

Matrix Conventions In Computer Graphics

Prefix versus Postfix conventions

There are two conventions used by mathematicians to compose functions (you might have learned this in your math courses): *prefix* and *postfix*. Note that these are conventions that are used to both think about and represent functions – the final result using either convention should give the same answer. In general, OpenGL, GLSL, and GLM can be thought of as using prefix convention while DirectX uses the postfix convention.

Prefix convention is more natural. When we want to apply a function f on some domain x , we write $f(x)$. If we want to apply a function g on the domain $f(x)$, we write $g(f(x))$ or $(gf)(x)$. This is how we treat the composition of mappings or transforms in class lectures and in OpenGL code - that is, first apply transformation matrix f and then apply transformation matrix g . So, if we use a prefix convention, points and vectors must be manifested as column matrices to satisfy the rules of matrix multiplication. In class lectures, notes, assignments, and in OpenGL code, we follow the prefix convention.

Consider the following code fragment that performs some matrix computations:

```

1 // compute model-to-world-to-view-clip transformation matrix ...
2
3 // first, compute perspective transform matrix:
4 glm::mat4 proj_mtx = glm::perspective(glm::radians(45.f),
5                                     (GLfloat)fb_width / fb_height,
6                                     1.f, 200.f);
7
8 // compute view transform matrix:
9 glm::mat4 view_mtx = glm::lookAt(glm::vec3(0.f, 0.f, 10.f),
10                                glm::vec3(0.f, 0.f, 0.f),
11                                glm::vec3(0.f, 1.f, 0.f));
12
13 // compute model transform matrix:
14 // scale transform followed by rotation followed by translation
15 glm::mat4 trans_mtx = glm::translate(glm::mat4(1.f), glm::vec3(1.f));
16 glm::mat4 rot_mtx = glm::rotate(glm::mat4(1.f),
17                                glm::radians(s_angular_displacement),
18                                glm::vec3(0.f, 0.f, 1.f));
19 glm::mat4 scl_mtx = glm::scale(glm::mat4(1.f), glm::vec3(2.f));
20
21 // now, compute model-to-world-to-view-clip transformation matrix ...
22 glm::mat4 mvp_mtx = proj_mtx * view_mtx * trans_mtx * rot_mtx * scl_mtx;
```

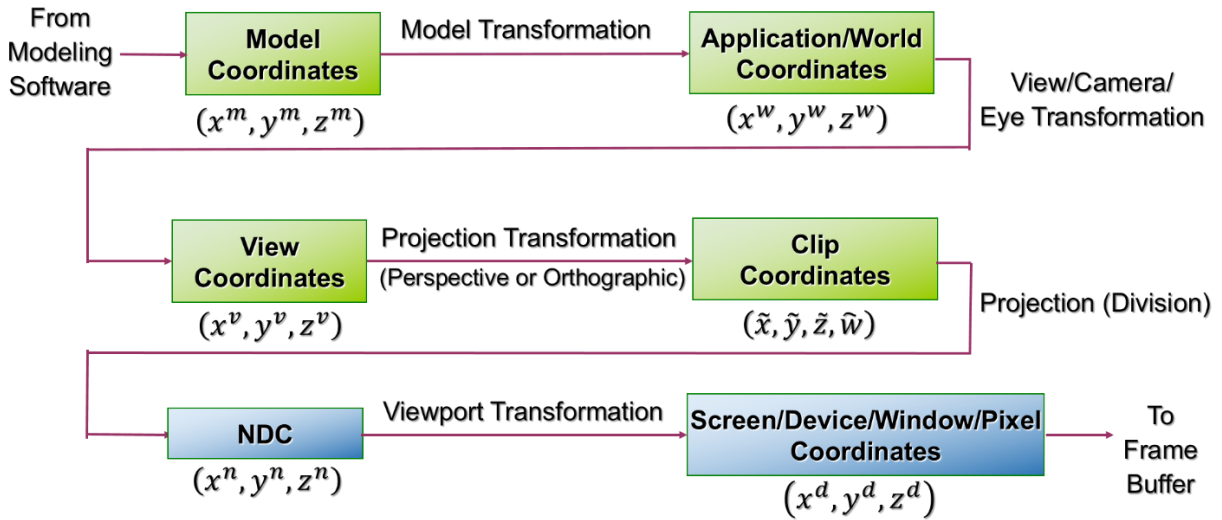
Five matrices are used in this code fragment: a perspective transformation, a view transformation, a translation transformation, a rotation transformation, and a scale transformation. The prefix convention is being used here to compute the final composite matrix. If P^m represents a point in the model coordinate system and \tilde{P} represents a point in the clip coordinate system, matrices are being applied on point P^m in the following prefix order:

$$M_{perspective} M_{view} M_{translation} M_{rotation} M_{scale} P^m = \tilde{P}$$

Since points are represented as column matrices, the following matrix multiplication holds:

$$M_{perspective} M_{view} M_{translation} M_{rotation} M_{scale} \begin{bmatrix} x^m \\ y^m \\ z^m \\ 1 \end{bmatrix} = \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \\ \tilde{w} \end{bmatrix}$$

First, model transformation consisting of scale, rotation, and translation (in that order), followed by the view and perspective transformations are being applied. This is borne out by Figure 1 which illustrates the passage of a vertex through the graphics pipeline:



The projection (division) and viewport transformation seen in Figure 1 are not included in the code fragment.

The thing to remember about prefix convention is that you write down the matrices in the order in which they'll be applied from right-to-left and compute the composite matrix from left-to-right.

On the other hand, if you're using the postfix convention, then points and vectors are manifested as row matrices. The five matrices would then be applied on points using postfix convention:

$$[x^m \ y^m \ z^m \ 1] M_{scale}^T M_{rotation}^T M_{translation}^T M_{view}^T M_{perspective}^T = [\tilde{x} \ \tilde{y} \ \tilde{z} \ \tilde{w}]$$

We use the notation M_{scale}^T to denote that this [postfix convention] matrix is the transpose of the scale matrix M_{scale} used in the prefix convention. The postfix convention is used by DirectX. Note that the above picture can also represent the DirectX graphics pipeline – the representation of the matrices is what makes DirectX matrices different from OpenGL, GLSL, and GLM matrices.

Row-Major or Column-Major layout of matrices

Graphics APIs represent 4×4 matrices as an array of 16 elements. These 16 values can be interpreted using either row-major or column-major layout. In row-major layout, all elements of the first row come before all elements of the second row, and so on, and, within each row, elements are ordered by column. In contrast, in column-major layout, all elements of the first column come before all elements of the second column, and so on, and, within each column, elements are ordered by row. Consider the following C/C++ array of 16 elements:

```

1 float mtx[16] = {
2     m0, m1, m2, m3,
3     m4, m5, m6, m7,
4     m8, m9, m10, m11,
5     m12, m13, m14, m15
6 };

```

In a graphics API using *row-major* layout, the first set of four values `m0, m1, m2, m3` in array `mtx` represent the 1st row of the corresponding matrix, the second set of four values `m4, m5, m6, m7` represent the 2nd row of the corresponding matrix, and so on. That is, the graphics API using row-major layout would interpret the 16 values in array `mtx` as the following matrix:

$$\begin{bmatrix} m0 & m1 & m2 & m3 \\ m4 & m5 & m6 & m7 \\ m8 & m9 & m10 & m11 \\ m12 & m13 & m14 & m15 \end{bmatrix}$$

In a graphics API using *column-major* layout, the first set of four values `m0, m1, m2, m3` in array `mtx` represent the 1st column of the corresponding matrix, the second set of four values `m4, m5, m6, m7` represent the 2nd column of the corresponding matrix, and so on. That is, the graphics API using column-major layout would interpret the 16 values in array `mtx` as the following matrix:

$$\begin{bmatrix} m0 & m4 & m8 & m12 \\ m1 & m5 & m9 & m13 \\ m2 & m6 & m10 & m14 \\ m3 & m7 & m11 & m15 \end{bmatrix}$$

Matrices in OpenGL

In the 1980s, Iris GL [the predecessor of OpenGL] used *postfix* convention to represent matrix transformations with points and vectors represented as row matrices and transformation matrices represented using the *row-major* layout [probably because the graphics library was implemented in C]. This means the 16 values in the following array:

```

1 GLfloat trans_mtx[16] = {
2     1.f, 0.f, 0.f, 0.f,
3     0.f, 1.f, 0.f, 0.f,
4     0.f, 0.f, 1.f, 0.f,
5     tx, ty, tz, 1.f
6 };

```

are interpreted by Iris GL as the matrix:

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ tx & ty & tz & 1.0 \end{bmatrix}$$

and a vertex $P(x, y, z)$ is represented as row matrix:

$$\begin{bmatrix} x & y & z & 1.0 \end{bmatrix}$$

The transformation of P by the matrix represented by array `trans_mtx` can be described as:

$$\begin{bmatrix} x & y & z & 1.0 \end{bmatrix} \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ tx & ty & tz & 1.0 \end{bmatrix} = \begin{bmatrix} x + tx & y + ty & z + tz & 1.0 \end{bmatrix}$$

OpenGL 1.0 [released in the early 1990s] was designed to use *prefix* convention and therefore represented points and vectors as column matrices. To enable OpenGL to be compatible with Iris GL data and programs, OpenGL designers were left with no choice but to use the column-major layout for interpreting the 16 values in an array. This means that the 1st four numbers in `trans_mtx` are stored in the 1st column of a matrix, the 2nd set of four numbers are stored in the 2nd column, and so on. Hence array `trans_mtx` represents the matrix:

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 & tx \\ 0.0 & 1.0 & 0.0 & ty \\ 0.0 & 0.0 & 1.0 & tz \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

and a vertex $P(x, y, z)$ is represented as column matrix:

$$\begin{bmatrix} x \\ y \\ z \\ 1.0 \end{bmatrix}$$

and the transformation of P by matrix defined by array `trans_mtx` [in column-major format] is represented as:

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 & tx \\ 0.0 & 1.0 & 0.0 & ty \\ 0.0 & 0.0 & 1.0 & tz \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1.0 \end{bmatrix} = \begin{bmatrix} x + tx \\ y + ty \\ z + tz \\ 1.0 \end{bmatrix}$$

Commands such as `glMultMatrix`, `glLoadMatrix`, `glOrtho`, and `glFrustum` are now deprecated from the core profile but are available for use in the compatibility profile. The `glMultMatrix` and `glLoadMatrix` commands take an array of 16 values (more precisely, a pointer to the first element of an array of values of type `GLfloat`) and interpret these 16 values using column-major layout.

Core profile OpenGL [or modern OpenGL] is not concerned with matrices except for the `glUniformMatrix*` commands. These commands allow the OpenGL application to transfer matrices to shaders in *either* column-major or row-major format.

Vectors and Matrices in GLSL

A *basic type* is a type defined by a keyword in GLSL. Basic types consist of many of the fundamental types present in C and C++. In addition, many other basic types such as vectors and matrices have been added to expose the internal GPU architecture to common graphics operations. Before describing matrix basic types, the more fundamental basic data types are enumerated.

void

Just as in C and C++, GLSL uses the `void` type to indicate that functions will not return a value.

Scalars

As the name implies, *scalar* types are used to represent single values. The following table describes the basic scalar types.

| Type | Meaning |
|---------------------|---|
| <code>bool</code> | as in C++ with literals <code>true</code> and <code>false</code> provided |
| <code>int</code> | 32-bit signed integer in two's complement form |
| <code>uint</code> | 32-bit unsigned integer |
| <code>float</code> | 32-bit single-precision floating-point matching IEEE754 specification |
| <code>double</code> | 64-bit double-precision floating-point matching IEEE 754 specification |

Vectors

Vectors are built-in types and reflect the vector-processing capabilities of GPU hardware. Vectors with elements of type `double`, `float`, `int`, `uint`, and `bool` having two, three, or four components are available.

| Type | Meaning |
|--------------------|--|
| <code>vec2</code> | two-component single-precision floating-point vector |
| <code>vec3</code> | three-component single-precision floating-point vector |
| <code>vec4</code> | four-component single-precision floating-point vector |
| <code>dvec2</code> | two-component double-precision floating-point vector |
| <code>dvec3</code> | three-component double-precision floating-point vector |
| <code>dvec4</code> | four-component double-precision floating-point vector |

`ivec2`, `ivec3`, and `ivec4` types specify two, three, and four component signed integer vectors, respectively. Types `uvec2`, `uvec3`, and `uvec4` represent two, three, and four component unsigned integer vectors, respectively. Finally, `bvec2`, `bvec3`, and `bvec4` types represent two, three, and four component Boolean vectors, respectively.

It is important for you to realize that vector calculations requiring multiple CPU operations are implemented as single atomic operations on GPUs. Suppose you want to calculate the dot product of two three-component vectors on the CPU. Typically, your code would declare a C/C++ structure for a vector of `float`s:

```

1 struct Vec3 {
2     float x, y, z;
3 };

```

The dot product of two vectors would require a sum of products that would be coded like this:

```

1 vec3 u { 1.f, 2.f, 3.f };
2 vec3 v { 1.1f, 2.2f, 3.3f };
3 float sop = u.x*v.x + u.y*v.y + u.z*v.z;

```

Calculating the dot product on the CPU requires 3 multiplications, 2 additions, and 1 assignment. In contrast, GLSL can use the basic vector types in conjunction with GPUs that are optimized to perform vector operations, to execute the dot product operation as an atomic operation:

```

1 vec3 u { 1.f, 2.f, 3.f };
2 vec3 v { 1.1f, 2.2f, 3.3f };
3 float sop = dot(u, v);

```

In addition to dot products, computing cross products is commonplace in computer graphics:

```

1 // CPU code ...
2 vec3 u { 1.f, 2.f, 3.f };
3 vec3 v { 1.1f, 2.2f, 3.3f };
4 vec3 w;
5 w.x = u.y*v.z - u.z*v.y;
6 w.y = u.z*v.x - u.x*v.z;
7 w.z = u.x*v.y - u.y*v.x;
8
9 // GLSL code ...
10 vec3 a = vec3(1.f, 2.f, 3.f);
11 vec3 b = vec3(1.f, 2.2f, 3.3f);
12 vec3 c = cross(a, b);

```

While the scalar operations on the CPU require 6 multiplications, 3 subtractions, and 1 assignment, GLSL can reduce these scalar computations to a single, atomic GPU operation.

Matrices

GLSL has built-in types for containing single- and double-precision floating-point numbers in square and non-square matrices up to 4×4 in size: 2×2 , 2×3 , 2×4 , 3×2 , 3×3 , 3×4 , 4×2 , 4×3 , and 4×4 . In general, matrix types in GLSL are defined to use *column-major format*. Matrix types beginning with `mat` have single-precision components while matrices beginning with `dmat` have double-precision components. The first number in the type is the number of columns, the second is the number of rows. If there is only one number, the matrix is square. The following table illustrates the full range of matrix data types:

| Type | Meaning |
|---------------------|--|
| <code>mat2x2</code> | single-precision floating-point matrix with 2 columns and 3 rows |
| <code>mat2x3</code> | single-precision floating-point matrix with 2 columns and 3 rows |
| <code>mat2x4</code> | single-precision floating-point matrix with 2 columns and 4 rows |
| <code>mat3x2</code> | single-precision floating-point matrix with 3 columns and 2 rows |
| <code>mat3x3</code> | single-precision floating-point matrix with 3 columns and 3 rows |
| <code>mat3x4</code> | single-precision floating-point matrix with 3 columns and 4 rows |
| <code>mat4x2</code> | single-precision floating-point matrix with 3 columns and 2 rows |
| <code>mat4x3</code> | single-precision floating-point matrix with 4 columns and 3 rows |
| <code>mat4x4</code> | single-precision floating-point matrix with 4 columns and 4 rows |
| <code>mat2</code> | same as <code>mat2x2</code> |
| <code>mat3</code> | same as <code>mat3x3</code> |
| <code>mat4</code> | same as <code>mat4x4</code> |

The `dmat` types are similar to the `mat` types with the exception of having double-precision components.

Matrix initialization

Initialization of matrix types is done with constructors. You would initialize the diagonal of a matrix with all other elements set to zero like this:

```
1 mat2 m0(1.0); // initialize 2 x 2 matrix as identity matrix
2 mat3 m1(2.2); // initialize diagonal with 2.2 with other elements set to 0
3 mat4 m2(1.0); // initialize 4 x 4 matrix as identity matrix
```

To initialize a matrix by specifying vectors or scalars, the components are assigned to the matrix elements in *column-major* order:

```
1 vec2 v02 = vec2(2.0, 3.0), v12 = vec2(4.0, 5.0);
2 mat2 m0(v02, v12); // initialize 2 x 2 matrix as [v02 v12]
3
4 vec3 v03 = vec3(2.0, 3.0, 4.0), v13 = vec3(5.0, 6.0, 7.0),
5     v23 = vec3(8.0, 9.0, 10.0);
6 mat3 m1(v03, v13, v23); // initialize 3 x 3 matrix as [v03 v13 v23]
7
8 vec4 v04 = vec4(2.0, 3.0, 4.0, 5.0); v14 = vec4(6.0, 7.0, 8.0, 9.0);
9 vec4 v24 = vec4(10.0, 11.0, 12.0, 13.0); v34 = vec4(14.0, 15.0, 16.0, 17.0);
10 mat4 m2(v04, v14, v24, v34); // initialize 4 x 4 matrix as [v04 v14 v24 v34]
```

```

11
12 // initialize m3 with following column vectors
13 // (2.0, 3.0, 4.0, 5.0), (6.0, 7.0, 8.0, 9.0),
14 // (10.0, 11.0, 12.0, 13.0), (14.0, 15.0, 16.0, 17.0)
15 mat4 m3(2.0, 3.0, 4.0, 5.0,
16          6.0, 7.0, 8.0, 9.0,
17          10.0, 11.0, 12.0, 13.0,
18          14.0, 15.0, 16.0, 17.0);

```

It is important for you to understand that initializing GLSL matrices is very different from C/C++. If you define and initialize a 2D matrix of 16 elements in C/C++, the elements are specified in row-major order. To create a C/C++ 2D array similar to GLSL matrix `m3` [defined in the previous code fragment], you would define the matrix as:

```

1 // C/C++ two-dimensional array m4 defined as:
2 GLfloat m4[4][4] = {
3     { 2.f, 6.f, 10.f, 14.f },
4     { 3.f, 7.f, 11.f, 15.f },
5     { 4.f, 8.f, 12.f, 16.f },
6     { 5.f, 9.f, 13.f, 17.f }
7 };
8 // is equivalent to GLSL matrix object m4 initialized as:
9 // mat4 m3(2.0, 3.0, 4.0, 5.0,
10 //         6.0, 7.0, 8.0, 9.0,
11 //         10.0, 11.0, 12.0, 13.0,
12 //         14.0, 15.0, 16.0, 17.0);

```

Matrix length

The column-major layout of GLSL's matrix type is further stressed by the `length()` method. When applied to objects of matrix type, the `length()` method returns the number of columns of the matrix:

```

1 mat23 m1;
2 int const L1 = m1.length(); // initializes L1 to 2
3 mat32 m2;
4 int const L2 = m2.length(); // initializes L2 to 3

```

Matrix components

Just as with C/C++, the components of a matrix can be accessed using the subscript `[]` operator:

```

1 mat32 m; // uninitialized 3 x 2 matrix with 3 columns and 2 rows
2 m[2][0] = 0.0; // sets upper-right element of matrix
3 m[2][1] = 1.5; // sets bottom-right element of matrix

```

However, there are important differences between C/C++ and GLSL syntax. Applying a single subscript to a GLSL matrix treats the matrix as an array of column vectors, and selects a single column, whose type is a vector of the same size as the row size of the matrix:


```

1 mat32 m; // uninitialized 3 x 2 matrix with 3 columns and 2 rows
2 m[0] = vec2(2.0); // sets all 3 elements of first column to 2.0
3 m[1] = vec2(1.1, 2.2); // sets 3 elements of 2nd column to 1.1, 2.2, 1.0

```

In contrast to GLSL syntax, applying a single subscript to a C/C++ 2D array gives you an array of row elements:

```

1 GLfloat m4[4][4] = {
2     { 2.f, 6.f, 10.f, 14.f }, // row 1
3     { 3.f, 7.f, 11.f, 15.f }, // row 2
4     { 4.f, 8.f, 12.f, 16.f }, // row 3
5     { 5.f, 9.f, 13.f, 17.f } // row 4
6 };
7
8 GLfloat *row_ptr = m4[2]; // pointer to 1st element of row 3
9 row_ptr[3] = 21.f; // update value of last element of row 3

```

Matrix multiplication

Whether prefix or postfix convention is used is dependent on whether the matrix is multiplied by a vector or whether the vector is multiplied by a matrix. Postfix convention is used when a matrix is multiplied by a vector and prefix convention when a vector is multiplied by a matrix. To avoid confusion, you need to consistently stick to one of the two conventions. If you are not sure which convention to pick, the prefix convention is recommended since it matches the convention used in class notes.

```

1 // postfix convention with u and v thought of as row matrices
2 vec3 u, v; // assume v is initialized
3 mat3 m; // assume m is initialized
4 u = v * m;
5 // u = v * m is equivalent to
6 // u.x = dot(v, m[0])
7 // u.y = dot(v, m[1])
8 // u.z = dot(v, m[2])
9
10 // prefix convention with u and v thought of as column matrices
11 u = m * v;
12 // u = m * v is equivalent to
13 // u.x = m[0].x * v.x + m[1].x * v.y + m[2].x * v.z
14 // u.y = m[0].y * v.x + m[1].y * v.y + m[2].y * v.z
15 // u.z = m[0].z * v.x + m[1].z * v.y + m[2].z * v.z

```

Matrices in GLM

As indicated [here](#), GLM is based on GLSL and therefore uses *column-major layout*. This means that the following initialization of a matrix using GLM type `glm::mat4`:

```

1 glm::vec3 displacement(10.f, 20.f, 30.f);
2 glm::mat4 trans_mtx(glm::vec4(1.f, 0.f, 0.f, 0.f),
3                     glm::vec4(0.f, 1.f, 0.f, 0.f),
4                     glm::vec4(0.f, 0.f, 1.f, 0.f),
5                     glm::vec4(displacement, 1.f));

```

represents a translation matrix in prefix convention:

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 & 10.0 \\ 0.0 & 1.0 & 0.0 & 20.0 \\ 0.0 & 0.0 & 1.0 & 30.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

The orthographic projection transformation matrix for the OpenGL graphics pipe has the form:

$$M_{ortho} = \begin{bmatrix} \frac{2}{r-l} & 0 & 0 & -\frac{r+l}{r-l} \\ 0 & \frac{2}{t-b} & 0 & -\frac{t+b}{t-b} \\ 0 & 0 & -\frac{2}{f-n} & -\frac{f+n}{f-n} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where l and r specify the coordinates for left and right vertical clipping planes; b and t specify coordinates for bottom and top horizontal clipping planes; n and f specify the distances to near and far depth clipping planes. You can use method `glm::ortho()` to define a matrix for an orthographic parallel viewing volume:

```

1 glm::mat4 ortho_mtx = glm::ortho(l, r, b, t, n, f);

```

Or, you could directly specify the orthographic projection matrix:

```

1 // don't forget that GLM uses column-major layout ...
2 glm::mat4
3 ortho_mtx(glm::vec4(2.f/(r-l),    0.f,    0.f,    0.f),
4           glm::vec4(0.f,    2.f/(t-b),    0.f,    0.f),
5           glm::vec4(0.f,    0.f,    -2.f/(f-n),    0.f),
6           glm::vec4(-(r+l)/(r-l), -(t+b)/(t-b), -(f+n)/(f-n), 1.f));

```

Final words

Matrices must be represented as 2D arrays in a computer program. While the C/C++ programming languages represent a 2D array using row-major order, there are other programming languages - FORTRAN comes to mind - that use column-major order. Since there are prefix and postfix conventions on paper and row- and column-major orders of representing matrices in memory, there are four possible combinations to represent mappings in a computer program. This document has explained the particular conventions used by OpenGL, GLSL, and GLM. As users of these APIs, you should not be concerned with the programming language in which these APIs are authored.

Instead, you should always use the interfaces provided by these APIs to initialize, update, and use these matrices.