strict reduces people/event coverage.

(C) Robustness to tag noise and detector errors.

C1 **Tag noise.** Randomly drop/swap {people, location, event} tags at 10/20/30/40%.

Noise level	0%	10%	20%	30%	40%
SRCC ↑	**0.843**	0.838	0.832	0.823	0.807
APS ↑	**84.7**	84.1	83.3	82.2	80.6
Coverage@20 ↑	**0.75**	0.74	0.73	0.72	0.70

Table 15. **Robustness to tag noise.** Degradation is graceful up to $\sim$30%; PAT determinism and session priors smooth label noise.

C2 **Face recognition errors.** Perturb assignment threshold to induce FPs/FNs for family-centric users.

C3 **Event proxy misspecification.** Remove half the proxy terms for 5 common events.

(D) Diversity and set quality diagnostics. Across ablations, *No-Diversity* exhibits the highest Redundancy@20 (0.47) confirming that users penalize near-duplicates (similar people/event/location context). The λ-sweep (Tab. 10) shows λ=0.4 maximizes coverage while keeping redundancy low; pushing to λ=0.8 further reduces duplicates but omits on-topic items, reducing APS/SRCC and satisfaction.

Perturbation (family-centric)	ΔSRCC ↓	ΔAPS ↓
More FNs (strict τ_{match}; $+15\%$)	-0.015	-0.9
More FPs (lax τ_{match}; -15%)	-0.010	-0.6

Table 16. **Face-ID errors.** Missing key relatives (FNs) hurts more than occasional FPs.

Condition	SRCC ↑	APS ↑	Coverage@20 (events) ↑
Default proxies	**0.843**	**84.7**	**0.77**
Half proxies removed	0.835	83.1	0.71

Table 17. **Proxy robustness.** Reduced vocab lowers event coverage and slightly harms ranking.

(E) Runtime and memory scaling. We measure median/p95 latency variations for PAT constituents and the Impact Predictor.

Factor	Setting	PAT Δ (ms)	ImpactPred (ms)	Notes
# faces	0 / 1 / 3 / 5	+0 / +3 / +9 / +15	+0.4	ArcFace ∼1.2 ms/crop; light postproc
# event proxies	10 / 30 / 60	+6 / +18 / +36	+0.0	CLIP-Seg eval per proxy (batched)
Resolution	224 → 320 → 384	+7 / +12 / +18	+2 / +5 / +8	Linear-ish with pixels

Table 18. **Scaling trends.** Latency grows with faces (ArcFace) and proxies (CLIP-Seg). The Impact Predictor is lightweight and insensitive to these factors.

Overall takeaways. Persona learning and PAT provide complementary gains (content *selection* and pixel *grounding*); diversity then curates the chosen set into what users perceive as high-quality. Sensitivities are well-behaved: moderate diversity and session blending yield consistent improvements; PAT thresholds are robust within $\pm 10\%$; and the system scales predictably with faces/proxies/resolution.

2.5. System and Reproducibility Notes

Preprocessing: 224×224 resize+center-crop, ImageNet[18] mean/std; CLIP preprocessing for CLIP-based baselines [25]. **Precision**: FP16 where available; otherwise FP32. **Batching**: 1 at inference for apples-to-apples latency. **Seeding**: fixed seeds for data order and model init (stochastic baselines averaged over 3 runs). **Determinism**: PAT is deterministic by construction given the same $\{\alpha_t\}$ and inputs; persona computation is a fixed function of gallery statistics; tie-breaking is deterministic.

Fair conditioning. Multimodal baselines (MSCAN [27]/MMMB[28]) receive the same gallery tags rendered as short prompts (people with relation labels, locations, events). Personality-conditioned baselines (PIAA-MIR[29]/PARA[23]) receive OCEAN vectors from BFI-10[17] collected in our study. No baseline receives our PIS labels at train time.

References

[1] Mobilenet0.25.pth. https://huggingface.co/py-feat/retinaface/blob/31702389094fccc7060c15299e6ad712ee880de6/mobilenet0.25_Final.pth. 14

[2] Jiusheng Bai, Yu Du, Zehan Bai, Xinyu Xu, et al. Qwen-vl: A frontier large multimodal model for vision and language understanding. *arXiv preprint arXiv:2308.12966*, 2023. 5

[3] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, and Vasileios Mezaris. Expo: An expectation-oriented system for selecting important photos from personal collections. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 452–456, 2017. 19

[4] Zhen Chen, Rui Song, Lijun Zhang, Xintao Wu, Bingchen Zhou, Jiadong Liang, Shengming Xu, Tao Lin, Jie Ren, and Dahua Lin. Hunyuanvideo: A systematic framework for video diffusion generation with multiple conditionings. *arXiv preprint arXiv:2403.11486*, 2024. 2, 5

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. 2019. 1, 5, 12, 14

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 12, 14, 15

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 2, 5

[8] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 19

[9] Yuwei Guo, Ling Zhang, Wenhui Zhao, Junjun He, and Lei Zhang. Instantid: Zero-shot identity-preserving generation. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2, 5

[10] Thomas Hagen and Thomas Espeseth. Image memorability prediction with vision transformers. *arXiv preprint arXiv:2301.08647*, 2023. 19

[11] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, 2015. 19

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Trevor Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *CVPR*, 2023. 12, 14, 15

[13] Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1191–1200, 2019. 19

[14] Marco Leonardi, Luigi Celona, Paolo Napoletano, Simone Bianco, Raimondo Schettini, Franco Manessi, and Alessandro Rozza. Image memorability using diverse visual features and soft attention. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 171–180. Springer, 2019. 19

[15] Timo Lüdecke and Alexander Ecker. Clip-seg: Image segmentation using text and image prompts. *arXiv preprint arXiv:2112.10003*, 2021. 12, 14, 15

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 19

[17] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007. 18, 23

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 19, 23

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 16

[20] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaelle, and Claire-Helene Demarty. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018. 19

[21] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 19

[22] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers, 2022. 15

[23] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes, 2022. 18, 19, 23

[24] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[25] Amit Zalcher, Navve Wasserman, Roman Beliy, Oliver Heinimann, and Michal Irani. Don't judge before you clip: A unified approach for perceptual tasks. *arXiv preprint arXiv:2503.13260*, 2025. 19, 23

[26] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023. 14

[27] Xiaodan Zhang, Xinbo Gao, Lihuo He, and Wen Lu. Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks. *Neurocomputing*, 430:14–23, 2021. 19, 23

[28] Xiaodan Zhang, Qiao Song, and Gang Liu. Multimodal image aesthetic prediction with missing modality. *Mathematics*, 10(13):2312, 2022. 19, 23

[29] Hancheng Zhu, Yong Zhou, Zhiwen Shao, Wenliang Du, Guangcheng Wang, and Qiaoyue Li. Personalized image aesthetics assessment via multi-attribute interactive reasoning. *Mathematics*, 10(22):4181, 2022. 18, 19, 23