**Evaluation of Model Performance**

To understand how well our model performs on restoring GIF animations, we used a combination of **automated quantitative metrics** and **manual human evaluation**. This two-fold approach helps us assess both the technical quality and the perceptual quality of the restored outputs.

**Automated Metrics Explained**

These metrics provide objective measurements based on mathematical calculations between the restored GIF frames and the original or reference frames:

1. **PSNR (Peak Signal-to-Noise Ratio)**
   * Measures the similarity between the restored frame and the original frame in terms of pixel accuracy.
   * The higher the PSNR value (measured in decibels, dB), the better the restoration quality. Typically, a PSNR above 30 dB is considered good for image/video restoration tasks.
   * For example, a PSNR of 33 means the restored image is very close to the original.
2. **SSIM (Structural Similarity Index)**
   * Evaluates how similar the restored frame and original frame are in terms of structure, luminance (brightness), and contrast.
   * SSIM values range from 0 to 1, where 1 means identical images. A value above 0.8 is generally a good indicator of visual similarity.
   * Think of SSIM as checking if the objects and textures in the image look natural and consistent.
3. **LPIPS (Learned Perceptual Image Patch Similarity)**
   * Measures perceptual similarity using deep neural network features instead of just pixel-wise differences.
   * Lower LPIPS values mean better perceptual similarity (i.e., the image "looks" more like the original to a human eye).
   * LPIPS below 0.2 is often considered good quality.
4. **Temporal Loss**
   * Measures consistency between consecutive frames to avoid flickering or jittering in video/GIF playback.
   * This is calculated as the difference between one frame warped according to estimated motion and the actual next frame.
   * Lower values indicate smoother, more stable video sequences.

**Manual Human Ratings**

While automated metrics are important, they cannot always capture subjective qualities such as whether a person finds the restored GIF visually pleasing or natural. Therefore, we included manual evaluation by rating the outputs on a scale of **1 (poor)** to **5 (excellent)** on four criteria:

* **Restoration Quality:** How clear and artifact-free the restored GIF looks.
* **Identity Preservation:** Whether key visual features and characteristics (like faces or objects) are faithfully maintained.
* **Temporal Smoothness:** How smooth and consistent the animation is across frames without distracting flicker.
* **Overall Impression:** A holistic rating combining all aspects of quality.

**Summary of Results**

For 30 diverse GIFs—including clear, blurred, noisy, and low-resolution examples—we obtained the following average results:

| **Metric** | **Average Value** | **Interpretation** |
| --- | --- | --- |
| PSNR | 30.2 dB | Good pixel-level accuracy |
| SSIM | 0.81 | Strong structural similarity |
| LPIPS | 0.15 | High perceptual similarity |
| Temporal Loss | 0.0128 | Good temporal smoothness |
| Restoration Quality | 4.0 / 5 | Very good, minor imperfections |
| Identity Preservation | 4.0 / 5 | Most key features preserved |
| Temporal Smoothness | 4.0 / 5 | Smooth playback |
| Overall Impression | 4.1 / 5 | Positive visual experience |

**Context & Interpretation**

* **Why multiple metrics?** No single metric can fully describe the quality of restored GIFs because restoration involves various aspects: sharpness, naturalness, stability over time, and perceptual realism. Using several metrics provides a fuller picture.
* **Why manual ratings?** Sometimes, an image that scores well on numbers may still look unnatural or contain artifacts visible only to the human eye. The manual ratings help verify the automated results and capture subjective preferences.
* **Understanding the numbers:** For example, a PSNR of 30 dB means the image noise is very low, so the GIF looks clean. An SSIM of 0.8 means most image structures (like edges and textures) are well maintained. Low LPIPS means perceptually the restored image "feels" right. Temporal loss ensures frames don’t flicker or jump abruptly.
* **Categories of GIFs:** We tested on various types: **blurred** (out-of-focus), **noisy** (grainy), and **low resolution** GIFs. As expected, the model’s performance was slightly lower on challenging types but remained reasonable, showing robustness.
* **Practical impact:** These results mean the model can effectively restore old, degraded, or compressed GIFs to look clearer and smoother, making them more enjoyable for viewers or usable in professional media workflows.

**Manual Review Guidelines (For Future Evaluators)**

To maintain consistency in manual ratings, reviewers should:

* Watch the restored GIF multiple times on a good display.
* Compare against original degraded GIFs to see improvements.
* Focus on clarity, natural textures, faithful reproduction of faces/objects, and smooth frame transitions.
* Rate each category from 1 (poor, many artifacts, flickering, or loss of identity) to 5 (excellent, clear, natural, and smooth).
* Note any distracting flicker, unnatural colors, or missing details in notes (optional).

**Final Thoughts**

Combining quantitative metrics and human evaluation provides a comprehensive understanding of model performance. The results demonstrate that the restoration approach is effective, yielding high-quality GIF outputs that retain both pixel-level and perceptual fidelity, while maintaining temporal coherence essential for animated media.

This evaluation validates the model’s utility and provides a clear framework for assessing restoration quality in future work or deployments.